Semantic Indexing of Soccer Audio-Visual Sequences: A Multimodal Approach Based on Controlled Markov Chains

Riccardo Leonardi, *Member, IEEE*, Pierangelo Migliorati, *Member, IEEE*, and Maria Prandini, *Member, IEEE*

***Abstract*—Content characterization of sport videos is a subject of great interest to researchers working on the analysis of multimedia documents. In this paper, we propose a semantic indexing algorithm which uses both audio and visual information for salient event detection in soccer. The video signal is processed first by extracting low-level visual descriptors directly from an MPEG-2 bit stream. It is assumed that any instance of an event of interest typically affects two consecutive shots and is characterized by a different temporal evolution of the visual descriptors in the two shots. This motivates the introduction of a controlled Markov chain to describe such evolution during an event of interest, with the control input modeling the occurrence of a shot transition. After adequately training different controlled Markov chain models, a list of video segments can be extracted to represent a specific event of interest using the maximum likelihood criterion. To reduce the presence of false alarms, low-level audio descriptors are processed to order the candidate video segments in the list so that those associated to the event of interest are likely to be found in the very first positions. We focus in particular on goal detection, which represents a key event in a soccer game, using camera motion information as a visual cue and the "loudness" as an audio descriptor. The experimental results show the effectiveness of the proposed multimodal approach.**

***Index Terms*—Controlled Markov chains, highlights detection, multimodal analysis, semantic indexing, sport videos.**

I. Introduction

THE efficient distribution of sport videos over various networks should contribute to the rapid adoption and widespread usage of multimedia services, because sport videos appeal to large audiences. The problem of designing effective automatic techniques for the semantic characterization of sport video documents has therefore attracted the attention of many researchers.

To face the problem of semantic characterization of a multimedia document, a human being uses his/her cognitive skills, while an automatic system can resort to a two-step procedure [1] where, in the first step, some low-level descriptors are extracted in order to represent low-level information in a compact way, and, in the second step, a decision-making algorithm is used to extract a semantic index from the low-level descriptors (see Fig. 1).

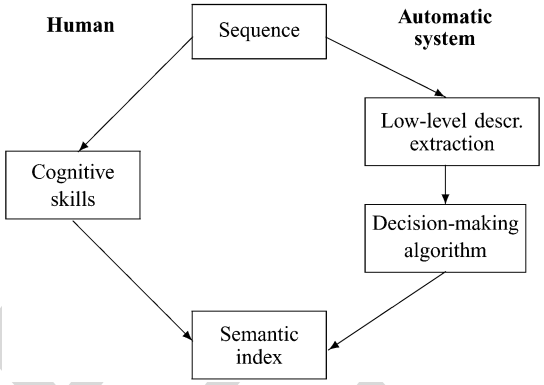

Fig. 1. Problem of multimedia content characterization: a two-step solution to the design of an automatic system for semantic indexing.

The use of many different audio, visual, and textual low-level descriptors have been proposed in the literature to characterize multimedia documents [2]–[4].

Compared to other videos such as news and movies, sport videos have a well-defined content structure and domain rules. The valuable semantics in a sport video generally occupy only a small portion of the whole content; in addition, its value drops significantly after a relatively short period of time [5].

A game of sport is often organized so as to define well-structured temporal references identifying specific components of the game which can form a complete hierarchy. 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 the formation in which players line up on two sides of the ball. A tennis game is divided into sets, each set into games, and each game into serves. In TV broadcasting of sport games, commercials or other special information are inserted between specific game segments [6]. In addition, in sport videos there is a fixed number of cameras in the field that can be selected to reproduce unique scenes according to the specific video segment it represents. 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. These *a priori* pieces of information on the presence of canonical scenes (the pitching scene in baseball or the serve scene in tennis) could simplify the problem of designing an automatic semantic indexing system. However, this is

Manuscript received April 30, 2003; revised September 11, 2003. This work was supported in part by the IST programme of the EU under Projects IST-2001-32 795 SCHEMA, and IST-2000-28 304 SPATION.

R. Leonardi and P. Migliorati are with the Department of Electronics for Automation of the University of Brescia, **[AUTHOR: PLEASE PROVIDE A COMPLETE ADDRESS.—ED.]**Italy (e-mail: leon@ing.unibs.it; pier@ing.unibs.it).

M. Prandini is with the Department of Electronics of the Politecnico of Milano, **[AUTHOR: PLEASE PROVIDE A COMPLETE ADDRESS.—ED.]**Italy (e-mail: prandini@elet.polimi.it).

Digital Object Identifier 10.1109/TCSVT.2004.826751

not the case for soccer videos. The video sequence of each play in a soccer game typically contains multiple shots with similar characteristics. Thus, simple clustering of shots according to the field of view would not reveal high-level play transition relations. Due to these considerations, specific techniques have to be developed for the analysis of this type of sport program at least when considering visual information. An overview of the corresponding literature is described in Section II.

Current approaches for audio-visual data segmentation and classification are mostly focused on visual cues. However, audio plays an important role in content parsing for various applications. Many different low-level audio features have in fact been proposed and discussed in the literature for video content characterization [2]–[4].

In the specific case of sport programs, the focus has been on excited/nonexcited commentary classification [7] and audio event spotting (such as baseball hits or football touchdowns) [8], [7]. The presence of specific characteristics of the audio corresponding to hits or touchdowns can facilitate the detection of highlights in baseball and football programs. In [7], the analysis of baseball programs was in fact carried out considering only audio-track features. In this case, the audio track consists of the presenter's speech mixed with crowd noise, remote traffic and music noises, and automatic gain control changing the audio level. Features related to the energy and the information complexity, phoneme level, and prosodic features were used to solve different problems, such as detecting human speech endpoints and building a temporal template for detecting baseball hits or excited human speech. In [9], the segmentation in three classes (speaker, crowd, and referee whistle) of the audio signal associated to a football audio-video sequence was studied. Specifically, a method using cepstral analysis and Hidden Markov Models (HMMs) was proposed.

In [10] and [11], a multimodal approach was adopted where audio is used jointly with video/text for content characterization. In [10], an integrated digest system was proposed to detect and classify highlights from baseball game videos in a 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. In particular, a method for distinguishing silence, speech, music, hail, and mixture of music and speech combining cepstral analysis and Gaussian mixture models was proposed.

