SEMANTIC INDEXING OF SPORTS PROGRAM SEQUENCES BY AUDIO-VISUAL ANALYSIS

R. Leonardi, P. Migliorati

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

M. Prandini

DEI - Politecnico di Milano, Italy
prandini@elet.polimi.it

ABSTRACT

Semantic indexing of sports videos is a subject of great interest to researchers working on multimedia content characterization. Sports programs appeal to large audiences and their efficient distribution over various networks should contribute to widespread usage of multimedia services. In this paper, we propose a semantic indexing algorithm for soccer programs which uses both audio and visual information for content characterization. The video signal is processed first by extracting low-level visual descriptors from the MPEG compressed bit-stream. The temporal evolution of these descriptors during a semantic event is supposed to be governed by a controlled Markov chain. This allows to determine a list of those video segments where a semantic event of interest is likely to be found, based on the maximum likelihood criterion. The audio information is then used to refine the results of the video classification procedure by ranking the candidate video segments in the list so that the segments associated to the event of interest appear in the very first positions of the ordered list. The proposed method is applied to goal detection. Experimental results show the effectiveness of the proposed cross-modal approach.

1. INTRODUCTION

The efficient distribution of sports videos over various networks should contribute to the rapid adoption and widespread usage of multimedia services, because sports video appeal to large audiences. The design of efficient automatic techniques suitable to semantically characterize sports video documents is therefore necessary and very important. Compared to other videos such as news and movies, sports videos have well defined content structure and domain rules. A long sports game is often divided into a few segments. Each segment in turn contains some sub-segments. For example, in American football, a game contains two halves, and each half has two quarters. Within each quarter there are many plays, and each play starts with a formation in which players line up on two sides of the ball. A tennis game is divided into sets, then games and serves. In addition, in sports video, there are a fixed number of cameras in the field that result in unique scenes during each segment. In tennis, when a serve starts, the scene is usually switched to the court view. In baseball, each pitch usually starts with a pitching view taken by the camera behind the pitcher [1].

In soccer video, automatic content extraction methods initially focused on shot classification [2] and scene reconstruction [3]. More recently the problems of segmentation and structure analysis have been considered in [4], whereas the automatic extraction of highlights and summaries have been analyzed in [5], [6], [7], [8], [9]. In [8], a higher level approach was proposed. Specifically, a method that tries to detect the complete set of semantic events which may happen in a soccer game is presented. This method uses the position information of the player and of the ball during the game as input, and therefore relies on a quite complex and accurate tracking system to obtain this information.

In [5] and [6] the correlation between low-level descriptors and the semantic events in a soccer game has been investigated. In particular, in [5], it is shown why low-level descriptors are not sufficient, individually, to obtain satisfactory results. In [6] it was instead demonstrated how to exploit the temporal evolution of the low-level descriptors in correspondence to semantic events. The proposed algorithm relies on the construction of a deterministic finite-state machine. This algorithm gives good results in terms of accuracy in the detection of the goals, though the number of false detections remains quite high.

In order to reduce the number of false detections, a video indexing algorithm based on statistical modelling was presented in [7]. It uses controlled Markov chains to model the statistical temporal evolution of low-level visual descriptors. The introduction of the statistical modelling improves the performance of the classification algorithm with respect to [6], even if there are still some false detections. Considering that the previous results were obtained using visual information only, it was decided to evaluate if a multi-modal approach would have improved the performance reducing further the number of false detections.

In literature, many different low-level audio features for video content characterization have been proposed and discussed. The reader is referred to [10] for an overview. To give some examples related to the specific case of sports programs, in [12] the focus has been on excited/non excited commentary classification, whereas audio event spotting (such as baseball hits or football touchdown) was proposed in [11], [12]. More recently, in [13], a multi-modal approach for content characterization of baseball sequences was adopted where audio is used jointly with video and text. In the case of soccer games audio-visual sequences, the sound track is composed mainly of foreground commentary coexisting with background sounds. The background sounds include ambient crowd noise, sparse happenings of excited segments of crowd noise, and special events such as whistles or clapping. The audio signal is therefore more complex to analyze than in other sports programs, and this possibly justifies the fact that there are very few significant examples where audio is used for content characterization of soccer audio-visual sequences.

