A METHOD FOR EXTRACTION OF AUDIO-VISUAL LEITMOTIF IN MOVIES BY CROSS MEDIA ANALYSIS.

Jenny Benois-Pineau, Myriam Desainte-Catherine, Nicolas Louis
LABRI UMR 5800, Université Bordeaux 1, France
e-mail : jenny.benois@labri.fr, myriam@labri.fr, louis@labri.fr

ABSTRACT

The joint analysis of video and audio components in multimedia documents has been widely used since the beginning of activities related to the new multimedia Standard MPEG7. In this context the paper is focused on a method for extraction of an emotional and structuring cue from artistic content we call "audio-visual leitmotif", which is a first step in characterization of an author style in producing the content. The method is based on joint motion –based video partitioning and on model– based music recognition.

1. INTRODUCTION

With the emergence of the new multimedia content description standard MPEG7 [1], a vast amount of works have appeared devoted to partitioning of video documents into meaningful segments and characterization of them [2]. The standard MPEG7 adopts a hierarchical recursive model of segments– a segment tree which can be mapped onto a semantic model of video content. Taking into account the way of composing an "author" video document (a movie, a documentary), this hierarchy can be represented as a three–level partition of a whole document into video shots, sequences and scenes [3]. A shot is the easiest segment to isolate if the problem of an automatic partitioning of video is addressed. A rich collection of methods to segment video into shots has been presented in literature and new methods still appear [4] handling complex transition effects and based on low level video signal analysis. Sequences and especially scenes are more difficult to distinguish, as they refer to semantics and namely are very much dependent on an individual approach of the content creator. In the pioneer works [5,6] a joint analysis of video and audio components in video documents was developed in order to resolve the problem of an over segmentation into shots and to propose more semantic logical partition of the content into story units corresponding to scenes. In the same context the works on segmentation of the audio stream according to classes Speech/Music/Noise [7] can be mentioned. Without addressing a vast area of speech recognition used for join audio and video analysis, another typical approach addresses the problem of jingle recognition, that is a specific sound related to the type of document. Thus in [8] the problem of laughter detection is addressed.
In this work we try to make the first step to the characterization of style, composition and to description of a macro-structure of "author" cinema and documentary. We introduce a notion of an "audio-visual leitmotif" which is basically a piece of musical content in audio stream combined with a characteristic visual information. This paper presents a method for its definition in a given document and its recognition along the time. The paper is organized as follows. Section 2 presents a general cross indexing framework for leitmotif definition, video analysis tools are presented in Section 3. Extraction of musical information is described in Section 4. Some results and discussion are presented in Section 5.

2. GENERAL CROSS-MEDIA INDEXING SCHEME

We suppose that in an artistic content an association of temporally distant segments can be done using the same (or similar) musical fragment associated with the presence of some characteristic elements in video. We call such a combination of audio and visual information "an audio-visual leitmotif" (AVL). In this paper we propose a first simplified model of AVL, which combines a model of music content and a simple temporally ordered collection of video shots. The temporal boundaries of this collection enclose musical content. The handling of AVL in a document requires solving 3 problems : definition of AVL fro the given document, recognition of AVL inside a document, visualization of AVL for retrieval in interactive systems. As far as visual component of AVL is concerned, the fundamental step for extraction and recognition is segmenting of video stream into shots and determining shots which enclose a specific musical content. For the visualization we propose to represent the visual part of AVL by key-frames corresponding to homogeneous camera work.
Musical component is characterized by the score which is played, by the instruments that are playing it and by the interpretation of the musicians. From a score, a musical

style should be extracted (baroque, romantic, jazz, etc.) as well as a musical form (sonata, concerto, etc.). Thus, modelling of an audio component can be done at several levels. Separating levels is important in order to permit a precise segmentation. Thus, an AVL should be represented by the three attributes in order to detect any ``variation'' of it. For example, an occurrence of a variation of an AVL can have the same score, but it can be played by a different instrument, or a different interpretation (faster or slower). In the same way, another occurrence of an AVL can have a different score of the same style.
The global indexing scheme we propose is semi-automatic. That is the audio-visual leitmotif is isolated when it appears for the first time in the document. The temporal boundaries are approximately indicated by operator based on audio perception. Then based on video analysis, the boundaries are automatically positioned to the boundaries of enclosed shots detected by video analysis and structuring method. After that the model of AVL musical content is automatically built and the AVL is searched for along the time in the document, every time adjusting its boundaries by video analysis. For the visualization of AVL, an automatic shot selection is done based on a segmentation into micro–shots.