In [11], an algorithm for the extraction of highlights from TV Formula 1 programs was presented. The extraction is carried out considering a multimodal approach that uses audio, video, and superimposed text annotation combined in a dynamic Bayesian network.

In the case of soccer 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 sport 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 [12], [13].

In this paper, we present a semantic indexing algorithm for soccer audio-visual sequences using jointly visual and audio information for semantic event detection. For this program category, the semantic content can in fact be related to the occurrence of interesting events such as, for example, goals, shots to goal, and free kicks. These events occur at the beginning or at the end of the game actions. Therefore, a good semantic index of a soccer video sequence could be the list of all game actions, each one characterized by its beginning and ending event. Such a summary could be very useful to satisfy various types of semantic queries.

We shall report simulation experiments for the detection of goals, which represent the key event in a soccer game. The associated algorithm which can be used to describe other salient events in soccer consists of a two-step procedure, where the video signal is processed first so as to produce a list of candidate video segments, and the candidate segments are then ranked based on the audio information. The algorithm could be extended, as well, to address the problem of detecting semantic events in other sport programs, for which a two-stage transition model of the visual cues is adequate.

The video processing step proposed here is inspired by [14] and [15]. In these papers, the temporal evolution of low-level motion descriptors is used to detect goals in a soccer program. A deterministic automaton is introduced, whose transitions between nodes are determined by the evolution in time of the motion descriptors. The automaton may reach a goal node whenever an appropriate timed sequence of motion-based events occurs.

The advantage of using only simple motion descriptors is that the motion vector field is directly available in the MPEG compressed bit stream, hence the algorithm is easy to implement. However, the algorithm in [14] and [15] is accurate in terms of the number of detected events, but presents a high number of false positives. Roughly speaking, in order to capture all goal events, the automaton timing requirements have to be chosen so as to cover all goal instances, according to a "worst-case" approach, with the undesired side effect of capturing also other no-goal situations. In addition, this is difficult to be robustly transferred on a variety of test material.

In order to reduce the number of false positives, the temporal evolution of the low-level visual descriptors was modeled by a controlled Markov chain in [16], on which this work is partly based. Controlled Markov chain models include standard Markov chains as a subclass, the added feature being the presence of a control input affecting the transition probabilities of the Markov chain. In our specific context, the use of controlled Markov chains allows to describe the temporal evolution of the motion descriptors through transition probabilities with different characteristics in different shots, with the control input modeling the occurrence of a shot-cut.

The key idea when adopting a controlled Markov chain model for goal detection is that a goal event typically affects a pair of shots (the shot where the goal actually takes place, and the subsequent one with peculiar crowd and players reaction), and these two shots have different characteristics.

Differently from [16], we suggest to add information contained in the audio signal to complement the information con-

tained in the video signal. In particular, we adopt as an audio descriptor the increase of the audio signal loudness between two shots: the higher the increase, the more likely is the occurrence of a goal in the associated shot pair.

The reason for choosing such a simple audio descriptor is twofold: simplicity of implementation of the resulting goal detection algorithm and intrinsic difficulty in analyzing the sound track of a soccer program.

The rest of the paper is organized as follows. An overview of the contributions on semantic indexing of soccer programs in the literature is given in Section II. The goal detection algorithm is presented in Section III, by describing first the adopted low-level descriptors (Section III-A) and then the video processing and audio processing stages (Sections III-B and III-C). Experimental results showing the effectiveness of the proposed approach are reported in Section IV. Finally, concluding remarks are drawn in Section V.

II. Overview of the Contributions on the Analysis of Soccer Programs

Specific techniques have been developed in the literature for the analysis of soccer programs based on visual features, starting from shot classification [17] and scene reconstruction [18], to address—though only more recently—the problems of segmentation and structure analysis [19], [20] and of extraction of highlights and summaries [21], [14], [16], [15], [12], [22], [23].

In [19], the authors proposed an algorithm for structure analysis and play-break segmentation of sport videos, which, differently from the shot-based analysis methods proposed in other papers, relies on the domain-specific temporal structure of the frame sequence for high-level content characterization. Frame-based domain-specific features are classified through unsupervised learning into different categories, each one identified by a label. The resulting temporal segmentation of the so-obtained mid-level label sequence is used to determine the high-level structure of the video. Fusion among multiple label sequences based on different features can be used to achieve higher performance. In the case of soccer game videos, each sample frame is classified into three categories of views labeled as "global," "zoom-in," and "close-up," by using the "dominant color ratio" feature representing the grass area fraction. Heuristic rules are then used to process the obtained mid-level labels sequence so as to identify the play or break status of the game. The temporal segmentation of the video sequence in play/break segments can then be used to detect high-level structures.

The work in [19] was further refined in [20], where the authors proposed a probabilistic algorithm for determining the play/break segmentation of a soccer video based on the domain-specific frame features "dominant color ratio" and "motion intensity." The play and break conditions are modeled by a set of HMMs with observations given by the "dominant color ratio" and "motion intensity" features. At each time t, a sliding sequence of observations is considered, and its maximum likelihood over each of the HMMs is evaluated. Dynamic programming is used to decide if the game is in a break or a play condition at any time t during the whole duration of the program, based on the two sequences of maximum-likelihood values and the correlation in time of the break and play conditions.

The automatic extraction of highlights and summaries of soccer videos have been analyzed in a few contributions [21], [14], [16], [15], [12], [22], [23].

In [22], a method that tries to detect a large set of semantic events which may happen in a soccer game is presented. This method uses the information of the position of the players and the ball during the game as input and therefore needs a quite complex and accurate tracking system.

In [23], the authors proposed a framework for the analysis and summarization of soccer videos using kinematic and object-based features. The proposed framework includes some novel low-level soccer video processing algorithms, such as dominant color region detection, shot boundary detection, and shot classification, as well as some higher level algorithms for goal detection, referee detection, and penalty-box detection. In particular, the authors introduced 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 grass color may vary from stadium to stadium, and depends on the time of the day within any given stadium. The algorithm proposed for goal detection is based solely on kinematic features resulting from common rules adopted after goal events by the producers 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. Accordingly, three types of summaries can be produced, where a list of the slow-motion segments in a game, the goals in a game, and the slow-motion segments classified according to object features. The first two types of summaries are based on kinematic features only for computational efficiency, while the summaries of the last type require higher level semantic characterization.