In this paper, we focus in particular on the detection of goals,

This research has been partially supported by the IST programme of the EU under projects IST-2001-32795 SCHEMA, and IST-2000-28304 SPATION.

which represent the key event in a soccer game. We use a simple audio descriptor (the "loudness" of the audio signal) to order the video segments identified by the analysis carried out on the video signal according to the algorithm we proposed in [7]. In this way, the segments associated to goals appear in the very first positions of the ordered list, and the number of false detections can be greatly reduced. The proposed cross-modal approach could be applied, in principle, to address the problem of detecting further semantic events of interest in a soccer game, following the same line of reasoning.

The rest of the paper is organized as follows. Section 2 describes the low-level features on which the goal detection algorithm is based. The goal detection algorithm is presented in Section 3. Experimental results showing the effectiveness of the proposed approach are reported in Section 4. Finally, concluding remarks are drawn in Section 5.

2. LOW-LEVEL DESCRIPTORS

In this section we describe the low-level visual and audio descriptors and the shot-cut detection method used in the proposed goal detection algorithm.

2.1. Visual descriptors

We consider the visual descriptors 'Lack of motion', 'Fast pan', and 'Fast zoom', which were originally proposed in [6]. These descriptors summarize the information on (i) lack of motion, and (ii) camera operations (panoramic and zoom views) contained in the video sequence, which are actually relevant to semantic indexing of soccer videos. Lack of motion usually occurs at the beginning or at the end of game actions, when no interesting action is taking place. Fast panoramic views occur in case of shots towards the goal-keeper or fast ball exchanges between distant players, whereas fast zoom views are used when interesting situations are likely to occur according to the perception of the camera operator. Each descriptor represents a binary variable taking values in the set $\{0, 1\}$, and is evaluated on each P-frame based on the motion vector field which is directly available in the MPEG compressed bit-stream. The descriptor 'Lack of motion' is evaluated by thresholding the mean value of the motion vectors module, and is set to 0 when the threshold is exceeded. The threshold value is set equal to 4. Camera motion parameters, represented by horizontal 'pan' and 'zoom' factors, are evaluated using a least-mean squares method applied to the P-frame motion field. The value of the descriptor 'Fast pan' ('Fast zoom') is then obtained by thresholding the pan factor (zoom factor). In this case, the descriptors are set to 1 when the threshold is exceeded. The threshold value is selected to be equal to 20 for the 'Fast pan' descriptor and to 0.002 for the 'Fast zoom' descriptor.

2.2. Shot-cut

The problem of shot-cut detection for video segmentation has been given a lot of attention in literature, and the reader is referred to [14] for an overview. In our implementation, shot-cuts are detected using low-level visual descriptors extracted from the MPEG compressed bit-stream [6]. Specifically, to detect if there is a shot-cut between two consecutive P-frames, we use the difference between the mean value of the motion vectors modules associated to the two P-frames. This parameter is likely to exhibit a high value in presence of a shot-cut characterized by an abrupt change in the motion field between the two considered P-frames. The information provided by this difference is suitably combined with the number of Intra-Coded Macro-Blocks of the P-frame. When the obtained value is greater than a prefixed threshold, we assume that a shot-cut has occurred [6].

2.3. Audio descriptor

When analysing the audio signal, the sound track is typically segmented in partially overlapping 'audio-frames'. In case of sampling frequency set to 44.1 kHz, each one is usually composed of 1024 consecutive audio samples, with 512 samples in common with the previous audio-frame [10]. The 'loudness' of frame k is given by

$$l(k) = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x_k^2(n)}$$

where $\{x_k(n), n = 0, \ldots, N-1\}$ is the set of N audio samples of frame k, with $N = 1024$ [10]. The elementary unit of audio signal for statistical analysis is a 'clip' which is a set of consecutive audio samples corresponding to 1.5 seconds. This value is selected as it is often reasonable to consider the audio signal to have quasi-stationary behavior in such time interval [10]. Each clip is composed of a number of frames that depends on the sampling frequency. For example, if the sampling frequency is 44.1 kHz, then each clip corresponds to 66150 samples, hence it corresponds to 128 frames. The evolution of the frame loudness within an audio clip follows the evolution in time of the amplitude of the audio signal. Therefore, it constitutes a fundamental parameter for audio signal characterization. This is especially true in our context where the final objective of audio analysis is goal detection, since the occurrence of a goal causes the commentator to increase the loudness of his/her voice and the crowd to clap or whistle.

