View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Archivio istituzionale della ricerca - Università di Brescia

Top-Down and Bottom-Up Semantic Indexing of Multimedia Documents

R. Leonardi, P. Migliorati

University of Brescia
{leon,pier}@ing.unibs.it

Abstract

The aim of this work consists in proposing a dual approach for the sake of semantic indexing of audio-visual documents. We present two different algorithms based respectively on a bottom-up and a top-down strategy. Considering the top-down approach, we propose an algorithm which implements a finite-state machine and uses low-level motion indices extracted from an MPEG compressed bit-stream. Simulation results show that the proposed method can effectively detect the presence of relevant events in sport programs. Using the bottom-up approach, the indexing is performed by means of Hidden Markov Models (HMM), with an innovative approach: the input signal is considered as a non-stationary stochastic process, modeled by a HMM in which each state is associated with a different property of audio-visual material. Several samples from the MPEG-7 content set have been analyzed using the proposed scheme, demonstrating the performance of the overall approach to provide insights about the content of audio-visual programmes. Moreover, what appears quite attractive instead is to use low-level descriptors in providing a feedback for non-expert users of the content of the described audio-visual programme. The experiments have demonstrated that, by adequate visualization or presentation, low-level features carry instantly semantic information about the programme content, given a certain programme category, which may thus help the viewer to use such low-level information for navigation or retrieval of relevant events.

1 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-video material represents therefore a very important challenge. Ideally 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 [1].

Traditionally, the most common approach to create an index of an audio-visual document has been based on the automatic detection of the changes of camera records and the types of involved editing effects. This kind of approach has generally demonstrated satisfactory performance and lead to a good low-level temporal characterization of the visual content. However the reached semantic characterization remains poor since the description is very fragmented considering the high number of shot transitions occurring in typical audio-visual programs.

Alternatively, there have been recent research efforts to base the analysis of audio-visual documents on a joint audio and video processing so as to provide for a higher level organization of information [2] [3]. In [3] these two sources of information have been jointly considered for the identification of simple scenes that compose an audio-visual program. The video analysis associated to cross-modal procedures can be very computationally intensive (by relying, for example, on identifying correlation between non-consecutive shots).

In this paper we propose and compare the performance of two different approaches for semantic indexing of audio-visual documents. In the first approach, the problem of identification of relevant situations in audio-visual programmes within specific contexts (e.g., sport) has been taken into account. In this context we have faced the problem of semantic video indexing using associated motion information. The idea has been to establish a correlation between semantic features and motion indices associated to a video sequence, by establishing a sequence of transitions between camera motion characteristics associated to series of consecutive shots, or portions of shots. For soccer video sequences, the semantic content can be identified with the presence of interesting events such as, for example, goals, shots towards goal, and so on; these events can be found at the beginning or at the end of game action. A good semantic index of a soccer video sequence can therefore be established by a summary made up of a list of all game actions, each characterized by its beginning and ending event [4].

In the second part of the work, the attention has been focused to suitably combine, using a complementary bottom-up approach, audio and visual descriptors and associated individual shots/audio segments to extract higher level semantic entities: scenes or even individual programme items [5]. In particular, the indexing is performed by means of Hidden Markov Models (HMM), used in an innovative approach: the input signal is considered as a non-stationary stochastic process, modeled by a HMM in which each state stands for a different class of the signal [6]. From the obtained results, it appears that a good characterization can be achieved also by providing a direct feedback of low-level audio and visual descriptors to the to viewer, rather than trying to automatically define the associated semantics of the information.

The paper is organized as follows. In Sections 2 and 3 the proposed algorithms are presented, whereas the simulation results are discussed in Section 4. Conclusions are drawn in Section 5.

2 A top-down approach for Soccer video indexing using motion information

As previously mentioned, the problem of semantic video indexing is of great interest for its usefulness in the field of efficient navigation and retrieval from multimedia databases (description and retrieval of images or shots, ...). This task, which seems very simple for the human being, is not trivial to implement in an automatic manner.

