An overview of multi-modal techniques for the characterization of sport programmes

N. Adami, R. Leonardi, P. Migliorati

DEA University of Brescia, Via Branze 38, Brescia, Italy

ABSTRACT

The problem of content characterization of sports videos is of great interest because sports video appeals to large audiences and its efficient distribution over various networks should contribute to widespread usage of multimedia services. In this paper we analyze several techniques proposed in literature for content characterization of sports videos. We focus this analysis on the typology of the signal (audio, video, text captions, ...) from which the low-level features are extracted. First we consider the techniques based on visual information, then the methods based on audio information, and finally the algorithms based on audio-visual cues, used in a multi-modal fashion. This analysis shows that each type of signal carries some peculiar information, and the multi-modal approach can fully exploit the multimedia information associated to the sports video. Moreover, we observe that the characterization is performed either considering what happens in a specific time segment, observing therefore the features in a "static" way, or trying to capture their "dynamic" evolution in time. The effectiveness of each approach depends mainly on the kind of sports it relates to, and the type of highlights we are focusing on.

Keywords: Sports video content characterization, semantic indexing, multi-modal analysis, audio-visual features

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 valuable semantics in a sports video generally occupy only a small portion of the whole content, and the value of sports video drops significantly after a relatively short period of time.[1] 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 start with the 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 create 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. Furthermore, for TV broadcasting, there are commercials or other special information inserted between game sections.[2]

To face the problem of semantic characterization of a multimedia documents, a human being uses his/her cognitive skills, while an automatic system can face it by adopting a two-step procedure: in the first step, some low-level features are extracted in order to represent low-level information in a compact way; in the second step, a decision-making algorithm is used to extract a semantic index from the low-level features.

To characterize multimedia documents, a lot of different audio, visual, and textual features have been proposed and discussed in literature,[3],[4].[5] Specifically the problem of sport content characterization has been given a lot of attention.

For soccer video, for example, the focus was placed initially on shot classification[6] and scene reconstruction.[7] More recently the problems of segmentation and structure analysis have been considered in,[8] ,[9] whereas the

Further authors information: send correspondence to Riccardo Leonardi, E-mail: Riccardo.Leonardi@ing.unibs.it; Pierangelo Migliorati, E-mail: Pierangelo.Migliorati@ing.unibs.it

automatic extraction of highlights and summaries have been analyzed in,[10] ,[11] ,[12] ,[13] ,[14] ,[15] .[16] In,[15] for example, 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 needs a quite complex and accurate tracking system to obtain this information.

As far as baseball sequences are concerned, the problem of indexing for video retrieval has been considered in,[17] whereas the extraction of highlights is addressed in,[18] ,[19] .[2]

The indexing of formula 1 car races is considered in,[20] ,[21] and the proposed approach uses audio, video and textual information.

The analysis of tennis videos can be found, for example, in,[2] ,[22] whereas basketball and football are considered in,[23] ,[24] ,[25] and[26] respectively, to give few examples.

In this paper we analyze some techniques proposed in literature for content characterization of sports videos. The analysis focus on the typology of the signal (audio, video, text, multi-modal, ...) from which the low-level features are extracted.

The paper is organized as follows. In Section 2 some general considerations on content characterization of sports videos are presented. In Section 3 we address the methods based on visual information, whereas in Section 4 we describe the techniques based on audio signal. Section 5 is devoted to the algorithms based on multi-modal analysis. A general discussion on the considered algorithms is given in Section 6, and concluding remarks are drawn in Section 7.

2. GENERAL CONSIDERATIONS ON THE CHARACTERIZATION OF SPORTS VIDEOS

The analysis of the methods proposed in literature for content characterization of sports documents could be addressed in various ways. A possible classification could be based, for example, on the type of sport considered, e.g., soccer, baseball, tennis, basketball, etc. Another possibility could be to consider the methodology used by the characterization algorithm, e.g., deterministic versus statistical approach, to give two possible examples.

In this paper we have analyzed the various techniques from the point of view of the typology of the signal (audio, video, text, ...) from which the low-level features involved in the process of document characterization are extracted.

Considering the audio signal, the related features are usually extracted in two levels: short-term frame-level, and long-term clip-level.[4]
The frame-level features are usually designed to capture the short-term characteristic of the audio signal, and the most widely used have been: 1) Volume ("loudness" of the audio signal); 2) Zero Crossing Rate, ZCR (number of times that the audio waveform crosses the zero axis); 3) Pitch (fundamental frequency of an audio waveform); 4) Spectral features (parameters that describes in a compact way the spectrum of an audio frame).
To extract the semantic content, we need to observe the temporal variation of frame features on a longer time scale. This consideration has lead to the development of various clip-level features, which characterize how frame-level features change over a clip.[4]
These clip-level features are based on the frame-level features, and the most widely used have been:
1) Volume based, mainly used to capture the temporal variation of the volume in a clip;
2) ZCR based, usually based on the statistics of ZCR;
3) Pitch based;
4) Frequency based, that reflect the frequency distribution of the energy of the signal.
Related to the audio signal, there are also the techniques which try to detect and interpret some specific keywords pronounced by the speaker that comments the sports video. This type of information is usually very useful, even if it is very difficult to obtain.

