Analysis of Video Content for a Multi-Layer Navigation of Multimedia Documents

M. Bonnet*, A. Bugatti°, R. Leonardi° & P. Migliorati°
*Laboratoires d'Electronique Philips, Paris - France
°Department of Electronics for Automation, University of Brescia
Via Branze 38, I-25123 Brescia – Italy
Email: leon@ing.unibs.it

Abstract

This paper describes a set of automatic extraction tools so as to generate a three-layer organization of video documents. The underlying coarse to fine description allows for a fast navigation throughout the document, depending on the degree of details which is desired. Once the time-codes of the individual segments for each layer of the hierarchy have been identified, it is possible to map them into a Description Scheme (DS), which maintains the hierarchy and linear structure of the video document. This structural DS serves the role of a table of content for the multimedia document, the same way it is done in books. The particular interest of the proposed approach lies in the automatic solutions that can be used to generate the different segments at each level of the DS, and in the browsing tool that can be easily derived to navigate throughout the document.

1. Introduction

On top of future digital video broadcast services, it is necessary to enable a better exploitation of multimedia information resources by non-IT experts. In order to facilitate this task, multimedia information must be enriched by a description of its content, so as to provide intuitive modalities for naive users to search audio-visual information. This should also enhance the quality of a broadcast service by providing new features and more interactivity.

If content description is desirable, it is clear that this should be achieved with a minimal increase in cost. Clearly only nearly automatic extraction techniques of the content description are likely to keep the costs at reasonable levels. As most TV programmes/movies typically have an associated script which specifies the structure of the associated audio-visual document from the programme level to individual programme items, it is necessary to generate intermediary descriptions of the content to reach the "physical" layer of individual frame play-back.

This is the scope of this work, for which a three layer intermediary decomposition of individual programme items is suggested: the scene, the shot, and the micro-segment level. The scene level tries to identify simple patterns of visual content, which are typical of actions, dialogues, stories…; at the shot level individual camera records can be found; finally, the micro-segment level provides a partition of individual shots into temporal segments exhibiting consistent camera motion characteristics.

2. Table of Content (ToC) DS

As each individual element of each layer can be decomposed into a set of elements of the next layer, it is clear that the structure of the decomposition can be mapped into a particular DS, which serves the role of a table of content of the video document (see Figure 1). The structure of this table of content can be easily represented using the current MPEG-7 segment DS (normal decomposition) which is part of the generic DS for audio-visual documents (generic AV DS) [1].

However, for maintaining the simplicity of the normal decomposition in the ToC DS, semantic information (such as scene type, camera motion, mosaics, editing effect type…) is directly added at each layer of the decomposition, by associating proper sub-DS, or descriptors. These include semantic information such as background and foreground mosaics, camera motion parameters, editing effects, … The X-schema representation of the necessary extensions to the normal segment DS will be provided to incorporate such changes.

It is to be noticed that the previous ToC DS can be extended thanks to an analytical index DS [2] which is able to highlight description items and order them through pre-defined mechanisms. These representative highlights then give links to a set of leaves or nodes of the ToC tree. For more details regarding this concept please refer to contribution M4906 that was presented at the MPEG meeting in Vancouver [3].

This work has been partially funded by the European Community ESPRIT programme (project 28798 AVIR), Philips LEP, and the Italian Ministry of the University and of Scientific and Technological Research.

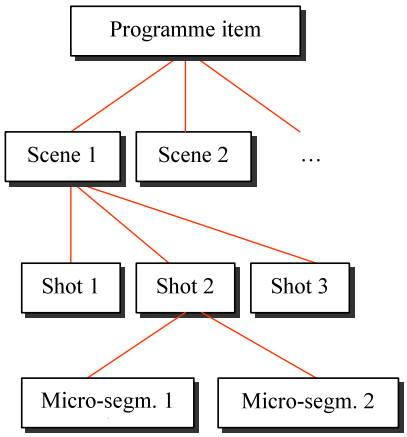

Figure 1. Table of content tree.