By estimating the mean value of the audio loudness for every clip, we obtain a low-level descriptor of the audio signal which we call 'clip loudness'. This descriptor exhibit a peculiar behavior in correspondence of two consecutive shots containing a goal event. It takes significantly higher values in the second of the two shots, whereas this is generally not the case in no-goal shot pairs. Figure 1 represents a typical plot of the clip loudness behavior in shot pairs associated to a goal and to a generic no-goal situation.

3. THE PROPOSED METHOD FOR GOAL DETECTION

In this section, we describe the proposed algorithm for goal detection. The algorithm consists of a two-step procedure. The video signal is processed first so as to produce a list of candidate video segments. These are then ordered in the second step based on the audio loudness transition between the consecutive candidate shot pairs.

3.1. Processing of the video signal

In this section, we briefly describe the adopted controlled Markov chain model (CMC). The reader is referred to [7] for more details. The components of a controlled Markov chain model are the state and input variables, the initial state probability distribution, and the controlled transition probability function. We suppose that the occurrence of a shot-cut event causes the system to change dynamics.

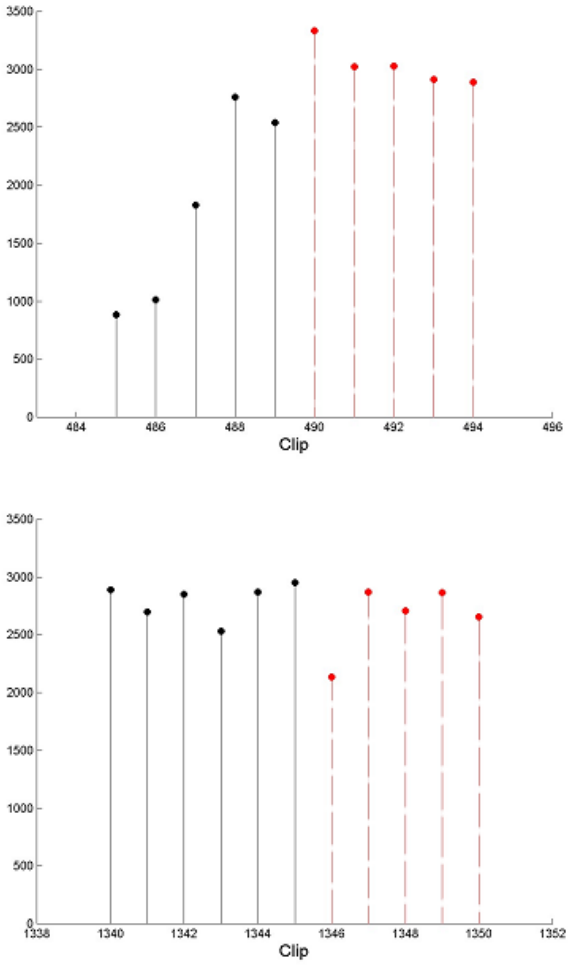

Fig. 1. Typical plots of the clip loudness in a shot pair associated to a goal (above), and to a generic no-goal situation (below). The dashed lines correspond to the second shot.

In order to model this fact, we describe the state of the system as a two-component state, with one component representing the operating mode and the shot-cut being the input that causes the system to change mode [7]. A schematic representation of the introduced model is given in Figure 2. In this figure, the symbol "$\sim$" is used for "distributed according to". We suppose that a goal event takes place over a two-shot block and that the evolution of the low-level visual descriptors during these shots can be modelled by a controlled Markov chain with the structure described in [7]. In the shot pairs when no-goal occurs, the low-level visual descriptors evolve according to a different controlled Markov chain model. In particular, we introduce 5 further models to capture different no-goal situations where either an interesting event (such as a corner kick or a free kick) occurs or no relevant event takes place.
On the basis of the derived six controlled Markov chain models, one can classify each pair of shots in a soccer game video sequence as the most likely event by using the maximum likelihood criterion. For each pair of consecutive shots (i.e., two consecutive sets of P-frames separated by shot-cut), one needs to i) extract