After the extraction of some low-level features, a decision making algorithm is needed for semantic indexing, and this algorithm can be implemented in two ways: modeling the human cognitive behaviour; using statistical approaches to relate features of some data descriptor with outcomes of the human decision process.

Our work is based on the latter approach; namely, we are attempting to make a semantic indexing of a video sequence starting from some descriptors of the motion field extracted from it. In particular, we have chosen three low-level motion indices and we have studied the correlation between these indices and the semantic events defined above [4].

Low-level motion descriptors

Motion data related to a video sequence are typically the motion vectors associated at each image. In our work, the motion information is directly extracted from the compressed MPEG2 domain, where each frame is divided into MacroBlocks (MB), and for each MB is given: the MB type; a motion vector obtained with a block-matching algorithm with respect to a reference frame. The MB type is strictly related to motion vectors as follows: if a MB is Intra-Coded, no motion vector is given; if a MB is No-Motion-Coded, the motion vector is null; otherwise a not-null motion vector is given. The motion field can be represented with various type of descriptors. These compact representations should anyway be combined with other indices suitable to state their reliability or to add other useful information.

In our work we have considered three low-level indices which represent the following characteristics: (i) lack of motion, (ii) camera operations (represented by pan and zoom parameters) and (iii) the presence of shot-cuts.

Lack of motion has been evaluated by thresholding the mean value of the motion vector module. Camera motion parameters, represented by horizontal "pan" and "zoom" factors, have been estimated using a least-mean square method applied to P-frame motion fields. We have thus detected fast horizontal pan (or fast zoom) by thresholding the pan value (or the zoom

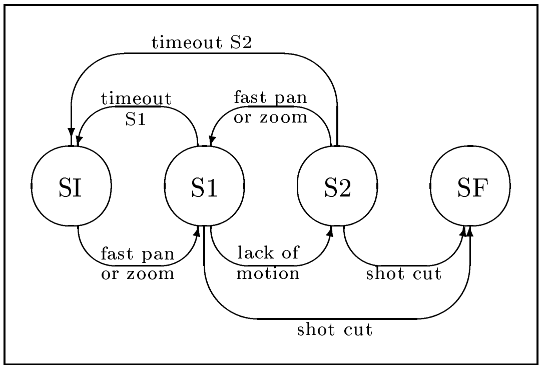

Figure 1: The proposed goal-finding algorithm.

factor) [4]. In our implementation, shot-cuts have been detected using only motion information too. In particular, we have used the sharp variation of the above mentioned motion indices and of the number of Intra-Coded MB of P-frames [7]. To evaluate the sharp variation of the motion field we have used the difference between the average value of the motion vector modules between two adjacent P-frames. This parameter will assume significantly high values in presence of a shot-cut characterized by an abrupt change in the motion vector field between the two considered shots. This information regarding the sharp change in the motion vector field has been suitably combined with the number of Intra-Coded MB of the current P-frames. When this parameter is greater than a prefixed threshold value, we assume that there is a shot-cut [4].

The proposed goal-finding algorithm

As it can be seen, the above mentioned low-level indices are not sufficient, individually, to reach satisfactory results. To find particular events, such as, for example, goals or shots toward goal, we have tried to exploit the temporal evolution of motion indices in correspondence with such events. We have noticed that in correspondence with goals we can find fast pan or zoom followed by lack of motion, followed by a shot cut. The concatenation of these low-level events have therefore been identified and exploited using the finite-state machine shown in Fig. 1.

From the initial state SI, the machine goes into state S1 if fast pan or fast zoom is detected for at least 3 consecutive P-frames. Then, from state S1 the machine goes into the final state SF, where a goal is decleared, if a shot-cut is present; from state S1 it goes into state S2 if lack of motion is detected for at least 3 consecutive P-frames. From state S2 the machine goes into final state SF if a shot-cut is detected, while it returns into state S1 if fast pan or zoom is detected for at least 3 consecutive P-frames (in this case the game action is probably still going on). Two "timeouts" are used to return into the initial state SI from states S1 and S2 in case nothing is happening for a certain number of P-frames (corresponding to about 1 minute of sequence).