3. Automatic extraction of the ToC DS

The tools needed for automatic extraction are applied in order to obtain:

1. Individual shot separation: by extraction of editing effects between consecutive camera records. This can be achieved by making use of the statistical independence of the two shots that are present on both sides of the editing effect; in the case of dissolves, fade-in, or fade-our, refer to the algorithm presented in [4].
2. Shot partitioning into micro-segments: by classification of the underlying consistent dominant camera motion present throughout the micro-segment duration [5].
3. Shot grouping into scenes: by identification of peculiar alternation of visual patterns between consecutive shots, so as to recognize characteristics situations such as dialogues, actions, … The visual correlation between non consecutive shots is established thanks to a vector quantization approach, which compares codebooks associated to the individual shot patterns [6].

A detailed presentation of these extraction methods is provided in the next session.

3.1 Shot cut detection and classification of editing effects

An image sequence is formed of a certain number of shots that correspond to a series of frames typically extracted from a single camera record. During the editing stages, each camera record is linked to other by means of editing effects such as cuts which simply define a sharp transition, wipes which correspond to a sliding window effect, mates which let a dark blob invade the frames and dissolves, which represent a gradual change from one camera record to another by simple linear combination of the frames involved in the dissolve process. The weights of the linear combination are typically set on the basis of the distance of the frame which is part of the dissolve with respect to the beginning and the end of the surrounding shots respectively. Other effects such as fade-in and fade-out are also used; in the first case, it simply corresponds to a dissolve from a continuous black shot to a real camera record, while in the second case, a camera record is progressively darkened. A reliable solution for the identification of a dissolve has been proposed recently, with a two stage approach: the first one is a detection process, while the second aims at locating with precision the time interval that the editing effect is covering.

It is assumed that a dissolve is created by overlapping two camera records for the duration of the dissolve and by weighting the contribution of each frame at any given instant n on the basis of its location with respect to the boundaries of the editing effect.

Assuming that the series of frames forming each shot are outcomes of a same stationary random process (at least for first order statistics), an estimate of the marginal pdf of each process is represented by the last frame of prior to the dissolve.

Normally the two frames are quite different so that their associated histograms are those of two independent random variables.

Accordingly for any frame being part of the dissolve, its associated first order statistics can be estimated by the convolution of the histograms of the boundary frames properly scaled.

This implies that the difference between the actual histogram and that obtained by convolution should ideally be zero. On the contrary, if the two frames were images of a same shot, the previous histogram difference would be different from zero. From this simple consideration, it is possible to obtain a simple criterion for dissolve detection [4].

3.2 Identification of micro-segments

In order to generate a Table of Content of a video program, it is necessary to perform a temporal segmentation of the video, to represent it hierarchically from its root (the whole video) to its leaves (its smallest temporal segments). In many cases, video shots, may not be appropriately considered as leaves of this representation. It may be necessary to over-segment them

into smaller entities. For instance, in applications such as sport, video shots can be very long (typically in soccer or tennis games). We thus proposed [5] a method for over segmenting them into micro-segments that have an homogeneity in terms of camera motion.

The procedure requires three stages to extract shots and micro-segments from a video sequence. The purpose of the first step of the process is to split the sequence into video shots, in order to be able to compute the camera motion parameters for each shot. It is important to remark that shots yielded by this step do not need to cover the whole sequence, because it makes no sense to estimate the camera motion parameters on those frames which belong to a special transition, like a fading. The algorithm developed in [4] takes these special effects into account to provide its segmented shots.

Once the sequence has been split into shots and camera motion parameters have been estimated on each shot, the third stage of the process is applied. The purpose of this block is to split each shot into a set of micro-segments which present a high level of homogeneity on camera motion parameters. The algorithm described in the sequel corresponds to this stage.