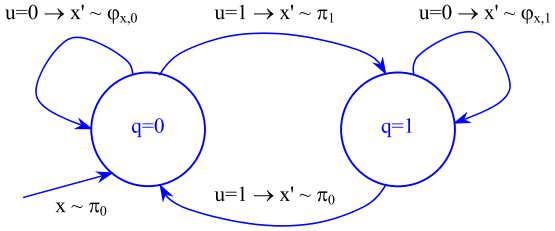

Fig. 2. Controlled Markov chain model.

the sequence of low-level descriptors, ii) determine the sequence of values the state and input variables take, and iii) determine the likelihood of such sequence of values according to each one of the six admissible models. The model that maximizes the likelihood function is then associated to the considered pair of shots.

3.2. Processing of the audio signal

The audio processing step takes as input the candidate shot pairs for goal detection identified by the video processing algorithm, and, based on the audio descriptor value produces an ordered list of the shot pairs so that a goal will appear in the first positions. As discussed in the introduction, the audio track of a soccer program is difficult to analyse because of the complex set of audio sources which are combined. For this reason, a simple criterion is used for ranking the candidate shot pairs, which relies on the observation that, typically, the occurrence of a goal causes the audio signal to increase its loudness. Specifically, for any candidate shot pair, the average value of the low-level audio descriptor 'clip loudness' (cf. Section 2.3) on each shot is computed. The larger is the increase of such average clip loudness between the two shots forming a candidate pair, the higher is the position of that shot pair in the list. A block diagram of the overall goal detection algorithm is represented in Figure 3.

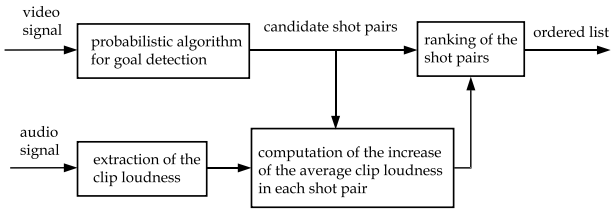

Fig. 3. Block diagram of the goal detection algorithm.

4. EXPERIMENTAL RESULTS

Experiments were run to test the algorithm for goal detection described in Section 3. We considered a few MPEG2 soccer audio-visual sequences and divided them into two sets of sequences: one (the training set) of about 10 hours for training the controlled Markov chain (CMC) models adopted in the video processing step of the algorithm, and the other (test set) of 4.5 hours for evaluating the performance of the algorithm. More specifically, all shot pairs of the training sequences were manually classified in goal and no-goal pairs, with the no-goal pairs distinguished in different categories (e.g., corner kick, free kick, plain action, etc.). Then, the

controlled transition probability functions and initial state probability distributions characterizing the different goal and no-goal CMC models were estimated based on the sequence of values taken by the state and input variables in the corresponding set of shot pairs. The goal detection algorithm was then used on the test set composed of the audio-visual sequences of the two plays of three soccer games. The sequences contain 13 goals and more than 2000 shot-cuts. The obtained results are summarized in Table 1.

Table 1. Performance of the proposed cross-modal analysis method (LIV-LIP:Liverpool-Lipsia; SPA-SLO:Spain-Slovenia; SPA-PAR:Spain-Paraguay) .

	n. of goals	n. of goals det.	n. of shots	n. of cand. shot pairs	goals pos. in ord. list
LIV-LIP 1	2	2	333	43	4, 5
LIV-LIP 2	3	3	355	62	18, 21, 22
SPA-SLO 1	1	1	385	56	3
SPA-SLO 2	3	3	358	65	2 ,18, 23
SPA-PAR 1	1	1	374	58	11
SPA-PAR 2	3	3	435	77	1, 12, 19

Note that all the shot pairs where a goal actually occurs are within the first 23 positions in the ordered list, irrespectively of the considered sequence. This means that if we take the first 23 shot pairs of the ordered lists, we are able to detect all the goals with a reduced number of false detections. In Table 2 we report the results obtained by ranking all shot pairs without first using the video-based selection procedure. Note that the shot pairs where a goal occurs typically drop down in the ordered list, thus confirming the importance of using the motion information contained in the video signal as indicated.