Considering the visual signal, the related features can be categorized into four groups, namely: color, texture, shape, and motion.[4]
1) Color: Color is an important attribute for image representation, and the color histogram, which represent the color distribution in an image, is one of the most used color features.

2) Texture: Texture also is an important feature of a visible surface where repetition or quasi-repetition of a fundamental pattern occurs.
3) Shape: Shape features, that are related to the shape of the objects in the image, are usually represented using traditional shape analysis such as moment invariants, Fourier descriptors, etc.
4) Motion: Motion is an important attribute of video. Motion features, such as moments of the motion field, motion histogram, or global motion parameter have been widely used.

Another important aspect of the analysis of the video signal is the basic segment used to extract the features, that can be composed by one or few images, or by an entire video shot.
Related to the image and video analysis, there are also the techniques in which the textual captions and logos superimposed on the images are detected and interpreted. This captions usually carry a significant semantic information that can be very useful if available,[27] .[28]

In the next sections we will describe some techniques based on visual information, then some methods that analyze audio information, and finally the techniques which consider both audio and visual information, in a multi-modal fashion.

3. TECHNIQUES BASED ON VISUAL INFORMATION

In this section we describe some techniques of content characterization of sports videos that uses features extracted mainly from the image and video signal. To have a more complete description of the features proposed for content analysis based on image and video, refer to,[4].[5]

3.1. Baseball and tennis video analysis

Di Zhong and Shih-Fu Chang at ICME'2001[2] proposed a method for the temporal structure analysis of live broadcast sport videos, using as examples tennis and baseball sequences. 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, a tennis game is divided into sets, then games and serves. 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.

The main objective of the work presented in[2] is the automatic detection of fundamental views (e.g., serve and pitch) that indicates the boundaries of higher level structures. Given the detection results, useful applications such as table of contents and structure summaries can be developed. In particular, in the considered work,[2] the re-current event boundaries, such as pitching and serving views are identified, by using supervised learning and domain-specific rules. The proposed technique for detecting basic units within a game, such as serves in tennis and pitching in baseball, uses the idea that these units usually starts with a special scene. Mainly a color based approach is used, and to achieve higher performance, an object-level verification to remove false alarms was introduced. In particular spatial consistency constraints (color and edge) are considered to segment each frame into regions. Such regions are merged based on proximity and motion. Merged regions are classified into foreground moving objects or background objects based on some rules of motions near region boundaries and long-term temporal consistency. One unique characteristic of serve scenes in tennis game is that there are horizontal and vertical court lines. The detection of these lines is taken into account to improve the performance of the identification algorithm.

The analysis of Tennis video is also carried out in 2001 by Petkovic et al..[22] They propose a method for automatic recognition of strokes in tennis videos based on Hidden Markov Model. The first step is to segment the player from the background, then HMMs is trained to perform the task. The considered features are dominant color, and shape description of the segmented player, and the method appear to lead to satisfactory performance.

The problem of highlights extraction in baseball game videos has been further considered in ICIP'2002 by P. Chang et al..[19] In particular a statistical model is built up in order to explore the specific spatial and temporal structure of highlights in broadcast baseball game videos. The proposed approach is based on two observations. The first is that most baseball highlights are composed of certain types of scene shots, which can be divided into a limited amount of categories. The authors identified seven important types of scene shots, with which most interesting highlights can be composed.

These types of shots are defined as: 1) pitch view, 2) catch overview, 3) catch close-up, 4) running overview, 5) running close-up, 6) audience view and 7) touch-base close-up. Although the exact video streams of the same type of scene shots differ from game to game, they strongly exhibit common statistical properties of certain measurements due to the fact that they are likely to be taken by the broadcasting camera mounted at similar locations, covering similar portions of the field, and used by the cameraman for similar purposes, for example, to capture the overview of the outer field, or to track a running player. As previously mentioned, most highlights are composed of certain types of shots, and the second observation is that the context of transition of those scene shots usually implies the classification of the highlights. In other words, same type of highlights usually have similar transition pattern of scene shots. For example, a typical home run can be composed of a pitch view followed by an audience view and then a running close-up view.
The features used in[19] have been edge descriptor, grass amount, sand amount, camera motion and player height. Of course the context of all the home runs can vary but they can be adequately modelled using a Hidden Markov Model (HMM). In the proposed system an HMM model for each type of highlights is learned. A probabilistic classification is then made by combining the view classification and the HMM model. In summary, the proposed system first segments a digitized game video into scene shots. Each scene shot is then compared with the learnt model, and its associated probability is then calculated. Finally given the stream of view classification probabilities, the probability of each type of highlights can be computed by matching the stream of view classification probabilities with the trained HMMs. In particular the following highlights have been considered: "home run", "catch", "hit", "infield play", and the simulation results appear quite satisfactory.