3 A bottom-up approach: Content Based Indexing using HMM

In the second part of this presentation the focus has been placed on providing tools for analyzing both audio and visual streams, for translating the signal samples into sequences of indices.

The whole processing system can be seen as composed of different steps of analysis, each of which extracts information at a defined level of semantic abstraction. First of all, the input stream is demultiplexed into the two main components, audio and video. An independent segmentation and classification of the two channels, audio and video, represent 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, for 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 a Hidden Markov Model (HMM) [8], in an innovative approach: the input signal is considered as a non-stationary stochastic process, modelled by a HMM in which each state stands for a different class of the signal. Given a sequence of unsupervised feature vectors, the correspondent most likely sequence of indices identifying particular signal classes can be generated using the Viterbi algorithm [6].

For audio classification we consider four classes, namely: music, silence, speech and background noise. The result of the audio classification consists in the association of one of this classes to each previously extracted feature vector. In other words, at the end of this analysis we get a segmentation of the audio signal into these four classes with a temporal resolution of 0.5 seconds (the resolution is given by the shift in time of each clip with respect to the previous one). The first step of the analysis is the segmentation of the audio file in equal length frames, partially overlapped in order to reduce the spectral distortion due to the windowing. The duration of each frame is N samples (typically N is set in order to have time duration of 30-40 ms), and each frame is overlapped for 2/3 of its duration with the next frame. In this way it is possible to associate a label (music, silence, speech or background noise) to each frame and these labels can be changed every N/3 samples. After the segmentation step, an audio class descriptor can be associated to each frame using an ergodic HMM.

In the video analysis we consider the problem of segmenting a video signal into elementary units, i.e., video shots; with this aim, we train a two-state HMM classifier, in which the state "1" is associated a "no detected transition" and the other state is associated a "detected shot transition". In this way the system is able to identify both abrupt transition (cut) and slow transition (fading) between consecutive shots. Again, these are obtained by applying the Viterbi algorithm to the sequence of feature vectors: when the second state is reached the system recognizes a shot transition. The same idea is used to establish a correlation between non-adjacent shots, in order to identify subsequent occurrence of a same visual content between non-consecutive shots.

The task of the following step of the analysis focuses on extracting content from segmented video shots, and then indexing them according to the initial audio and video classification. First the results of the previous audio and video classifications must be time-aligned, obtaining a joint list of indices with audio and video classification information associated to each one. We introduce another semantic entity called "scene", which is composed by a group of consecutive shots. Scenes represent a level of semantic abstraction in which audio and visual signals are jointly considered in order to reach some meaningful information about the associated stream of data.

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 (e.g., 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 non 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 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 previously 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 sequence of descriptors obtained with a long-term analysis on multimedia data. This way we examine the linear separability of different scene classes evaluating two indices of disorder and the associated scattering matrices. This recognition system is very flexible and does not require to define a priori the type of scene.

4 Experimental results

In this section the simulation results obtained using the proposed methods are described and discussed.

Top-down approach: Indexing of soccer games sequences

The performance of the proposed goal-finding algorithm have been tested on 2 hours of MPEG2 sequences containing the semantic events reported in Table 1. Almost all live goals are detected, and the algorithm is able to detect some shots to goal too, while it gives poor results on free kicks. The number of false detection is quite relevant, but we have to take into account that these results are obtained using motion information only, so these false detection will probably be eliminated using other type of media information (e.g., audio information).

Events	**Present**			**Detected**		
	Live	Replay	Total	Live	Replay	Total
Goals	20	14	34	18	7	25
Shots to goal	21	12	33	8	3	11
Penalties	6	1	7	1	0	1
Total	**47**	**27**	**74**	**27**	**10**	**37**
False						**116**

Table 1: Performance of the proposed goal-finding algorithm.