3. VIDEO ANALYSIS AND STRUCTURING

The goal of video analysis and structuring is to segment the whole video document into shots and micro-shots, characterized by homogeneous camera motion.
Taking into account a vast amount of artistic content already available in a compressed form, we propose to fulfill this task on MPEG2 or MPEG1 compressed documents. In [8] we proposed a shot change detection algorithm based on macroblock optical flow. In this paper we apply it and also complete this algorithm by characterization of global camera motion which allow a segmentation into micro-shots.
This approach is based on estimation of global 2D motion model in video which is a projection of a 3D camera motion into 2D image plane.
In MPEG1,2 video stream a part of images is predicted by motion compensation (P-frames). This means that a motion vector $(dx_i,dy_i)^T$ is available for i-th macro-block in image plane. A part of macro-blocks is intra-coded.
We suppose that the motion vector field follows a global affine model expressed as

$$\begin{pmatrix} dx_i \\ dy_i \end{pmatrix} = \begin{pmatrix} a_1 + a_2 x_i + a_3 y_i \\ a_4 + a_5 x_i + a_6 y_i \end{pmatrix},$$

where $(x_i,y_i)^T$ is the co-ordinate vector of the center of i-th block. In [10] a physical interpretation of the six parameters is referred to the camera motion. Namely pan, tilt, zoom, rotation and hyperbolic terms corresponding to perspective deformation :
$pan=a_1$ $tilt=a_4$
$zoom=(a_2 + a_6)/2$ $rot=(a_5 + a_3)/2$
$hyp_1=(a_2 - a_6)/2$ $hyp_2=(a_3 - a_5)/2$
Taking into account the noisiness of MPEG macro-block optical flow, in [9] we introduced a weighted least-square robust motion estimation scheme with outlier rejection based on Tuckey estimator.
Here the parameter vector $q = (a_1,\ldots a_6)^T$ is estimated as

$$\hat{q} = \arg \min \sum r(r),$$

where r is Tuckey estimator[10], r is the residual between motion vector from MPEG optical flow and that one derived from the global model. The shot change detection algorithm is based on the assumption of a change in the following characteristics. Firstly, the number Q of macro-blocks intra-coded in P-frames can indicate a change in the spatial content of the scene. The change of motion parameters A shows a break in camera work, which can be naturally expected in different shots. Finally the mean squared error of model estimation MSE also increases if the model change.
The global measure we proposed in [9] to detect shot changes is

$$D(k) = (Q(k) + a)(MSE(k) + a)(A(k) + a)$$

with $A(k) = \sum_{n=1}^{6} \Delta a_n(k)$

$\alpha = 1$ and Δ a relative absolute difference. It has peaks in shot boundaries.
In order to segment shots into micro-shots we now propose to use a significance test which we adapt from [9]. The significance of each motion parameter a_i is tested, then for a given video segment only significant parameters are retained and the corresponding camera work (see the relationships above) is deduced.
For each parameter a_i two competing hypotheses are considered. The first one H0 assumes that the considered parameter is significant, the second one H1 considers that on the contrary, $a_i=0$. The five other parameters are left free in both cases. For each hypothesis, an associated likelihood function is defined, where the considered random variables are the residues r. They are assumed to be independent and to follow zero-mean Gaussian law of the same variances, which are a posteriori estimated as

$$\hat{s}^2_{m_l} = \frac{1}{N_d} \sum_{i \in S_d} r_{il}^2, \quad l=0 \text{ or } 1,$$

with N_d and S_d the number of measures and S_d the estimation domain.

Based on Gaussian distributions, the logarithmic likelihood test is expressed as :

$$\frac{N_d}{2}\left(\ln\left(\sum_{i\in S_d} r_{i0}^2\right)-\ln\left(\sum_{i\in S_d} r_{i1}^2\right)\right)\begin{matrix}H0\\ \lessgtr\\ H1\end{matrix} l_A$$

These scheme supposes to estimate motion parameter vector q twice. Firstly for the hypothesis H0 the full weighted least square estimation is done. For the H1, we make benefit from the knowledge of dominant estimation support S_d that is only MPEG motion vectors which were not rejected as outliers by H0 estimation. Thus N_d and S_d correspond to non-outliers in initial measures.
The likelihood threshold l_A is not critical for the translational components, as canceling of one of them causes a dramatic rise in the sum of squared errors r^2 if the component is significant. In order to chose l_A components $a_3..a_6$ we use an a priori probability of hypotheses. After the test is fulfilled, we compute the dominant "physical" motion and select representative key frames according to the following simple rules. If a translational or rotational motion is dominant, then only a middle image in the shot is taken as key frame, if the zoom component is significant in any combination with other parameters, we select the first and last frame in the shot as key-frames.