3.2. Soccer video analysis

Particular attention has been devoted in the literature to the problem of content characterization of soccer video. In soccer video, for example, each play typically contains multiple shots with similar color characteristics. Simple clustering of shots would not reveal high-level play transition relations. Moreover, soccer video does not have canonical views (e.g., pitch) indicating the event boundaries. Due to these considerations, specific techniques have been developed for the analysis of this type od sport video sequence.

In ICME'2001, S.-F. Chang et al.[8] proposed an algorithm for structure analysis and segmentation of soccer video. Some works on sport video analysis and video segmentation are using shot as the base for analysis. However, such approach is often ineffective for sports video due to errors in shot detection, and the lack of or mismatch of domain-specific temporal structure. Starting from this consideration, in,[8] instead of using the shot-based framework, a different approach is proposed, where frame-based domain-specific features are classified into mid-level labels through unsupervised learning, and temporal segmentation of the label sequences is used to automatically detect high-level structure. Moreover, fusion among multiple label sequences based on different features are used to achieve higher performance. In particular, the high level structure of the content is revealed using the information related to the fact that the ball is in play or not. The first step is to classify each sample frame into 3 kinds of view (mid-level labels: global, zoom-in, and close-up) using a unique domain-specific feature, grass-area ratio. Then heuristic rules are used in processing the view label sequence, obtaining play/break status of the game.

The previously described work have been further refined in,[9] where an algorithm for parsing the structure of produced soccer programs is proposed. At first two mutually exclusive states of the game are defined, play and break. A domain-tuned feature set, dominant color ratio and motion intensity, is selected, based on the special syntax and content characteristic of soccer videos. Each state of the game has a stochastic nature that is modelled with a set of hidden Markov models. Finally standard dynamic programming techniques are used to obtain the maximum likelihood segmentation of the game into the two states.

Ekin and Tekalp[16] in SPIE'2003 proposed a framework for analysis and summarization of soccer videos using cinematic and object-based features. The proposed framework includes some novel low-level soccer video processing algorithm, such as dominant color region detection, robust shot boundary detection, and shot classification, as well as some higher-level algorithms for goal detection, referee detection and penalty-box detection. The system can output three types of summaries: 1) all slow motion segments in a game; 2) all goals in a game; 3) slow-motion segments classified according to object features.
The first two types of summaries are based on cinematic features only for speedy computational efficiency, while

the summaries of the last type contain higher-level semantics. In particular the authors propose new dominant color region and shot boundary detection algorithms that are robust to variations in the dominant color, to take into account the fact that the color of the grass field may vary from stadium to stadium, and also as a function of the time of the day in the same stadium. Moreover the algorithm proposed for goals detection is based solely on cinematic features resulting from common rules employed by the producers after goal events to provide a better visual experience for TV audiences. Distinguishing jersey color of the referee is used for referee detection. Penalty box detection is based on the three-parallel-line rule that uniquely specifies penalty box area in a soccer field. Considering for example the algorithm for goal detection they define a cinematic template that should satisfy the following requirements. Duration of the break: a break due to a goal lasts no less than 30 and no more than 120 seconds. The occurrence of at least one close-up/out of field shot: this shot may either be a close-up of a player or out of field view of the audience. The existence of at least one slow-motion replay shot: the goal play is most often replayed one or more times. The relative position of the replay shot: the replay shot follow the close-up/out of field shot.

The problem of highlights extraction in soccer video has been considered also by Leonardi, Migliorati et al.,[10] ,[11] ,[12] .[13]

In,[10] and[11] the correlation between low-level descriptors and the semantic events in a soccer game have been studied. In particular, in,[10] it is shown that low-level descriptors are not sufficient, individually, to obtain satisfactory results (i.e., all semantic events detected with only a few false detections). In,[11] and[13] the authors have tried to exploit the temporal evolution of the low-level descriptors in correspondence with semantic events, by proposing an algorithm based on a finite-state machine. This algorithm gives good results in terms of accuracy in the detection of the relevant events, whereas the number of false detections remains still quite large.
The considered low-level motion descriptors, associated to each P-frame, represent the following characteristics: lack of motion, camera operations (pan and zoom parameters), and the presence of shot-cuts. The descriptor "Lack of motion" has been evaluated by thresholding the mean value of motion vector module. Camera motion parameters, represented by horizontal "pan" and "zoom" factors, have been evaluated using a least-mean squares method applied to P-frame motion fields. Shot-cuts have been detected through sharp transition of motion information and high number of Intra-Coded Macroblocks of P-frames.[29]