Bottom-up approach: Content Based Indexing using HMM

Audio classification results

This analysis have been based on 60 minutes of audio, on which we have compared the supervised classification and the results obtained using the HMM. Each simulation has been carried on audio segment with duration of 3'.

Table 2 shows the classification percentages of each class (music, silence, speech, and noise): the values in the main diagonal are the percentages of correct recognition.

	Rec. Music	Rec. Silence	Rec. Speech	Rec. Noise	Recall	Prec.
Music	80.5	3.6	11.4	4.5	0.805	0.843
Silence	2.4	95.8	1.8	0	0.958	0.867
Speech	11.3	2.3	85.4	1	0.854	0.856
Noise	3.1	2.9	2.5	91.5	0.915	0.928

Table 2: Recognition percentages, and Recall and Precision.

The classifier performance are summarized using a couple of indices evaluated for each class. These indices are called precision and recall, and are defined as follows: $precision = \frac{N_c}{N_c+N_f}, \quad recall = \frac{N_c}{N_c+N_m}$

where N_c is the number of correct detection, N_m is the number of missed detection and N_f is the number of wrong detection. Both these indices are in the range [0,1]. The various performance indices have been evaluated and the results are shown in Table 2.

It is clear from Table 2 that with noise and silence the algorithm show the best performance, while results in music and speech detection are poorer. This result is due to the high level of misclassification between music and voice. These errors can be due to the following problems: i) the number of data in the training set is too low, ii) the used audio features can be insufficient in order to have a correct classification, iii) limits in HMM approach.

Video classification results

The performance analysis of the video classificator is based on a video stream with duration of 60 minutes. In the same way of audio analysis, each simulation has been carried out on video segment with duration of 3' and then all the results have been combined to obtain the whole classification The video classification is composed from two steps: shot segmentation by means of transition recognition, and then correlation between-shots in order to obtain the correlation among non contiguous shots. While the initial segmentation has made using the HMM classificator, the between-shots analysis uses only the probability density functions of such model, where vector elements of initial state and of transition matrix are set to 1/N.

The results of both analysis are shown in Table 3, where state 1 stands for "no detected transition" and state 0 stands for "detected shot transition".

	Rec. State 0	Rec. state 1
State 0	95	5
State 1	1.1	98.9

Table 3: Results in video shot classification.

We have two possible errors: state 1 recognized as 0 and state 0 recognized as 1. As Table 3 indicates, it is more probable that the system reveals false shot change rather than it does not detect an existing ones. The reduced value of precision is due to a relatively high number of false shot detections. These false detections are due to very fast camera motion, luminance changes and motion of big objects in the scene.

Scene identification results

Using the results of audio and video classification, it is possible to evaluate the performance of the scenes identification system by means of segmentation rules, as described in the previous chapters. The first step is to align audio and video descriptors, creating a descriptor shot sequence. Using this sequence it is possible to search the scene categories already defined (dialogs, actions, stories e generic scenes). For each kind of scene, four different performance indices have been calculated: recall, precision, $recall_{cover}$ e $precision_{cover}$. The first two are defined as in the case of the audio and video classification simulations, while the latter two have been introduced to consider situations when some shots are recognized correctly and other are recognized improperly. A scene is considered to low correctly recognized if at least one of its shots is correctly identified (i.e., it belongs to the right kind of scene). Moreover, when two consecutive scenes are recognized as a unique scene, only one is considered correct (if it has been correct classified) while the other is declared missed. The second couple of indices has been introduced because sometimes the identified scene is only partially overlapping with the real scene (some shots belonging to the identified scene do not belong to the real scene). Let N_c be the number of shots belonging to one kind of correctly identified scene, N_a the number of shot belonging to this kind of scene, and N_r the whole number of shots belonging to the identified scenes of this kind of scene, then we define $precision_{cover} = \frac{N_c}{N_r}$, $recall_{cover} = \frac{N_c}{N_a}$

Table 4 shows the results for each type of scene and for each index.