Table 2. Performance of the algorithm where all shot pairs are ordered based on the audio signal.

	number of goals	number of shots	goals position in the ordered list
LIV-LIP 1	2	333	20, 41
LIV-LIP 2	3	355	112, 121, 123
SPA-SLO 1	1	385	18
SPA-SLO 2	3	358	9, 254 , 88
SPA-PARA 1	1	374	62
SPA-PARA 2	3	435	1, 72, 104

5. CONCLUSIONS

In this paper we have presented a semantic video indexing algorithm based on Controlled Markov Chain models that exploits the temporal evolution of low-level visual descriptors extracted from a MPEG-2 compressed bit-stream representing soccer video programmes. To reduce the number of false detections given by an initial video processing unit, a second stage has been added to incorporate audio properties. In particular we have evaluated the "loudness" associated to each video segment identified by the analysis carried out on the video signal. The intensity of the "loudness" has been used to order the candidate video segments. Consequently, the segments associated to interesting events appear in the very first positions of the ordered list, and the number of false detections can be greatly reduced. Further work needs to be done to better exploit other more sophisticated properties of the audio signal, in order to further reduce the number of false detections. Preliminary experiments have however demonstrated that a joint characterization of audio-visual descriptors with a controlled Markov chain model is inadequate, demonstrating the need to have separate processing for this two information sources as suggested in this paper.

6. REFERENCES

[1] D. Zhong, S.F. Chang, "Structure Analysis of Sports Video Using Domain Models", Proc. ICME'2001, pp. 920-923, Aug. 2001, Tokyo, Japan.

[2] Y. Gong, L.T. Sin, C.H. Chuan, H. Zhang, M. Sakauchi, "Automatic parsing of TV soccer programs", Proc. ICMCS'95, May 1995, Washington DC, USA.

[3] D. You, B.L. Yeo, M. Yeung, G. Liu, "Analysis and presentation of soccer highlights from digital video", Proc. ACCV 95, Dec. 1995, Singapore.

[4] L. Xie, S.F. Chang, A. Divakaran, H. Sun, "Structure Analysis of Soccer Video with Hidden Markov Models", Proc. ICASSP'2002, May 2002, Orlando, FL, USA.

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

[6] A. Bonzanini, R. Leonardi, P. Migliorati, "Event Recognition in Sport Programs Using Low-Level Motion Indices", Proc. ICME'2001, pp. 920-923, Aug. 2001, Tokyo, Japan.

[7] R. Leonardi, P. Migliorati, M. Prandini, "Modeling of Visual Features by Markov Chains for Sport Content Characterization", Proc. EUSIPCO'2002, pp. 349-352, Sept. 2002, Toulouse, France.

[8] V. Tovinkere, R. J. Qian, "Detecting Semantic Events in Soccer Games: Toward a Complete Solution", Proc. ICME'2001, pp. 1040-1043, Aug. 2001, Tokyo, Japan.

[9] A. Ekin, M. Tekalp, "Automatic Soccer Video Analysis and Summarization", Proc. SST SPIE03, Jan. 2003, CA, USA.

[10] Y. Wang, Z. Liu, J.C. Huang, "Multimedia Content Analysis Using Audio and Visual Information", IEEE Signal Processing Magazine, vol. 17, no. 6, pp. 12-36, Nov. 2000.

[11] Y. Chang, W. Zeng, I. Kamel, R. Alonso, "Integrated image and speech analysis for content-based video indexing", Proc. 3rd IEEE Int. Conf. Multimedia Computing and Systems, pp. 306-313, June 1996, Hiroshima, Japan.

[12] Y. Rui, A. Gupta, A. Acero, "Automatically extracting highlights for TV Baseball programs", Proc. ACM Multimedia 2002, pp. 105-115, Oct. 2000, Los Angeles, CA, USA.

[13] M. Han, W. Hua, W. Xu, Y. Gong,"An integrated Baseball Digest System Using Maximum Entropy Method", Proc. ACM Multimedia 2002, Dec. 2002, Juan Les Pins, France.

[14] I. Koprinska, S. Carrato, "Temporal Video Segmentation: a survey", Signal Processing: Image Communication, Vol. 16, No. 5, pp. 477-500, Jan. 2001.