The above mentioned low-level indices are not sufficient, individually, to reach satisfactory results. To find particular events, such as, for example, goals or shot toward goal, it is suggested to exploit the temporal evolution of motion indices in correspondence with such events. Indeed in correspondence with goals a fast pan or zoom often occurs followed by lack of motion, followed by a nearby shot cut. The concatenation of these low-level events is adequately modelled with the finite-state machine shown in Fig. 1.

From the initial state SI, the machine goes into state S1 if a fast pan or a fast zoom is detected for at least 3 consecutive P-frames. Then, from state S1 a transition occurs toward the final state SF (goal detection), if a shot-cut is present; from state S1 it goes instead into state S2 if lack of motion is detected for at least 3 consecutive P-frames. From state S2 the final state SF can also be reached if a shot-cut occurs, otherwise a transition toward the initial state S1 occurs 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 initial state SI from states S1 and S2 in case nothing is happening for a certain number of P-frames (corresponding to 1 minute).

The performance of the proposed algorithm have been tested on 2 hours of MPEG2 sequences. Almost all live goals are detected, and the algorithm is able to detect some shots toward goal too, while it gives poor results on free kicks. The number of false detection remains high.

3.3. Replay segment identification

In the analysis of sports videos it is sometimes important the identification of the presence of a slow-motion replay.

In ICASSP'2001, Sezan et al.[28] presented and discussed an algorithm for the detection of slow-motion replay segments in sports video. Specifically the proposed method localizes semantically important events by detecting slow-motion replays of these events, and then generates highlights of these events at different levels. In particular a Hidden Markov Model is used to model the slow-motion replay, and an inference algorithm is introduced which

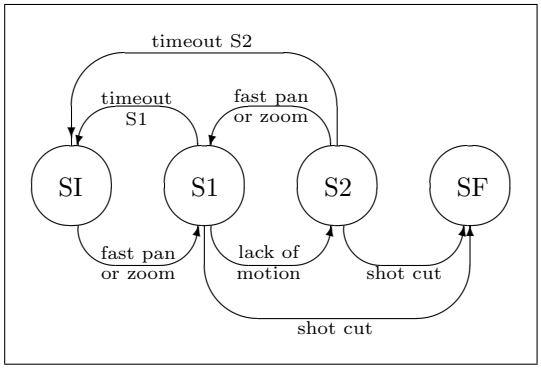

Figure 1. The classification algorithm based on Finite State Machines.

computes the probability of a slow motion replay segment, and localizes the segment boundaries.
Four features are used in the HMM, three of which are calculated from the pixel-wise mean square difference of the intensity of every two subsequent fields, and one of which is computed from the RGB color histogram of each field. The first three features describe the still, normal motion replay, and slow motion fields. The fourth feature is for capturing the gradual transition in editing effects.

An evolution of the previously described work has been presented at ICASSP'2002,[30] where an automatic algorithm for replay segment detection by detecting frames containing logos in the special scene transition and sandwich replays. The proposed algorithm first automatically determines the logo template from frames surrounding slow motion segments, then, it locates all the similar frames in the video using the logo template. Finally the algorithm identifies the replay segments by grouping the detected logo frames and slow motion segments.

4. TECHNIQUES BASED ON THE AUDIO SIGNAL

While current approaches for audiovisual data segmentation and classification are mostly focused on visual cues, audio signal may actually play an important role in content parsing for many applications. In this section we describe some techniques of content characterization of sports videos that uses features extracted mainly from the audio signal associated to the multimedia document. To have a more complete description of the techniques proposed for content analysis based on audio signal, please refer to.[3]

The first example that we consider is related to the characterization of baseball videos, and was proposed in ACM Multimedia 2000, by Rui et al.[18]

In this work the detection of highlights in baseball programs is carried out considering audio-track features alone without relying on expensive to compute video-track features. The characterization is performed considering a combination of generic sports features and baseball specific features, combined using a probabilistic framework. This way highlights detection can even be done on the local set-top box using limited computing power.