The scene identification procedure is based on deterministic rules rather than on a stochastic classifier, and it provides limited results when it is used in order to understand the audio-visual assembly modality that the director used to create the scenes. The errors can be due to inaccuracies in a priori rules, and obviously to errors in previous steps of classification.

Figures 2,3,4 show the results of the classifications of three TV programmes representing respectively 3 minutes of a talk-show, of a music program, and of a scientific documentary. As we can see, the different statistical evolution of audio and video indices can be effectively used to infer the semantic content of the analyzed signals.

From the conducted studies, it can therefore be concluded that semantic characterization of the content can rarely be achieved at the highest level, if not with a top-down approach,

	Recall	Precision	Recall-cover	Precision-cover
Dialog	0.893	0.791	0.810	0.673
Action	0.884	0.801	0.789	0.743
Story	0.790	0.755	0.759	0.693
Generic scene	0.745	0.715	0.690	0.668

Table 4: Values for each index and for each kind of scene evaluated using the results of the previous steps of audio and video classification.

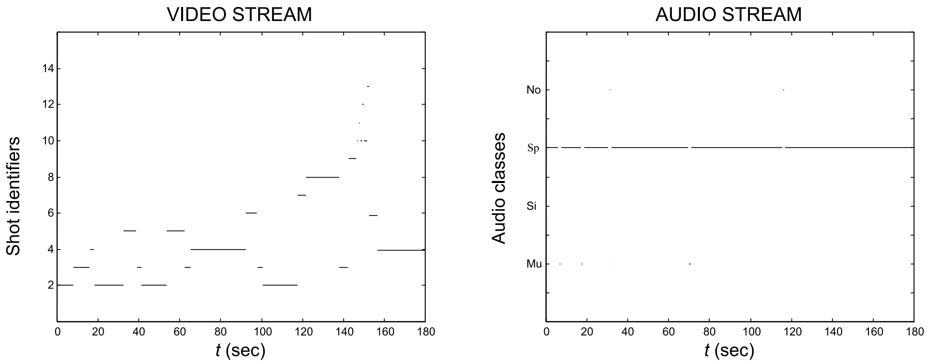

Figure 2: Audio and video classification results of 3 minute of a talkshow. In the video stream diagram, the shot label is associated to the visual content of the relative shot, in such a way that shots with similar visual content are denoted by the same label.

having predefined in a specific application context the high-level semantic instances of events of interest (e.g., goals in a soccer game). Only intermediary semantic characterization is obtainable otherwise, to identify macro-scenes defining dialogue, story or action situations.

What appears quite attractive instead is to use low-level descriptors in providing a feedback for non-expert users of the content of the described audio-visual programme. The experiments have demonstrated that, by adequate visualization or presentation, low-level features carry instantly semantic information about the programme content, given a certain programme category, which may thus help the viewer to use such low-level information for navigation or retrieval of relevant events.

5 Conclusion

In this paper we presented two different indexing algorithms based respectively on bottom-up and top-down approaches. Regarding the top-down approach, we propose a semantic indexing algorithm based on finite-state machines and low-level motion indices extracted from the MPEG compressed bit-stream. Considering the bottom-up approach, the signal classification is performed by means of Hidden Markov Models (HMM). Several samples from the MPEG-7 content set have been analyzed using the proposed classification schemes, demonstrating the performance of the overall approach to provide insights of the content of the audio-visual material. The classification results are quite satisfactory, and it appears that a joint visualization of audio and visual properties provide reach insights on the audio-visual programme content.

References

[1] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati, L. Rossi, "Describing Multimedia Documents in Natural and Semantic-Driven Ordered Hierarchies", Proc. ICASSP'2000, pp. 2023-2026, Istanbul, Turkey, 5-9 June 2000.