The problem of highlights extraction in soccer video was considered also in [21], [14], [16], and [15]. In these papers, the correlation between low-level visual descriptors and the semantic events in a soccer game was studied. In particular, in [21], it was shown that individual low-level descriptors are not sufficient to obtain satisfactory results if they are used individually. In [14] and [15], the authors proposed an algorithm based on a finite-state machine, where the sequencing in time of the low-level visual descriptors is exploited to detect semantic events. This algorithm gives good results in terms of accuracy in the detection of the relevant events, but the number of false detections remains still quite large.

III. Proposed Multimodal Analysis Method for Goal Detection

In the proposed algorithm for goal detection, audio and visual information are processed according to the following two steps:

Step 1) Low-level visual descriptors associated with motion information and camera panoramic/zoom views are directly extracted from the MPEG bit stream. The goal event is supposed to take place over a shot pair, i.e., two consecutive sets of P-frames separated

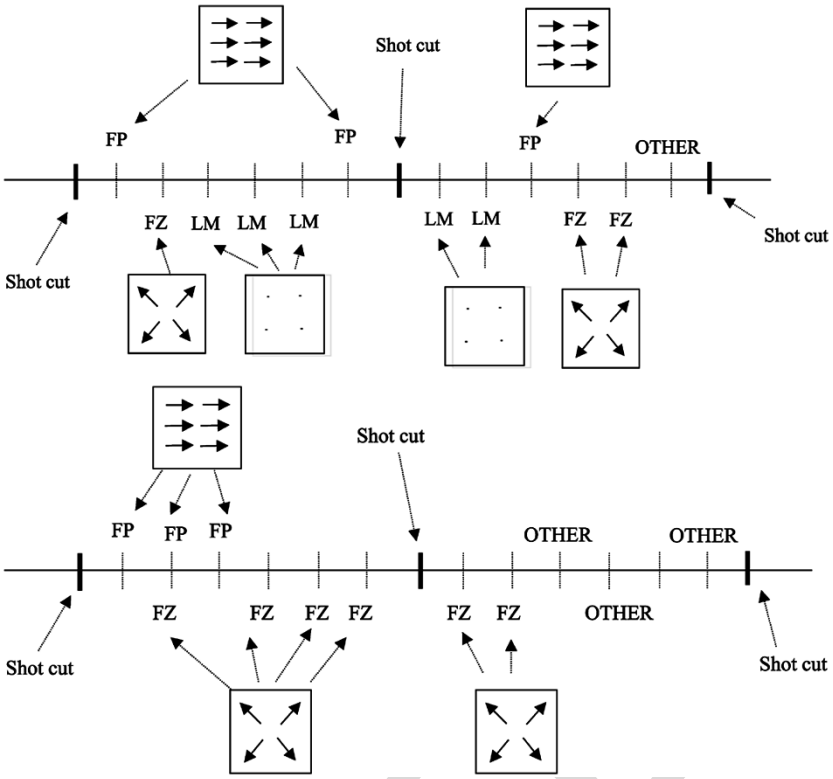

Fig. 2. Evolution of the visual descriptors in a shot pair where no interesting event occurs (above) and a goal event occurs (below). LM: "Lack of motion"; FP: "Fast pan"; and FZ: "zoom"; OTHER: types of motion.

by a shot-cut. During the two shots when a goal takes place, the temporal evolution of the extracted low-level visual descriptors is modeled by a controlled Markov chain (CMC) with the shot-cut representing the control input affecting the CMC transition probabilities. If no goal occurs in a shot pair, the temporal evolution of the descriptors is governed by various CMCs that can model either relevant events (e.g., corner kicks or free kicks) or situations where no particular event takes place. The shot pairs of the soccer sequence where a goal is likely to take place can thus be identified on a maximum-likelihood (ML) criterion;

Step 2) The low-level audio descriptor "loudness" of the audio signal associated with shot pairs extracted in step 1) is evaluated. The candidate shot pairs are ordered in a list based on the difference between the intensity of the loudness signal in the second shot and that in the first shot: the larger the difference, the higher the position of the shot pair in the list. In this way, the segments associated with a goal should appear in the first positions of the ordered list.

Note that the proposed algorithm is shot-pair-based. In our implementation, we use a method for shot-cut detection that relies on the low-level visual descriptor mentioned at step 1) and was introduced in [14].

We next describe the adopted low-level descriptors and then detail steps 1) and 2) briefly outlined above.

A. Low-Level Descriptors

In this section, we describe the low-level descriptors used in the proposed automatic semantic indexing system and the method adopted for their evaluation.

1) Visual Descriptors: We consider the visual descriptors "Lack of motion," "Fast pan," and "Fast zoom," which were originally proposed in [14]. These descriptors summarize the information on: 1) lack of motion and 2) 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 instead in the case of shots toward the goal-keeper or fast ball exchanges between distant players. Finally, 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 from the MPEG bit stream.

In Fig. 2, we report an example of the evolution of the visual descriptors associated with two shots when no interesting event occurs (top of Fig. 2) and when a goal occurs (bottom of Fig. 2). We shall see in the sequel how the evolution in time of the visual descriptors is easily captured through a different probabilistic model in both situations.

The descriptor "Lack of motion" is evaluated by thresholding the mean value μ of the motion vectors module as follows:

$$\mu = \frac{1}{M_x M_y - N_I} \sum_{i=0}^{M_x-1} \sum_{j=0}^{M_y-1} \sqrt{v_x^2(i,j) + v_y^2(i,j)} \qquad (1)$$

where M_x and M_y are the horizontal and vertical dimensions (in macro-blocks) of the P-frame, N_I is the number of intra-coded macro-blocks, and $v_x(i,j)$ and $v_y(i,j)$ are the horizontal and vertical components of the motion vector at position (i,j). The "Lack of motion" descriptor is set to 0 when μ exceeds a given threshold. The selected threshold value is 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 [24]. 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 set equal to 20 for the "Fast pan" descriptor and to 0.002 for the "Fast zoom" descriptor.

2) Shot-Cut: The problem of shot-cut detection for video segmentation has been given a lot of attention in literature. The reader is referred to [25] for an overview. In our implementation, shot-cuts are detected using the low-level visual descriptors extracted from the MPEG bit stream [14].