The audio track consists of the presenter's speech, mixed with crowd noise, mixed with remote traffic and music noises, and automatic gain control changing the audio level. To have an idea of the feature taken into account, they use the bat-and-ball impact detection to adjust likelihood of a highlight segment, and therefore the same technology could in principle be used also for other sports like golf. In particular, the audio features considered have been: Energy related features, Phoneme-level features, Information complexity features, Prosodic features. These features are used for solving different problems. Specifically some of them are used for human speech endpoint detection, other are used to built a temporal template to detect baseball hits or to model exited human speech. These features have been suitably modelled using a a probabilistic framework.

The performance of the proposed algorithm are evaluated comparing its output against human-selected highlights for a diverse collection of baseball games. They appear very encouraging.

To give another example, we consider the segmentation in three classes of the audio signal associated to a Football audio-video sequence proposed in IWDSC'2002, by Lefevre et al.[26]

In this paper the audio data is divided into short sequences (typically with duration of one or half a second) which will be classified into several classes (speaker, crowd, referee whistle). Every sequence can then be further analyzed depending on the class it belongs to. Specifically the method proposed uses Cepstral analysis and Hidden Markov Models. The results presented in terms of accuracy in the three classes segmentation are good.

5. TECHNIQUES BASED ON MULTI-MODAL ANALYSIS

In the previous sections we have considered some approaches based on the analysis of the audio signal or the image and video signal alone. In this section we will consider some examples of algorithms that use both audio and visual cues, in order to exploit the full potentiality of the multimedia information. To have a more complete description of the features proposed for content analysis based on both audio and visual signal, please refer to,[4,5]

5.1. Baseball audio-visual analysis

Gong et al.[31] at ACM Multimedia 2002 proposed an integrated baseball digest system. The system is able to detect and classify highlights from baseball game videos in TV broadcast. The digest system gives complete indices of a baseball game which cover all status changes in a game. The result is obtained by combining image, audio and speech clues using a maximum entropy method.

The image features considered are the color distribution, the edge distribution, the camera motion, the player detection, and the shot length. Considering the color distribution, the authors observe that every sport game has a typical scene, such as pitch scene in baseball, corner-kick scene in soccer, serve scene in tennis. Color distribution in individual image games are highly correlated for similar scenes. Given the layout of grass and sand in a scene shot of a baseball video, it is easy to detect where the camera is shooting from. Considering the edge distribution, this feature is useful to distinguish audience scenes from field scenes. The edge density is always higher in audience scene, and this information is used as an indicator of the type of the current scene.
Another feature that is considered is the camera motion, estimated using a robust algorithm. Also the player detection is considered as a visual feature. In particular the players are detected considering color, edge and texture information, thus the maximum player size and the number of players are the features associated to each scene shot.

The authors consider also some audio features. In particular the presence of silence, speech, music, hail and mixture of music and speech in each scene shot is detected. To perform this task they use the Mel-cepstral coefficients as the features modelled using a Gaussian Mixture Models. Considering that closed captions provide hints for the presence of highlights, the authors suggest to extract informative words or phrases from closed captions. From the training data, a list of 72 informative words are chosen, such as field, center, strike, etc. The multimedia features are then fused using an algorithm based on the Maximum Entropy Method to perform the highlights detection and classification.

5.2. Formula 1 car races audio-visual analysis

Petkovic et al.[20] at ICME'2002 proposed an algorithm for the extraction of highlights from TV Formula 1 programs. The extraction is carried out considering a multi-modal approach that uses audio, video and superimposed text annotation combined by Dynamic Bayesian Networks (DBN).

In particular, four audio features are selected for speech endpoint detection and extraction of excited speech, namely: Short Time Energy (STE), pitch, Mel-Frequency Cepstral Coefficients (MFCC) and pause rate. For the recognition of specific keywords in the announcer's speech a keyword-spotting tool based on a finite state grammar has been used.

In the visual analysis, color, shape and motion features are considered. First the video is segmented into shots based on the differences of color histogram among several consecutive frames. Then the amount of motion is estimated and semaphore, dust, sand and replay detectors are applied in order to characterize passing, start and fly-out events.
The third information source used in the processing is the text that is superimposed on the screen. This is

another type of on-line annotation done by the TV program producer, which is intended to help the viewer to better understand the video content. The superimposed text often brings some additional information that is difficult or even impossible to deduce solely by looking at the video signal.
The details of this algorithms are described in.[21] The reported results show that the fusion of cues from the different media has resulted in a much better characterization of Formula 1 races. The audio DBN was able to detect a great number of segments where the announcer raised his voice, which corresponds to only the 50% of all interesting segments, i.e., highlights in the race. The integrated audio-visual DBN was able to correct the result and detect about 80% of all interesting segments in the race.

