LOW LEVEL PROCESSING OF AUDIO AND VIDEO INFORMATION FOR EXTRACTING THE SEMANTICS OF CONTENT

N. Adami, A. Bugatti, R. Leonardi, P. Migliorati
DEA - University of Brescia , Via Branze, 38 - 25123, Brescia, Italy
Tel: +39 030 3715434; Fax: +39 030 380014
e-mail: leon@ing.unibs.it

Abstract - The problem of semantic indexing of multimedia documents is actually of great interest due to the wide diffusion of large audio-video databases. In the first part of this paper we briefly describe some techniques used to extract low-level features (e.g., shot change detection, dominant color extraction, audio classification, ...). Then the TOCAI framework for content description of multimedia material is presented, together with an application which implement it. Finally we propose two algorithms suitable to extract the high level semantics of a multimedia document. The first is based on finite-state machines and low-level motion indices, whereas the second uses Hidden Markov Models.

INTRODUCTION

Effective navigation through multimedia documents is necessary to enable widespread use and access to richer and novel information sources. Design of efficient indexing techniques to retrieve relevant information is another important requirement. Allowing for possible automatic procedures to semantically index audio-visual material represents a very important challenge. Such methods should be designed to create indices of the audio-visual material, which characterize the temporal structure of a multimedia document from a semantic point of view.

In the first part of this paper we address the problem of low-level features extraction showing different types of extraction algorithms, and some examples of possible use of them to perform simple tasks. Then a very powerful tool for browsing and retrieval of interesting events inside a programme is shown, and the Description Scheme (ToCAI) on which is based is also described. Finally we show some interesting results in joint audio-video analysis based on the low-level features previously extracted.

EXTRACTION OF LOW-LEVEL FEATURES

We have adopted several tools in order to obtain automatic feature extraction for describing an audio-visual document. These methods can be divided in two classes: audio and video.

For video analysis we developed the following methods.

- Individual shot separation is achieved by extraction of editing effects between consecutive camera records. This can be obtained 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 [1].

- Shot grouping into scenes is obtained by identification of peculiar alternation of visual patterns between consecutive shots, so as to recognize characteristics situations such as dialogues, actions and so on. The visual correlation between non consecutive shots is established thanks to a vector quantization approach, which compares the codebooks associated to the individual shot patterns [2].

- Dominant color extraction is very useful to have content-based retrieval for color, either for the whole image or for an arbitrarily shaped region. It is calculated using the color histogram and a measure of confidence which is high in cases where pixels of dominant color represent the majority of the pixels in the object, and it is low in the other cases.

Considering the audio information, we developed some simple but effective methods in order to have a significant segmentation of the audio stream that can be useful for browsing or retrieval of interesting events.

- Audio segmentation in homogeneous segments of speech and music is obtained using two different approaches: the first approach, based mainly on Zero Crossing Rate (ZCR) and Bayesian Classification, is very simple from a computational complexity point of view [3]. The second approach, based on Neural Networks (specifically a Multi Layer Perceptron, MLP), allows better performance at the expenses of an increased computational complexity [4].

- The average audio loudness extraction related to each shots. The shots having the higher audio loudness turned out to be very meaningful from the semantic point of view (but only in some kinds of programme: for instance, in a soccer match there are goals, roar crowd after goals, and so on).

TOCAI: A TOOL FOR BROWSING AND RETRIEVAL

All the previous information can be used inside our application developed to provide a tool for improved browsing and retrieval of multimedia documents. This tool is based on a framework called ToCAI [5], [6], which is composed by two parts. The first one, called ToC (Table of Contents), characterizes the temporal structure of a multimedia document from a semantic point of view at multiple levels of abstraction, so as to have a series of consecutive segments which are coherent in terms of the semantic of information at that level. The second one, called AI (Analytical Index), allows an easy way to effectively retrieve relevant information, such as objects appearing in the video, or identify specific events of interest (e.g., a murder in a thriller movie or a goal in football match). To these ends, it is important that the objects or events are arranged in an appropriately designed index, according to criteria meaningful for the application context. A more detailed explanation can be found on the web [7].