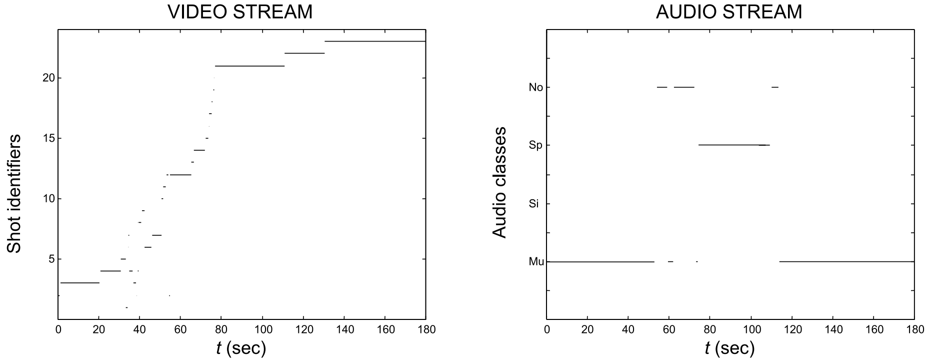

Figure 3: Audio and video classification results of 3 minute of a music program. In the video stream diagram, the shot label is associated to the visual content of the relative shot, in such a way that shots with similar visual content are denoted by the same label.

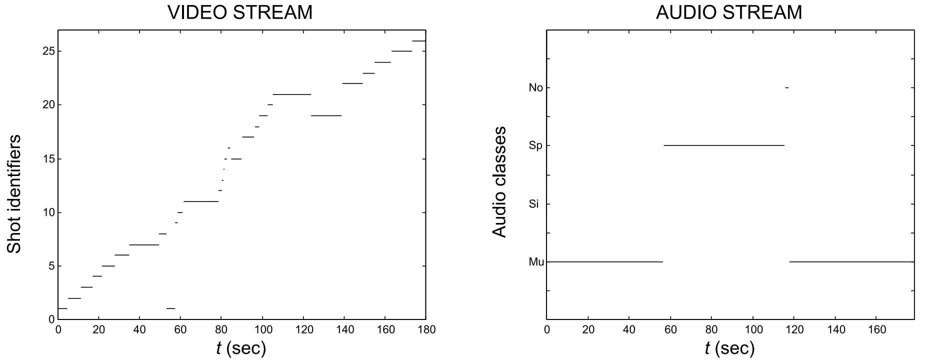

Figure 4: Audio and video classification results of 3 minute of a scientific documentary. In the video stream diagram, the shot label is associated to the visual content of the relative shot, in such a way that shots with similar visual content are denoted by the same label.

[2] Yao Wang, Zhu Liu, Jin-cheng Huang, "Multimedia Content Analysis Using Audio and Visual Information", To appear in Signal Processing Magazine.

[3] C. Saraceno, R. Leonardi, "Indexing Audio-Visual Databases Through a Joint Audio and Video Processing", International Journal of Imaging Systems and Technology, Vol. 9, No. 5, pp. 320-331, Oct. 1998.

[4] A. Bonzanini, R. Leonardi, P. Migliorati, "Semantic Video Indexing Using MPEG Motion Vectors", Proc. EUSIPCO'2000, pp. 147-150, 4-8 Sept. 2000, Tampere, Finland.

[5] R. Zhao, W.I. Grosky, "From Features to Semantics: Some preliminary Results", Proc. of IEEE International Conference ICME2000, New York, NY, USA, 30 July - 2 August 2000.

[6] F. Oppini, R. Leonardi, "Audiovisual Pattern Recognition Using HMM for Content-based Multimedia Indexing", Proc. of Packet Video 2000, Cagliari, Italy.

[7] Yining Deng, B. S. Manjunath, "Content-Based Search of Video Using Color, Texture, and Motion", Proc. of IEEE International Conference ICIP-97, Santa Barbara, California, USA, pp. 534-536, October 26-29, 1997.

[8] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", IEEE Proc., vol. 77, no. 2, pp. 257-286, Feb. 1989.