5.3. Soccer audio-visual analysis

Leonardi et al.[14] at WIAMIS'2003 presented a semantic soccer-video indexing algorithm that uses controlled Markov chains[32] to model the temporal evolution of low-level video descriptors.[12] To reduce the number of false detections given by the proposed video-processing algorithm, they add the audio signal characteristics. In particular they have evaluated the "loudness" associated to each video segments identified by the analysis carried out on the video signal. The intensity of the "loudness" has then been used to order the selected video segments. In this way, the segments associated to the interesting events appear in the very first positions of the ordered list, and the number of false detections can be greatly reduced.

The low-level binary descriptors, associated to each P-frame, represent the following characteristics: lack of motion, camera operations (pan and zoom parameters), and the presence of shot-cuts, and are the same descriptors used in.[11] Each descriptor takes value in the set $\{0, 1\}$.

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. In order to model this fact, we describe the state of the system as a two-component state, and also, we impose a certain structure on the controlled transition probability function.

A schematic representation of the introduced model is given in Figure 2. In this figure, the symbol "$\sim$" is used for "distributed according to".

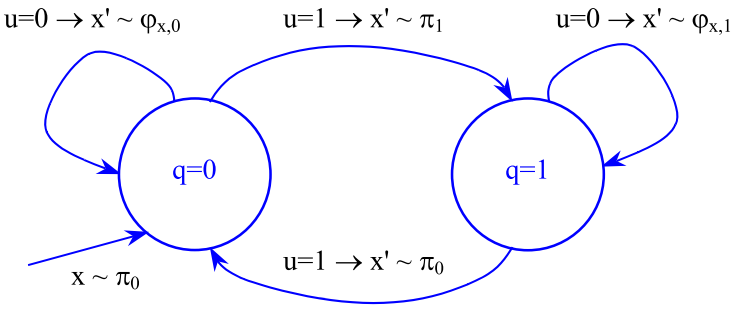

Figure 2. Controlled Markov chain model.

In our context, $\mathcal{T}$ represents the set of time instants associated with the P-frames sequence. As for $\mathbf{x}(t)$, it is state of the P-frame observed at time t. In particular, $\mathbf{x}(t)$ can take the following values: "LM", "FP", "FZ", "FPZ", and "Other", hence the set $\mathcal{X}$ has cardinality 5. The value taken by $\mathbf{x}(t)$ is evaluated by means of the previously described low-level descriptors.

We suppose that each semantic event takes place over a two-shot block and that it can be modeled by a controlled Markov chain with the structure described above. Each semantic event is then characterized by the two sets of probability distributions over the state space. Specifically, we have considered 6 models denoted by A, B, C, D, E, and F, where model A is associated to goals, model B to corner kicks, and models C, D, E, F describe other situations of interest that occur in soccer games, such as free kicks, plain actions, and so on.
On the basis of the derived six Controlled Markov models, one can classify each pair of shots in a soccer game video

	Soccer	Baseball	Tennis	Formula 1
Color	SFC,[8] SFC,[9] Tek[16]	SFC,[2] PCh,[19] Gong[31]	SFC,[2] Pet2[22]	Pet1[20]
Texture		PCh,[19] Gong[31]		
Shape	Tek[16]	PCh,[19] Gong[31]	Pet2[22]	Pet1[20]
Edge, line	SFC,[8] Tek[16]	SFC,[2] PCh,[19] Gong[31]	SFC[2]	
Motion	SFC,[8] SFC,[9] Tek,[16] LM,[11] LMP[14]	SFC,[2] PCh,[19] Gong[31]	SFC[2]	Pet1[20]
Caption, text				Pet1[20]
Volume	LMP[14]	Rui[18]		Pet1[20]
ZCR		Rui,[18] Gong[31]		Pet1[20]
Pitch		Rui,[18] Gong[31]		Pet1[20]
Spectral		Rui,[18] Gong[31]		Pet1[20]
Word spotting		Gong[31]		Pet1[20]

Table 1. Overview of the principal features used by the considered algorithms.

sequence by using the maximum likelihood criterion. For each pair of consecutive shots (i.e., two consecutive sets of P-frames separated by shot-cuts), one needs to i) extract the sequence of low-level descriptors, ii) determine the sequence of values assumed by the state variable, and iii) determine the likelihood of the sequence of values assumed by the low-level descriptors 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.

The performance of the proposed algorithm have been tested considering about 2 hours of MPEG2 sequences containing more than 800 shot-cuts, and the results are very promising. The number of false detections are still quite relevant. As the results are obtained using motion information only, it was decided to reduce the false detection associating to the candidates pairs shots the audio loudness.