Specifically, to detect if there is a shot-cut between two consecutive P-frames, say P-frame $k-1$ and P-frame k, we use the difference between the mean value of the motion vectors modules associated with the two P-frames [26] as follows:

$$\Delta\mu(k) = \mu(k) - \mu(k-1)$$

where $\mu(k)$ and $\mu(k-1)$ are computed according to (1) in reference to P-frames k and $k-1$, respectively. This parameter is likely to exhibit a high value in the presence of a shot-cut, which leads to an abrupt change in the motion field between the two considered P-frames.

The information provided by $\Delta\mu(k)$ is suitably combined with the number of intra-coded macro-blocks of the P-frame $k, \mathrm{Intra}(k)$, as follows:

$$\mathrm{Cut}(k) = \mathrm{Intra}(k) + \beta\Delta\mu(k)$$

where β is a weighting factor. When the parameter $\mathrm{Cut}(k)$ is greater than a predefined threshold, we assume that a shot-cut has occurred [14]. The β parameter is set to 20 and the threshold to 700.

3) Audio Descriptor: We refer to the "clip" as elementary unit for the audio signal processing. A "clip" is a set of consecutive audio samples corresponding to a 1.5-s time interval, during which the audio signal can be considered quasi-stationary [3]. Each clip is composed of a certain number of partially overlapping "audio-frames," the exact number depending on the sampling frequency. In our experiments where the sampling frequency is 44.1 kHz, a clip is composed of 128 audio-frames, each frame containing 1024 consecutive audio samples with 512 samples in common with the previous frame [3].

The "loudness" of frame k is given by

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

where $\{y_k(n), n = 0, \ldots, N-1\}$ is the set of $N = 1024$ audio samples of frame k. By estimating the mean value of the frame loudness over each clip, we obtain a low-level descriptor of the audio signal which we call "clip loudness."

The evolution of the clip loudness 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 roar in the crowd.

Fig. 3 represents a typical plot of the clip loudness behavior in a shot pair associated respectively to a goal and to a generic no-goal situation. Note that the clip loudness exhibits a peculiar behavior corresponding to 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 for no-goal shot pairs.

B. Video Processing Step

As mentioned earlier, it is assumed that a goal event takes place over two shots. The evolution of the low-level visual descriptors during these two shots is governed by a discrete time CMC, with the shot-cut playing the role of control input, and the discrete time reference given by the sequence of P-frames composing the shot pair. In those shot pairs when no goal occurs, the visual descriptors evolve according to a different discrete time CMC.

We next describe a general discrete time CMC model [27], and then apply it to our context.

The components of a CMC model are the state and input variables, the initial state probability distribution, and the controlled transition probability function. Here, we consider homogeneous models with state and input variables taking values in finite sets.

Denote by $\mathbf{s}(t)$ the random variable representing the state of the CMC at time $t \in \mathcal{T} := \{0, 1, 2, \ldots\}$. At each $t \in \mathcal{T}$, the state $\mathbf{s}(t)$ takes its value in a discrete state set $\mathcal{S}$. At time $t = 0$, the initial state $\mathbf{s}(0)$ is described in terms of its probability distribution, say P_0, over the space set $\mathcal{S}$. The evolution of $\boldsymbol{s}(t)$ from time $t \in \mathcal{T}$ to time $t+1$ is governed by a transition probability. This probability is affected by an input signal, denoted by $\mathbf{u}(t)$, taking value in a discrete input set $\mathcal{U}$, which in general is a random variable. The probability of transition is only a function of the input $u \in \mathcal{U}$ applied at time t. By this and given the Markovianity of the model, we mean that $\mathbf{s}(t+1)$ is a random variable conditionally independent of all other random variables at times smaller or equal to t, given $\mathbf{s}(t)$ and $\mathbf{u}(t)$. Here we assume a stationary transition probability, i.e.,

$$\mathrm{P}(\mathbf{s}(t+1) = s' \,|\, \mathbf{s}(t) = s, \quad \mathbf{u}(t) = u) = p(s, s', u)$$

$\forall s, s' \in \mathcal{S}, u \in \mathcal{U}, t \in \mathcal{T}$, where $p : \mathcal{S} \times \mathcal{S} \times \mathcal{U} \to [0, 1]$ is the *controlled transition probability function*.