4. EXTRACTION OF MUSICAL INFORMATION

Actually several works are carried out in the domain of extracting informations from music. Extraction of low-level descriptors [11] should lead to instruments characterization and classification based on the audio signal. But such extractions use analysis methods which always depend on the type of sound. High-level descriptors have a perceptual character. A musical component should be described as a combination of abstractions based on perception, as well as a visual scene can be segmented into forms, positions and textures. But, for now, we are very far from that goal, because we lack a general perceptual model as well as general and robust methods for analysis. For example, separation of several instruments that are playing together is still an open problem. In order to characterize the musical content of an audio signal, two methods are possible: extract the score and analyze it at a symbolic level [12] and rely high-level descriptors for music (style, rythm) to low-level descriptors of the signal [12] . We choose the first method because it is more powerful (once the score is known, several symbolic analysis can be performed on it) and the musical signal of the video was simple enough to permit extraction of the score and modelling of the instrument.
In this paragraph, we speak about the analysis and modelling of audio content of an AVL, the audio signal of the musical segment we studied is a sequence of chords of two notes. In our study, we admit that the instrument is harmonic and percussive. Then, under this hypothesis, we proceeded to segmentation of the audio signal into a sequence of chords by analysing the rms -amplitude evolution. Each attack of a chord corresponds to a great peak in energy (see Fig 2). This segmentation provides the durations (intervals between two peaks) and the volumes (amplitudes peaks) of the chords. Then, each chord is analysed with the DFT1 algorithm [13] in order to get the partials of the resonnance part and extract the fundamental. In this part, the end of the previous chord must be subtracted. For that purpose, we examine the evolution of the amplitude of each partial in order to determine weither it belongs to the previous chord (its amplitude continues to decrease) or to the new chord (a percussive peak can be detected). Then, a chord is separated into notes. The first note is obtained by extracting the harmonics of the lower fundamental, then other notes are obtained in the same way with the remaining partials. Thus, from each note of the score, we extract a duration, a volume, a fundamental frequency and the normalized amplitudes of harmonics. Those note occurrences are grouped together into note models, corresponding to instrument notes. A note model is composed of a pitch (fundamental frequecency) and the sets of characteristics extrated from its occurrences, that is, volumes and normalized amplitudes of harmonics of its occurrences. The model of the instrument is constituted with a series of notes models with increasing pitch.
The audio component of AVL modelling is constituted of two parts : the instrument model and the score. The latter is a series of chords with duration, whose notes are references to occurrences notes of the instrument model. At least, we study detection of AVL occurrences, once the AVL is analysed and modeled, detection of an exact occurrence of it (same score, same instrument and same interpretation) is based on the recognition of the rms -evolution, then of each note, and then of the serie of chords constituting the score. Recognition of the notes is performed by comparing harmonics of the note occurrence with notes of the instrument model.

5. RESULTS AND PERSPECTIVES

In this work, we limited the experiments on AVL which contain harmonic musical content. Namely, the harpsichord melody was chosen in the "La joueuse de Tympanon" movie ® SFRS. The acoustic model was constructed on the first chosen fragment and then the content was analyzed in order to recognize the AVL. In order to test the performance of the algorithm we selected "positive" fragments in the document containing the AVL "negative" examples where the AVL was not present and "uncertain"

examples where the AVL was disturbed by background noise.

The sound analysis was initially tested on two fragments of “La joueuse de Tympanon” ® SFRS, the global duration of the analysis was 20 seconds, that is 60 notes. We determined if a note was played or not by a tympanon using the spectrum analysis .

The video analysis and structuring method was widely tested on fragments of “L’homme de Tautavel”, “Le temps de Lagunes” and “La Joueuse de Tympanon” ® SFRS . The global duration of the processed content was 481,8 seconds, that is 12045 frames at Pal frame rate were processed. The performance of the shot change detection algorithm was measured in recall and precision terms. The tests showed that the shot change detection algorithm fails if a scene change happens on the frame I or on the neighbouring B-frames in MPEG stream. Otherwise all shot cuts and progressive dissolves were detected. The worst recall and precision observed on “La joueuse de tympanon” were of 74,2% and 88,4% respectively. The characterisation of global motion by proposed significance test totally corresponded to the visual interpretation of apparent motion in the document. An example of the shot change detection and key-frame selection is given in Figure 1 The significance of motion parameters is depicted in gradations of grey value.

The immediate perspective of this work is in extending the recognition algorithm based on similarity of visual descriptors of the video component of AVL and also constructing an acoustic model using a larger database.