To extract the relevant features, we have divided the audio stream of a soccer game sequence in consecutive clips of 1.5 seconds, in order to observe a quasi-stationary audio signal in this window.[4] For each frame the "loudness" is estimated as the energy of the sequence of audio samples associated to the current audio-frame.
The evolution of the "loudness" in an audio clip follows the variation in time of the amplitude of the signal, and it constitutes therefore a fundamental aspect for audio signal classification. We estimate the mean value of the loudness for every clip. In this way we obtain, for each clip, a low-level audio descriptor represented by the "clip loudness".
The false detections given by the candidate pairs of shots obtained by video processing are reduced by ordering them according to the average value of the "clip loudness" along the time span of the considered segment. In this way, the video segments containing the goals appear in the very fist positions of this ordered list. The simulation results appear to be very encouraging, reducing the number of false detection by an order of magnitude.

6. DISCUSSION ON THE CONSIDERED ALGORITHMS

In Table 1 an overview of the principal features used in the considered algorithms is presented. As we can see, most of them are based on the visual signal, and, from one sport to the other, the type of features analyzed can differ significantly. Moreover, some visual features such as color and motion features are used in almost all the cases taken into account. Considering the algorithms that uses a multi-modal approach, we can see that the basic idea is to exploit as much as possible all the available features. A very powerful information is that related to the textual caption superimposed to the video signal, or the recognition of some keywords pronounced by the presenter; anyway this information is difficult to obtain, and therefore it is not yet widely used.

The analysis of the techniques for sports video characterization suggests an important consideration about the modality in which the interesting event are captured by the algorithms. We can clearly see that the characterization is carried out either considering what happens in a specific time segment, observing therefore the features in a "static" way, or trying to capture the "dynamic" evolution of the features in the time domain.

In tennis, for example, if we are looking for a serve scene, we can take into account the specific, well known situation, in which we have the player in great evidence in the scene, and so we can try to recognize its shape,

as suggested in.[22]
In the same way, if we are interested in the detection of the referee in a soccer game video, we can look for the color of the referee jersey, as suggested in.[16] Also the detection of a penalty box can be obtained by the analysis of the parallel field lines,[16] which is clearly a "static" evaluation.
Considering for example formula 1 car races, calculating the amount of motion, and detecting the presence of the semaphore, dust, and sand, in,[20] the start and fly-out events are characterized.
In all these examples, it is clear that is a "static" characteristic of the interested highlights that is captured by the automatic algorithm. Similar situation occurs in baseball, if we are interested for example in pitch event detection, and we consider that each pitch usually starts with a pitching view taken by the camera behind the pitcher.

On the other hand, if we want for example to detect the goals in a soccer game, we can try to capture the "dynamic" evolution of some low-level features, as suggested in,[11] ,[13] or try to recognize some specific cinematic patterns, as proposed in.[16] In the same way, we are looking to a "dynamic" characteristic if we want to determine automatically the slow-motion replay segments, as suggested in.[30]

The effectiveness of each approach depends mainly on the kind of sports considered, and from the type of highlights we are interested in.

7. CONCLUSIONS

In this paper we have analyzed various techniques proposed in literature for content characterization of sports videos, focusing on the type of signal from which the low-level features are extracted. First we have considered the techniques based on visual information, then some methods based on audio signal, and finally some techniques based on audio-visual information. To give an example related to soccer video indexing, we have analyzed in more detail a semantic indexing algorithm based on controlled Markov chain models that captures the temporal evolution of low-level motion descriptors. To reduce the number of false detections given by the video processing algorithm, audio signal information has been added. In particular the "loudness" associated to each selected video segments has been evaluated. The intensity of the "loudness" has then been used to order the selected video segments, allowing the most interesting events to appear in the very first positions of the ordered list.
This analysis shows that each type of signal carries some peculiar information, and the multi-modal approach can efficiently exploit the multimedia information associated to the sports video. Moreover, we observe that the characterization is usually performed either considering the features in a "static" way, analyzing therefore what happens in a specific time segment, or trying to capture their "dynamic" evolution in time. The effectiveness of each approach depends on the kind of sports it relates to, and the type of highlights we are focusing on.

ACKNOWLEDGMENTS

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

REFERENCES

1. S.-F. Chang, "The holy grail of content-based media analysis," *IEEE Multimedia* **9**, pp. 6–10, Apr.-June 2002.
2. D. Zhong and S.-F. Chang, "Structure analysis of sports video using domain models," in *Proc. ICME'2001*, pp. 920–923, Aug. 2001, Tokyo, Japan.
3. T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. on Speech and Audio Processing* **9**, pp. 441–457, 2001.
4. Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis using audio and visual information," *IEEE Signal Processing Magazine* **17**, pp. 12–36, 2000.
5. C. Snoek and M. Worring, "Multimodal video indexing: a review of the state-of-the-art," in *ISIS Technical Report Series, Vol. 2001-20*, Dec. 2001.
6. Y. Gong, L. Sin, C. Chuan, H. Zhang, and M. Sakauchi, "Automatic parsing of TV soccer programs," in *Proc. ICMCS'95*, May 1995, Washington DC, USA.