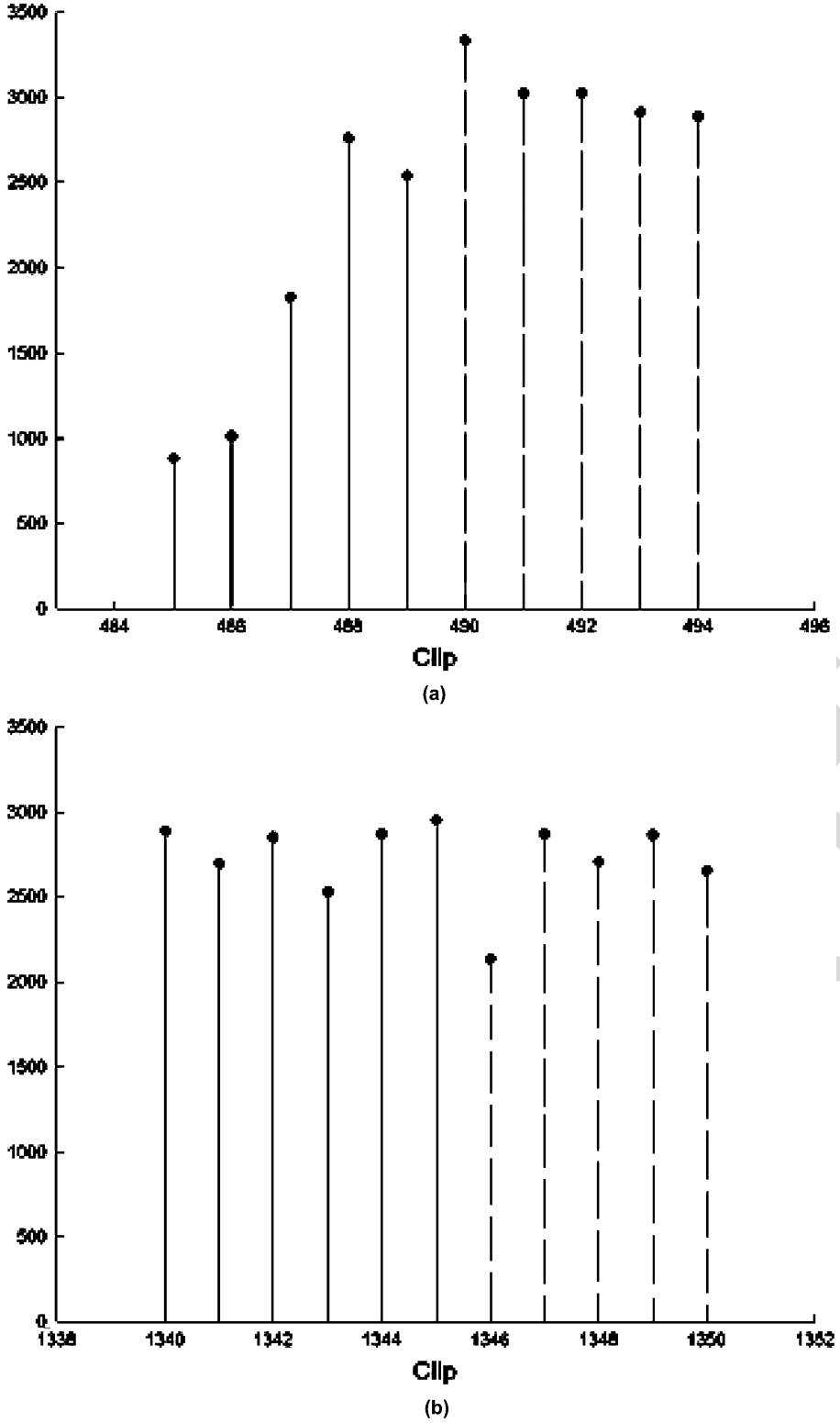

Fig. 3. Typical plots of the clip loudness in a shot pair associated (a) with a goal and (b) a generic no-goal situation. The dashed lines correspond to the second shot.

If the input space $\mathcal{U}$ has cardinality one, say $\mathcal{U} = \{\bar{u}\}$, then the CMC reduces to a standard homogeneous Markov chain with transition probabilities $\{p(s, s', \bar{u}), s, s' \in \mathcal{S}\}$.

In our context, $\mathcal{T}$ represents the P-frame number in the ordered P-frames sequence of a shot pair, and the input $\mathbf{u}$ is introduced to model the occurrence of a shot-cut event. The control set is then defined as $\mathcal{U} = \{0, 1\}$ with the understanding that if a shot-cut event happens at P-frame t, then $\mathbf{u}(t) = 1$, otherwise $\mathbf{u}(t) = 0$.

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, i.e.,

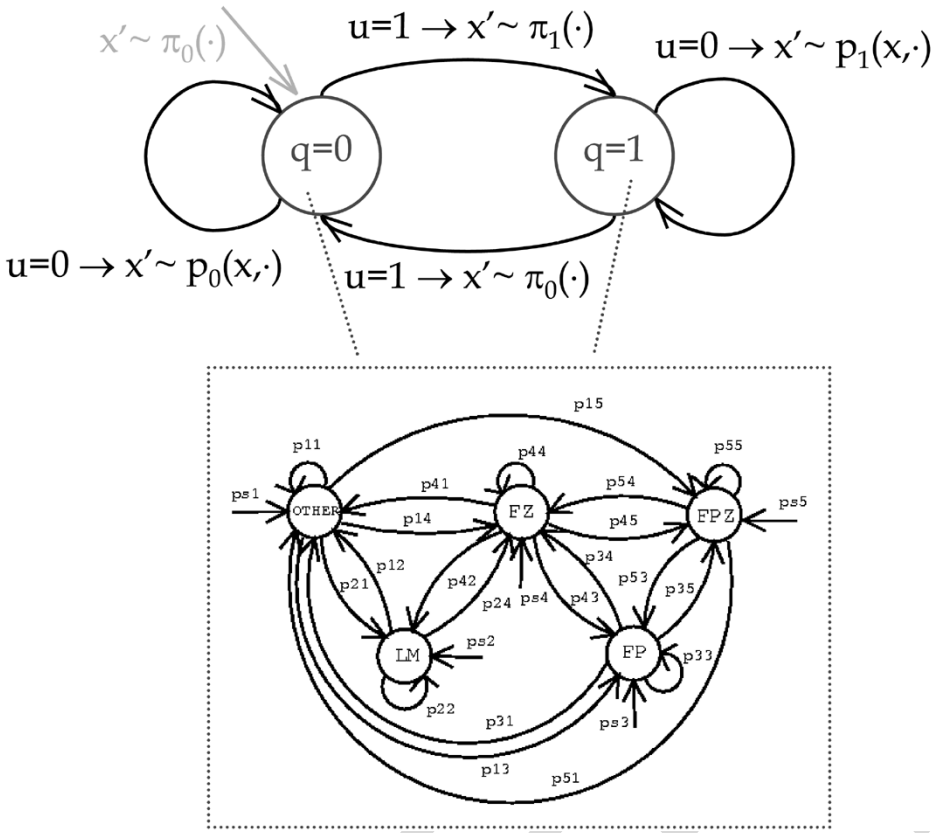

Fig. 4. Schematic representation of the adopted CMC model.

$\mathbf{s}(t) = (\mathbf{x}(t), \mathbf{q}(t)) \in \mathcal{S} = \mathcal{X} \times \mathcal{Q}$, where $\mathbf{q}(t) \in \mathcal{Q} := \{0, 1\}$ is the operative mode of the system and $\mathbf{x}(t)$ is the state of the P-frame observed at time t. In particular, $\mathbf{x}(t)$ can take values in the set $\mathcal{X} = \{\mathrm{LM}, \mathrm{FP}, \mathrm{FZ}, \mathrm{FPZ}, \mathrm{Other}\}$, whose elements 'LM', 'FP', 'FZ', 'FPZ', and 'Other' are defined through the following procedure, based on the visual descriptors introduced in Section III-A1.

Fix a time instant $t \in \mathcal{T}$ and consider the corresponding P-frame. The state variable $\mathbf{x}(t)$ is said to take the value 'LM' if the descriptor "Lack of motion" is equal to 1. If that is not the case, then, $\mathbf{x}(t)$ can take one of the other four values. Specifically, $\mathbf{x}(t)$ is equal to 'FP' if the value of the descriptor "Fast pan" is 1 and that of the descriptor "Fast zoom" is 0. In the opposite case, i.e., when "Fast pan" is equal to 0 and "Fast zoom" is equal to 1, then, $\mathbf{x}(t)$ takes the value 'FZ'. In the case when both the "Fast pan" and "Fast zoom" descriptors are equal to 1, $\mathbf{x}(t)$ assumes the value "FPZ." In all the other cases, $\mathbf{x}(t)$ is said to take the value "Other."