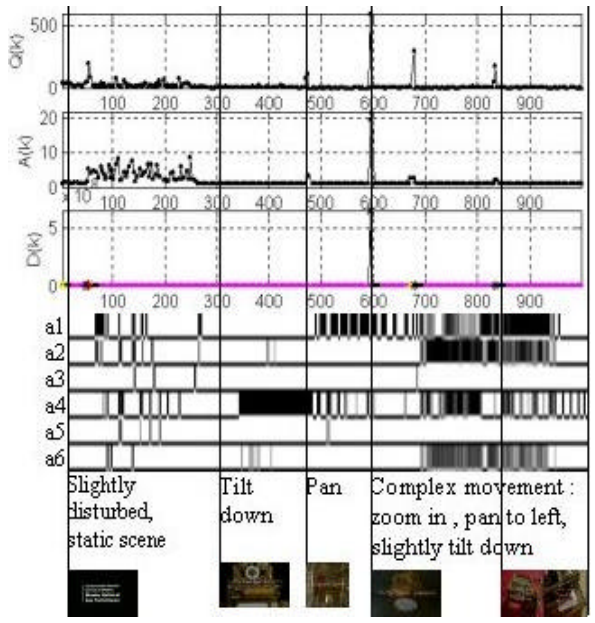

Fig 1 : Shot Change Detection on Tympanon

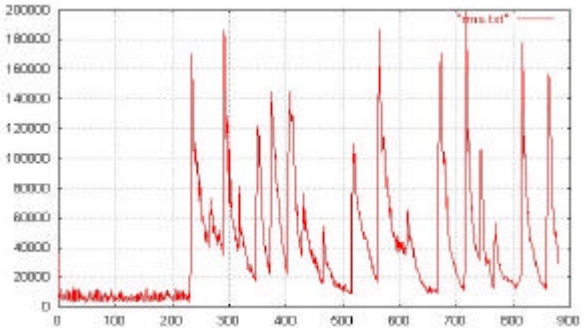

Fig 2 : Energy of piece of an audio signal

REFERENCES

[1] MPEG-7: Context, Objectives and Technical Raodmaps (V.12), ISO/IEC JTC1/SC29/WG11/N2861, July 1999

[2] PH. Salembier, J.R.Smith, “MPEG-7 Multimedia Description Schemes”, *IEEE Trans on CSVT*, vol 11, N6, 2001, pp. 748 –759

[3] Ph. Joly., H.-K. Kim “Efficient automatic analysis of camera work and microsegmentation of video using spatio-temporal images “ *SP:IC* 8, 1996, pp. 295-307

[4] C-L Huang, B-Y Liao, “A robust Scene-Change Detection Method for Video Segmentation”, *IEEE Trans. on CSVT*, vol 11, n°12, December 2001, pp. 1281 - 1288

[5] C. Saraceno, R. Leonardi, ”Audio as a support to scene change detection and characterization of video sequences” *in Proc ICASSP’97, Munich, Germany, April 1997*, pp 2597-2600

[6] C. Saraceno, R. Leonardi, “Identification of Story Units in Audio-visual Sequences by Joint Audio and Video Processing” in *Proc ICIP’98*, Chicago, IL, USA, Oct, 1998

[7] J.Pinquier, Ch. Sénac-Dours, R. André-Obrecht . « Indexation de la bande sonore : recherche des composantes Parole et Musique » in 13th (RFIA'2002) , Angers, France. AFRIF-AFIA, January 8-10 , 2002, pp. 163-170

[8] K. Kaneda, M. Sugiyama, “Laughter Detection for video caption generation”, in *Proc. CBMI’2001*, Brescia, Italy, 19- 21 September 2001, pp151 – 158

[9]M. Durik, J. Benois-Pineau, “ Robust motion characterisation for video indexing based on MPEG2 optical flow”, in *Proc. CBMI’2001*, Brescia, Italy, 19- 21 September 2001, pp57 – 64

[10] P. Bouthemy, M. Gelgon, F. Ganansia, “Aunified approach to shot change detection and camera motion characterization”, *IEEE CSVT*, vol. 9, N7, 1999, pp 1030-1044

[11] S . Rossignol et al, “Automatic characterisation of musical signals : feature extraction and temporal segmentation”, Journal of New Music Research, 2000

[12] Antti Eronen, “Automatic musical instrument recognition”, MSc Thesis, Tampere University of Techonology, 2001

[13] M .Desainte-Catherine and S .Marchand, “High precision Fourrier analysis Using Signal Derivatives”, Journal of the Audio Engineering Society, 48, p654-667, July/August 2000