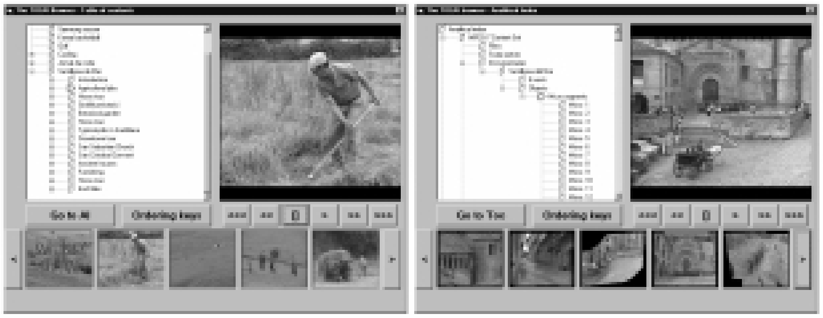

Figure 1: An interface view that shown the ToC and AI part.

The ToC part has been built using the shot and scene separation algorithms explained above. For the AI part some heuristic methods have been used in order to obtain objects ordering useful for retrieval operations. Another feature used is a measure of the camera motion, which has been very useful in sport documents to extract the shots with a player having fast movement.

To improve the capabilities of this tool some approaches based on joint audio-video analysis have been studied, as showed in the next paragraph.

JOINT AUDIO-VISUAL ANALYSIS

We studied two promising approaches to extract more useful information from audio and video low-level features. The first approach uses MPEG motion vectors and finite state machines in order to extract relevant events in a soccer game [8], whereas the second one uses Hidden Markov Model to achieve a better scene segmentation [9].

Soccer video indexing using MPEG motion vectors

The algorithm is focused on analysing descriptors (mainly MPEG2 motion vectors) and related data to search for particular features in correspondence of interesting scenes in soccer games, for example the event of a scored goal. When a goal is scored, the following features are present:

- a fast pan;
- a fast zoom, used to focus the attention on a player or on the ball;
- a sequence of static frames, while the goal keeper is framed (this is optional);
- a scene cut.

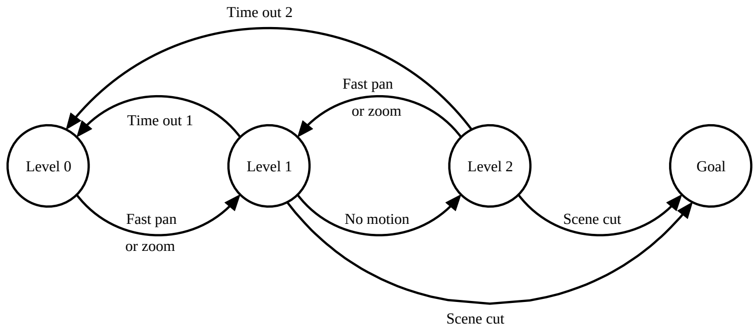

Figure 2: The final algorithm.

The detection of the above mentioned features is implemented as follows:

- Camera motion parameters, represented by horizontal ”pan” and ”zoom” factors, have been evaluated using a least-mean square method applied to P-frame motion fields. We have detected fast horizontal pan (or fast zoom) by thresholding the pan value (or the zoom factor). When a fast pan or a fast zoom is present in 3 consecutive P-frames, level1 is set;

- Static frames are detected using average magnitude of motion vectors m: when m takes low values for 3 consecutive P-frames, and level1 is present, level2 is set;

- Scene cuts are detected using a weighted sum of intra-coded Macro-Blocks and difference of average magnitude between the current and the future frame. When this sum takes a high value, a shot cut is present and if level1 or level2 are present, level3 is set.

The final algorithm is shown in Fig. 2.

All levels are mutually exclusive; when a level is present and the event necessary to set the next level doesn’t occur for a certain number of frames, the system is resetted (level0 is set). The algorithm to find penalties and corners works in a similar manner. The first simulation results show effectiveness of the proposed algorithm. The goals are detected in the 90% of the considered cases. This percentage is quite good and in order to improve the effectiveness of this algorithm also audio loudness information is used. In fact, in this particular case, a sudden increase of audio loudness (due both of the speaker and the crowd) can be a cue that a goal event is happening.

Scene classification by Hidden Markov Model

In this algorithm the focus is placed on providing tools for analyzing both audio and visual streams, for translating the signal samples into sequences of indices. The final objective is to add a recognition task in order to identify statistical patterns of such indices so as to enable a content-based audio-visual description (more at a semantic level) of the whole material (see figure 3).