7. D. You, B. Yeo, M. Yeung, and G. Liu, "Analysis and presentation of soccer highlights from digital video," in *Proc. ACCV 95*, Dec. 1995, Singapore.
8. P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," in *Proc. ICME'2001*, pp. 928–931, Aug. 2001, Tokyo, Japan.
9. L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden markov models," in *Proc. ICASSP'2002*, May 2002, Orlando, FL, USA.
10. A. Bonzanini, R. Leonardi, and P. Migliorati, "Semantic video indexing using MPEG motion vectors," in *Proc. EUSIPCO'2000*, pp. 147–150, Sept. 2000, Tampere, Finland.
11. A. Bonzanini, R. Leonardi, and P. Migliorati, "Event recognition in sport programs using low-level motion indices," in *Proc. ICME'2001*, pp. 920–923, Aug. 2001, Tokyo, Japan.
12. R. Leonardi, P. Migliorati, and M. Prandini, "Modeling of visual features by markov chains for sport content characterization," in *Proc. EUSIPCO'2002*, Sept. 2002, Toulouse, France.
13. R. Leonardi and P. Migliorati, "Semantic indexing of multimedia documents," *IEEE Multimedia* **9**, pp. 44–51, Apr.-June 2002.
14. R. Leonardi, P. Migliorati, and M. Prandini, "A markov chain model for semantic indexing of sport program sequences," in *Proc. WIAMIS'03*, Apr. 2003, London, UK.
15. V. Tovinkere and R. J. Qian, "Detecting semantic events in soccer games: Toward a complete solution," in *Proc. ICME'2001*, pp. 1040–1043, Aug. 2001, Tokyo, Japan.
16. A. Ekin and M. Tekalp, "Automatic soccer video analysis and summarization," in *Proc. SST SPIE03*, Jan. 2003, CA, USA.
17. T. Kawashima, K. Takeyama, T. Iijima, and Y. Aoki, "Indexing of baseball telecast for content based video retrieval," in *Proc. ICIP'98*, pp. 871–874, Oct. 1998, Chicago, IL., USA.
18. Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. ACM Multimedia 2002*, pp. 105–115, 2000, Los Angeles, CA, USA.
19. P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden markov models," in *Proc. ICIP'2002*, pp. 609–612, Sept. 2002, Rochester, NY.
20. M. Petrovic, V. Mihajlovic, W. Jonker, and S. Djordievic-Kajan, "Multi-modal extraction of highlights from tv formula 1 programs," in *Proc. ICME'2002*, Aug. 2002, Lausanne, Switzerland.
21. V. Mihajlovic and M. Petrovic, "Automatic annotation of formula 1 races for content-based video retrieval," in *Technical report, TR-CTIT-01-41*, Dec. 2001.
22. M. Petkovic, W. Jonker, and Z. Zivkovic, "Recognizing strokes in tennis videos using hidden markov models," Marbella, Spain, 2001.
23. W. Zhou, A. Vellaikal, and C.-C. J. Kuo, "Rule based video classification system for basketball video indexing," in *Proc. ACM Multimedia 2000*, Dec. 2002, Los Angeles, CA, USA.
24. D. Saur, Y. Tan, S. Kulkarni, and P. Ramadge, "Automated analysis and annotation of basketball video," in *SPIE Vol. 3022*, Sept. 1997.
25. G. Sudhir, J. lee, and A. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in *IEEE Multimedia*, 1997.
26. S. Lefevre, B. Maillard, and N. Vincent, "3 classes segmentation for analysis of football audio sequences," in *Proc. ICDSP'2002*, July 2002, Santorin, Grece.
27. M. Bertini, C. Colombo, and A. D. Bimbo, "Automatic caption localization in videos using salient points," in *Proc. ICME'2001*, pp. 69–72, Aug. 2001, Tokyo, Japan.
28. H. Pan, B. Li, and M. Sezan, "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transition," in *Proc. ICASSP'2002*, May 2002, Orlando, FL, USA.
29. T. Sikora, "Mpeg digital video-coding standards," *IEEE Signal Processing Magazine* **14**, 1997.
30. H. Pan, P. Beek, and M. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. ICASSP'2001*, May 2001, Salt Lake City, USA.
31. M. Han, W. Hua, W. Xu, and Y. Gong, "An integrated baseball digest system using maximum entropy method," in *Proc. ACM Multimedia 2002*, Dec. 2002, Juan Les Pins, France.
32. M. L. Puterman, *Markov Decision Processes*, Wiley, New York, 1994.