Also, we impose a certain structure on the controlled transition probability function. Specifically, the controlled transition probability function is supposed to satisfy the following condition:

$$p((x,q),(x',q'),u) = 0, \quad \text{if } (u = 1 \text{ and } q = q') \quad \text{or} \quad (u = 0 \text{ and } q \neq q'), \quad \forall x, \quad x' \in \mathcal{X}, q, q' \in \mathcal{Q} \tag{2}$$

which says that a shot-cut event forces the CMC to change operating mode, whereas if no shot-cut event occurs, then the CMC remains in the same mode.

The evolution within a single mode, say $q \in \mathcal{Q}$, is characterized by the set of probabilities

$$\mathcal{P}_q := \{p((x,q),(x',q),0), x, x' \in \mathcal{X}\}. \tag{3}$$

We can then associate with each mode $q \in \mathcal{Q}$ a fictitious Markov chain with state space $\mathcal{X}$ governed by the transition probability function $\{p_q(x,x'), x, x' \in \mathcal{X}\}$, where $p_q(x,x') := p((x,q),(x',q),0), \forall x, x' \in \mathcal{X}$. Here, we suppose that each Markov chain admits a stationary probability distribution and denote by π_q the one associated with mode $q \in \mathcal{Q}$.

When a shot-cut event occurs, then the operating mode of the system changes. As for the state component $\mathbf{x}$, we suppose that it is reinitialized as a random variable with a certain fixed distribution. Specifically, we assume that

$$p((x,q),(x',q'),1) = \pi_{q'}(x'), \quad \forall x' \in \mathcal{X}, q, q' \in \mathcal{Q}, q \neq q'. \tag{4}$$

As a consequence of (2) and (4), the transition probability function is completely characterized by the sets of transition probabilities $\mathcal{P}_q, q \in \mathcal{Q}$, defined in (3).

As for the initial state probability distribution, we assume that at time $t = 0$, when we start observing the system evolution, the system is in mode $q = 0$ and in stationary conditions, i.e.,

$$\mathrm{P}_0((x,q)) = \begin{cases} \pi_q(x), & \text{if } q = 0 \\ 0, & \text{otherwise} \end{cases}$$

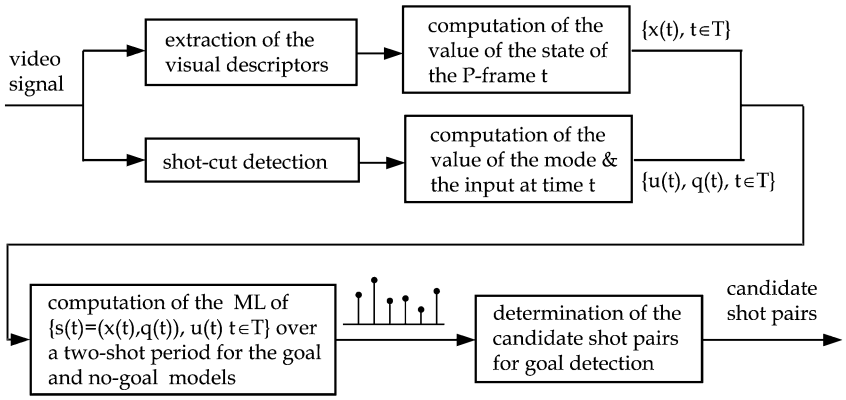

Fig. 5. Block diagram of the video processing step.

$\forall q \in \mathcal{Q}, x \in \mathcal{X}$. In this case, the initial state probability distribution is completely characterized by the set of transition probabilities $\mathcal{P}_0$ defined in (3) with $q = 0$.

A schematic representation of the introduced model is given in Fig. 4. In this figure, the symbol "$\sim$" is used for "distributed according to."

As a consequence of the discussion above, the goal event is associated with a CMC model characterized by the two sets of probability distributions $\mathcal{P}_0$ and $\mathcal{P}_1$ over the state space $\mathcal{X}$, governing the evolution of the component of the state $\mathbf{x}(t)$ within mode $q = 0$ and $q = 1$, respectively.

The no-goal event is associated with further CMC models, each one characterized by the sets $\mathcal{P}_0$ and $\mathcal{P}_1$, trying to capture different no-goal situations where either relevant events such as corner kicks or free kicks occur or no relevant event takes place. \medskip On the basis of these set of CMC models, one can analyze all pairs of shots in a soccer game video sequence and generate a list of shot-pair candidates for goal detection.

The procedure consists of a few steps. First, the sequence of low-level visual descriptors is extracted from the video signal and the shot-cuts are evaluated. Then, each shot pair is analyzed by determining: 1) the sequence of values assumed by the state variable $\mathbf{x}$ (based on the visual descriptors); 2) the sequence of values assumed by the state variable $\mathbf{q}$ and the input $\mathbf{u}$ (based on the detected shot-cut in the shot pair under consideration); and 3) the likelihood of the extracted sequences according to the goal and the no-goal models. If the goal model maximizes the likelihood function, then a goal is likely to be found in the examined shot pair, which is then added to the list of candidate shot pairs for goal detection.

A block diagram of the video processing step is represented in Fig. 5.

We now make precise how the likelihood function is computed for a CMC model.