It can be summarized as follows: first, thanks to the data provided by the camera motion extraction algorithm, the shot is over segmented into several micro-segments which must present a perfect homogeneity. Second, a merging process is applied while the homogeneity level of the set of micro-segments remains over a predefined threshold [5].

On one side we assume that a segment is perfect when it presents a single combination of camera motion parameters in all its frames. On the other side, we assume that a segment is bad when it presents important variations on these parameters. Thus, in order to measure the homogeneity of a segment we rely on its histogram. If a segment is perfect, then the bins of its histogram will be equal either to 0.0 (the considered motion does not appear at all) or to 1.0 (the motion appears on the whole segment). If the segment is not perfect, then the bins of its histogram present intermediate values.

In order to measure the homogeneity of a segment we measure how much its histogram differs from the ideal one.

3.3 Visual scene segmentation

In order to generate indices that can be used to access a video database, a description of each video sequence is necessary. The key to understand the content of a video sequence is based upon a temporal scene segmentation and classification. Shot cut detection has been used as a technique to segment video sequences. Unfortunately, visual information even when organized into consecutive shots does not always convey semantically meaningful information. A more adequate representation can be obtained when groups of consecutive shots can be merged into semantically more coherent entities called scenes.

With this objective in mind, recent works have demonstrated the potential of analyzing the associated audio signal for video scene identification and classification [6]. A cross-modal analysis of low-level visual and audio features has often brought intermediate but satisfactory semantic description of an audio-visual sequence. In this work the emphasis is placed on the identification of four different types of scenes (dialogues, stories, actions and events), by a joint audio-visual analysis.

The identification and characterization of scenes is an important step for speeding up the retrieval process. Normally, the process of understanding audio-visual material performed by a human being requires a joint analysis of both visual and audio signals. At a first level of abstraction, there is an unconscious grouping of segments of audio-visual material into semantically consistent scenes.

On one hand, the audio signal is processed in order to separate consecutive groups of audio frames into:

Silence segments which define those audio frames which only contain a quasi-stationary background noise, with a low energy level with respect to signals belonging to other classes.

Speech segments which contain voiced, unvoiced and plosive sounds.

Music segments which contain a combination of sounds with peculiar characteristics of periodicity.

Miscellaneous sound segments which correspond to all other categories, i.e. everything which does not belong to the previous classes.

On the other hand, visual information is analyzed first in order to detect shot cuts. For each shot, a visual feature processing unit is used: a VQ codebook is designed so as to reconstruct the associated shot with a certain distortion with respect to the original visual information associated with the corresponding shot.

The outputs of both the audio classifier and the visual feature processing units are then passed to a scene detection and classification (SDC) module, which tries to merge together consecutive shots, according to some rule. The SDC module starts to describe each shot S_i with a set of meta-data. These are the time interval in which the shot resides, the dominant associated audio class, the percentage of samples belonging to such a class (evaluated for the entire shot), the reference to the codebook associated to the shot and a label identifying the visual content of the shot.

Once such meta information (which serves the role of abstract keywords) has been extracted, it can be processed so as to identify four different types of scenes with the following rules:

- Dialogues: The audio signal is mostly speech and the change of the associated visual information occurs in

an alternated fashion, that is, the associated visual labels (which should ideally reflect a change of speaker) follow a pattern of the type ABABAB... ;

- Stories: The audio signal is mostly speech while the associated visual information exhibits the repetition of a given visual content, to create a shot pattern of the type ABCADEFGAH...;
- Actions: The audio signal belongs mostly to one class (which is not speech) and the visual information exhibits a progressive pattern of shots with contrasting visual contents of the type ABCDEF...;
- Generic scenes: Consecutive shots which do not belong to one of the aforementioned scenes but their associated audio is of a single consistent type.

4. The visual interface

The navigation tool through the ToC DS maintains the hierarchy of abstraction, which is provided by this description scheme. Thus, it is possible to move backward and forward through each individual micro-segment, shot or scene.