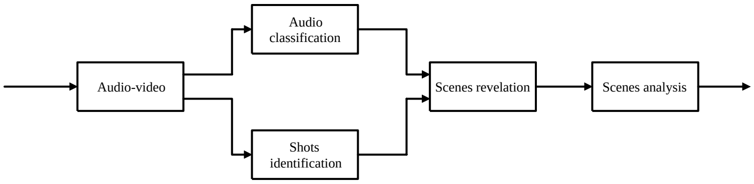

Figure 3: The algorithm using HHM.

First of all, the input stream is demultiplexed into the two components, audio and video. An independent segmentation and classification of the two channels, audio and video, represents the next step of the analysis. On one hand the audio stream is segmented into clips, and a feature vector is extracted from the low-level acoustic properties of each clip. On the other hand in the video analysis channel a feature vector is calculated by comparison of each couple of adjacent frames, in terms of luminance histograms, motion vectors and pixel-to-pixel differences. Each sequence of feature vectors extracted from the two streams is then classified by means of an Hidden Markov Model (HMM), used in an innovative approach: the input signal is considered as a non-stationary stochastic process, modeled by an HMM in which each state stands for a different class of the signal. This defines an adaptive classification scheme for which a set of new training algorithms was developed. Given a sequence of unsupervised feature vectors, the correspondent most likely sequence of indices identifying particular signal classes could be generated using the Viterbi algorithm.

A first approach to scene identification consists in the definition of four different types of scenes: dialogues, in which the audio signal is mostly speech and the change of the associated visual information occurs in an alternated fashion (for example: ABAB...); stories, in which 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 ABCADEFAG...; actions, when 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...; finally consecutive shots which do not belong to any one of the aforementioned scenes, but their associated audio is of a consistent type, are classified as generic scene. Once we have defined these kinds of scenes, we can look for them in the time-aligned sequence of descriptors obtained as mentioned.

A second and more general approach to scene classification is represented by a statistical pattern recognition analysis applying a clustering procedure on the basis of the sequences of descriptors obtained with a long-term analysis on multimedia data. This recognition system is very flexible and does not require defining the types of scenes a priori.

CONCLUSIONS

In this work we have presented some techniques for indexing of multimedia documents, starting from low-level features and using joint audio-video information in order to achieve an high level semantic information, useful for browsing and retrieval in audio-video databases. We have also described an application that uses the results obtained from the previous algorithms in order to allow an easy navigation through a multimedia document and some basic retrieval capabilities.

References

[1] N. Adami, R. Leonardi, ”Identification of editing effects in image sequences by statistical modeling”, Proc. Picture Coding Symposium ’99, pp. 157-160, *Portland, OR, U.S.A., Apr. 1999.*

[2] C. Saraceno, R. Leonardi, ”Identification of story units in audio-visual sequences by joint audio and video processing”, Proc. International Conference on Image Processing 1998, *Chicago, IL, U.S.A., Oct. 1998*

[3] A. Bugatti, R. Leonardi, L.A. Rossi, ”A video indexing approach based on audio classification”, In Proc. International Workshop on Very Low Bitrate Video Coding ’99, pp. 75-78, *Kyoto, Japan, Oct. 1999*.

[4] A. Bugatti, A. Flammini, R. Leonardi, D. Marioli, P. Migliorati, C. Pasin, ”Audio Classification in Speech and Music: A Comparison of Different Approaches”, Proc. Workshop on Image Analysis For Multimedia Interactive Services 2001, pp. 153-158, *Tampere, Finland, May 2001*.

[5] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati, L.A. Rossi, ”The ToCAI Description Scheme for Indexing and Retrieval of Multimedia Documents”, Multimedia Tools and Application Journal, Kluwer Academic Publishers, Vol. 14, No. 2, pp. 153-173, 2001.

[6] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati, L. Rossi, ”Multimedia documents description by ordered hierarchies: the ToCAI description scheme”, Proc. International Conference on Multimedia and Expo 2000, *New York, U.S.A., Aug. 2000*

[7] http://www.extra.research.philips.com/euprojects/avir/.

[8] A. Bonzanini, R. Leonardi, P. Migliorati, ”Semantic Video Indexing Using MPEG Motion Vectors”, Proc. EUSIPCO’2000, pp. 147-150, *Tampere, Finland, Sept. 2000*.

[9] F. Oppini, R. Leonardi, ”Audiovisual Pattern Recognition Using HMM for Content-Based Multimedia Indexing”, Proc. of Packet Video 2000, *Cagliari, Italy, 2000*.