Consider a shot pair and suppose that it is composed of m P-frames. Let $\{s_i = (x_i, q_i)\}_{i=0}^{m}$ and $\{u_i\}_{i=0}^{m}$ be the sequences of values taken by the state variable and the input, respectively. Then, it is easily seen that the value of the likelihood function can be expressed as follows:

$$\begin{aligned}&\mathrm{P}(\mathbf{s}(t) = s_t, \mathbf{u}(t) = u_t, t = 0, \ldots, m)\\ &= \prod_{t=1}^{m} \mathrm{P}(\mathbf{u}(t) = u_t \,|\, \mathbf{s}(0)\\ &= s_0, \mathbf{u}(0) = u_0, \ldots, \mathbf{s}(t) = s_t) p(s_{t-1}, s_t, u_{t-1})\\ &\quad \times \mathrm{P}(\mathbf{u}(0) = u_0 \,|\, \mathbf{s}(0) = s_0) \mathrm{P}_0(s_0)\end{aligned}$$

where p is the controlled transition probability function and P_0 is the initial state probability distribution. As for the terms $\mathrm{P}(\mathbf{u}(t) = u_t \,|\, \mathbf{s}(0) = s_0, \mathbf{u}(0) = u_0, \ldots, \mathbf{s}(t) = s_t)$, they depend on the "control policy," which is the (possibly probabilistic) rule adopted to decide which control input to apply based on the information on the past values of the state and control variables. Here we suppose that the adopted control policy is the same for all the CMC models. Under this assumption, the model corresponding to the maximum of the likelihood function is not affected by the specific form taken by the control policy.

C. Audio Processing Step

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, 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 analyze 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 III-A3) 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 Fig. 6. The block named "video processing step" is detailed in Fig. 5.

IV. Experimental Results

Experiments were run to test the algorithm for goal detection described in Section III. We considered a few MPEG2 soccer audio-visual sequences and divided them into two sets of sequences: one of about 10 h for estimating the CMC models adopted in the video processing step of the algorithm (training

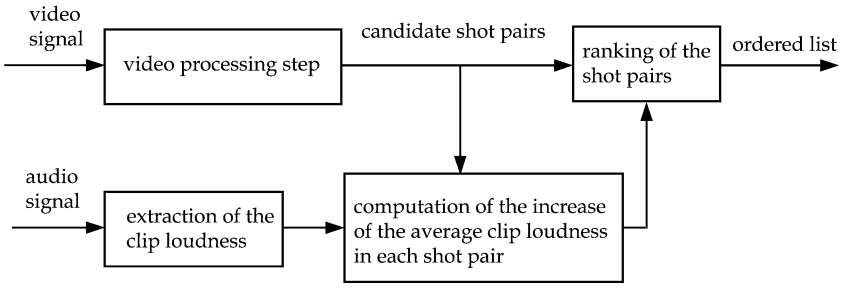

Fig. 6. Block diagram of the goal detection algorithm.

TABLE I
PERFORMANCE OF THE PROPOSED MULTIMODAL ANALYSIS METHOD (LIV-LIP: LIVERPOOL-LIPSIA; SPA-SLO: SPAIN-SLOVENIA; SPA-PAR: SPAIN-PARAGUAY)

	number of goals	number of goals detected	number of shots	number of candidate shot pairs	goals position in the ordered 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-PARA 1	1	1	374	58	11
SPA-PARA 2	3	3	435	77	1, 12, 19

set), and the other of 4.5 h for evaluating the performance of the algorithm (test set).

More specifically, all the shot pairs of the training sequences were manually classified as goal and no-goal pairs, with the no-goal pairs distinguished in different categories (e.g., corner kick, free kick, or plain action). Then, the transition probability sets $\mathcal{P}_0$ and $\mathcal{P}_1$ characterizing the different goal and no-goal CMC models were estimated based on the sequence of values taken by the state variable $\mathbf{x}(t)$ in the first shot $(\mathcal{P}_0)$ and in the second shot $(\mathcal{P}_1)$ of 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 I.

Note that all of the shot pairs where a goal actually occurs are within the first 23 positions in the ordered list, irrespective 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 of the goals with a reduced number of false detections.

The use of the video information only would have lead to a significantly higher number of false detections, as it can be seen in Column 4 of Table I.

In Table II, we report the results obtained by ranking all shot pairs on the sole basis of the audio clip loudness without first using the video-based candidate selection procedure. Note that the shot pairs where a goal occurs drops down significantly in the ordered list, thus confirming the importance of modeling with a CMC the information contained in the video signal and using it jointly with the audio information.

TABLE II
PERFORMANCE OF THE ALGORITHM WHERE ALL SHOT PAIRS ARE ORDERED BASED ON THE AUDIO SIGNAL (LIV-LIP: LIVERPOOL-LIPSIA; SPA-SLO: SPAIN-SLOVENIA; SPA-PAR: SPAIN-PARAGUAY)

	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

V. CONCLUSION

In this paper, we have presented a semantic indexing algorithm which uses both audio and visual information for event detection in soccer programs. We focused in particular on the detection of goals, which represents the key event in a soccer game. The proposed algorithm consists of a two-step procedure. 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 both a goal and a no-goal event is supposed to be governed by a controlled Markov chain. This allows to determine a list of those video

segments where a goal is likely to be found, based on the ML criterion. The audio information is then analyzed to rank the video segments in the list. Specifically, we evaluate the "loudness" associated with each video segment identified by the analysis carried out on the video signal. The intensity of the "loudness" is then used to order the video segments. Consequently, the segments associated to goals appear in the first positions of the ordered list. Experiments show that the number of false detections obtained by using the proposed multimodal approach is significantly smaller than that obtained by using either the audio or the visual information only. This confirms the intuition that in soccer programs each signal (audio and video) carries its peculiar information, which can be effectively exploited for semantic indexing by adopting an appropriate multimodal approach. Current research focuses the attention on more sophisticated properties of the audio signal, in order to further reduce the number of false detections. Moreover, we are considering the possibility to extend the proposed approach to other types of sport video programs.

REFERENCES

[1] R. Lagendijk, "A position statement for panel 1: Image retrieval," in *Proc. VLBV99*, Kyoto, Japan, Oct. 29–30, 1999, pp. 14–15.

[2] **[AUTHOR: PLEASE PROVIDE THE MONTH OF ISSUE.—ED.]**T. Zhang and C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 441–457, 2001.

[3] **[AUTHOR: PLEASE PROVIDE THE MONTH OF ISSUE.—ED.]**Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using audio and visual information," *IEEE Signal Processing Mag.*, vol. 17, pp. 12–36, 2000.

[4] C.-G.-M. Snoek and M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-Art,", vol. 2001–20, ISIS Tech. Rep. Series, Dec. 2001.

[5] S.-F. Chang, "The holy grail of content-based media analysis," *IEEE Multimedia*, vol. 9, pp. 6–10, Apr.–June 2002.

[6] D. Zhong and S.-F. Chang, "Structure analysis of sports video using domain models," in *Proc. ICME'2001*, Tokyo, Japan, Aug. 2001, pp. 920–923.