At the lowest levels (micro-segments, shots), a K-frame provides the necessary visual summary of the content. Indication of the type and extent of editing effects are also provided. Mosaics may be shown for certain micro-segments or shots, so as to give an instantaneous view of the background that will be covered during the extent of the corresponding video segment. These are extracted automatically by blending after appropriate warping, the background information of each individual frame.

At the scene level, an icon and textual summary indicates the type of visual scene, which is occurring. Finally, it possible to playback the video sequence at any level of the ToC hierarchy for the entire duration of the corresponding segment (scene, shot, micro-segment).

Figure 2 presents the typical display at the scene level of this navigation tool.

5. Conclusion

This paper describes a set of automatic extraction tools so as to generate a three-layer organization of multimedia documents. This creates a sort of Table of Content (ToC) description of the document, similarly to what is done for technical books. Further studies are being conducted to allow a more accurate description of the programme content by automatic extraction methods, especially at the highest level of the ToC, i.e. the scene. Efforts are also being devoted to map this ToC representation using the guidelines provided in MPEG-7 for the Generic Audio-Visual DS.

References

[1] MPEG Description Scheme Group. MPEG-7 Description Schemes (v 0.5). *ISO/IEC JTC1/SC29/WG11* Doc. **N2844**, Vancouver, Canada, Jul. 1999.

[2] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati, L.A. Rossi, and C. Saraceno. The TOCAI description scheme for indexing and retrieval of multimedia documents. to appear in European Workshop on Content Based Multimedia Indexing '99, Toulouse, France, Oct. 1999.

[3] L. Rossi and R. Leonardi. A Possible Extension of the Generic AV DS to Incorporate Highlighting and Ordering Functionalities. Proposal M4906 submitted to the MPEG meeting Vancouver, Jul. 1999.

[4] N. Adami and R. Leonardi. Identification of editing effects in image sequences by statistical modeling. In *Proc. Picture Coding Symposium '99*, Portland, OR, U.S.A., Apr. 1999.

[5] J. Llach i Pinsach and P. Salembier: Analysis and indexing of video sequences: table of content and index creation. submitted to *VLBV'99*, Japan.

[6] C. Saraceno and R. Leonardi: Indexing audio-visual databases through a joint audio and video processing. *Intern. Journal of Imaging Systems and Technology*, 9(5):320-331, Oct. 1998.

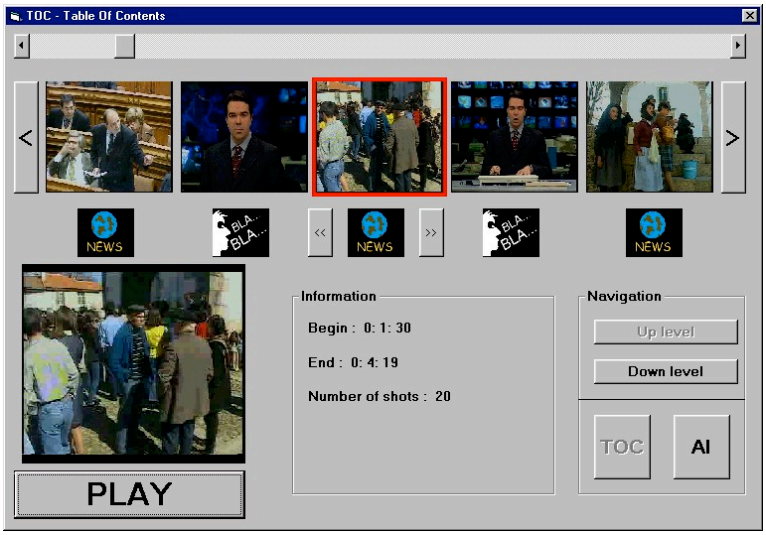

Figure 2. Navigation tool display throughout the video programme table of content