[7] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. ACM Multimedia*, Los Angeles, CA, 2000, pp. 105–115.

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

[9] **[AUTHOR: PLEASE PROVIDE PAGE NUMBERS.—ED.]**S. Lefevre, B. Maillard, and N. Vincent, "3 classes segmentation for analysis of football audio sequences," in *Proc. ICDSP'2002*, Santorini, Grece, July 2002.

[10] **[AUTHOR: PLEASE PROVIDE PAGE NUMBERS.—ED.]**M. Han, W. Hua, W. Xu, and Y. Gong, "An integrated baseball digest system using maximum entropy method," in *Proc. ACM Multimedia*, France, Dec. 2002.

[11] **[AUTHOR: PLEASE PROVIDE PAGE NUMBERS.—ED.]**M. Petrovic, V. Mihajlovic, W. Jonker, and S. Djordievic-Kajan, "Multi-modal extraction of highlights from TV formula 1 programs," in *Proc. ICME'2002*, Lausanne, Switzerland, Aug. 2002.

[12] R. Leonardi, P. Migliorati, and M. Prandini, "A Markov chain model for semantic indexing of sport program sequences," in *Proc. WIAMIS'03*, London, U.K., Apr. 2003, pp. 519–528.

[13] ——, "Semantic indexing of sports program sequences by audio-visual analysis," in *Proc. ICIP'03*, Barcelona, Spain, Sept. 2003. to appear.

[14] A. Bonzanini, R. Leonardi, and P. Migliorati, "Event recognition in sport programs using low-level motion indices," in *Proc. ICME*, Tokyo, Japan, Aug. 2001, pp. 920–923.

[15] R. Leonardi and P. Migliorati, "Semantic indexing of multimedia documents," *IEEE Multimedia*, vol. 9, pp. 44–51, Apr.–June 2002.

[16] R. Leonardi, P. Migliorati, and M. Prandini, "Modeling of visual features by Markov chains for sport content characterization," in *Proc.*, Toulouse, France, Sept. 2002, pp. 349–352.

[17] Y. Gong, L.-T. Sin, C.-H. Chuan, H. Zhang, and M. Sakauchi, "Automatic parsing of TV soccer programs," in *Proc.* , Washington, DC, May 1995.

[18] **[AUTHOR: PLEASE PROVIDE PAGE NUMBERS.—ED.]**D. You, B.-L. Yeo, M. Yeung, and G. Liu, "Analysis and presentation of soccer highlights from digital video," in *Proc. ACCV*, Singapore, Dec. 1995.

[19] 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*, Tokyo, Japan, Aug. 2001, pp. 928–931.

[20] **[AUTHOR: PLEASE PROVIDE PAGE NUMBERS.—ED.]**L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models," in *Proc. ICASSP*, Orlando, FL, May 2002.

[21] A. Bonzanini, R. Leonardi, and P. Migliorati, "Semantic video indexing using MPEG motion vectors," in *Proc. EUSIPCO*, Tampere, Finland, Sept. 2000, pp. 147–150.

[22] V. Tovinkere and R. J. Qian, "Detecting semantic events in soccer games: Toward a complete solution," in *Proc. ICME*, Tokyo, Japan, Aug. 2001, pp. 1040–1043.

[23] **[AUTHOR: PLEASE PROVIDE PAGE NUMBERS AND THE CITY.—ED.]**A. Ekin and M. Tekalp, "Automatic soccer video analysis and summarization," in *Proc. SST SPIE03*, CA, Jan. 2003.

[24] P. Migliorati and S. Tubaro, "Multistage motion estimation for image interpolation," *Signal Processing: Image Commun.*, vol. 7, pp. 187–199, July 1995.

[25] I. Koprinska and S. Carrato, "Temporal video segmentation: A survey," *Signal Processing: Image Commun.*, vol. 16, no. 5, pp. 477–500, Jan. 2001.

[26] Y. Deng and B. S. Manjunath, "Content-based search of video using color, texture, and motion," in *Proc. ICIP*, Santa Barbara, CA, Oct. 26–29, 1997, pp. 534–536.

[27] M. L. Puterman, *Markov Decision Processes*. New York: Wiley, 1994.

Riccardo Leonardi (S'80–M'83) received the Diploma and Ph.D. degrees in electrical engineering from the Swiss Federal Institute of Technology, Lausanne, in 1984 and 1987, respectively.

After one year at the University of California, Santa Barbara, as a Postdoctoral Fellow, he joined for three years AT&T Bell Laboratories as a Member of Technical Staff, performing research activities on visual communication. In March 1992, he joined the University of Brescia, Brescia, Italy, to lead research and teaching in the field of telecommunications, where he holds the Signal Processing Chair. His main research interests cover the field of digital signal processing applications, with a specific expertise on visual communications, and content-based analysis of audio-visual information. He has published more than 100 papers on these topics. He was responsible for setting up the Ph.D. program in information engineering at the University of Brescia, which has been active since November 1994. Since 1997, he has acted as an evaluator and auditor on several European Commission RTD programs.

Pierangelo Migliorati (M'96) received the Laurea (*cum laude*) degree in electronic engineering from Politecnico di Milano, Milan, Italy, in 1988, and the M.Phil. degree in information technology from the CEFRIEL Research Centre, Milan, in 1989.

From 1989 to 1995, he was with the CEFRIEL Research Center as Researcher on the Technical Staff. From 1989 to 1993, he worked in the Digital Signal Processing Area, focusing mainly on motion compensated image interpolation. From 1993 to 1995, he led the Transmission Systems Area, focusing on nonlinear channel modeling and equalization. Since 1995, he has been an Assistant Professor with the Department of Electronics for Automation, University of Brescia, Brescia, Italy, where he is interested in content-based indexing and transmission of multimedia documents. His teaching activities focus mainly on digital signal processing and digital communications.

Maria Prandini (M'00) received the Laurea degree in electrical engineering from the Politecnico of Milano, Milan, Italy, in 1994 and the Ph.D. degree in electrical engineering from the University of Brescia, Brescia, Italy, in 1998.

From March to July 1998, she was a Visiting Scholar with the Delft University of Technology, Delft, The Netherlands. From August 1998 to July 1999, she was a Visiting Postdoctoral Researcher with the Electrical Engineering and Computer Sciences Department, University of California, Berkeley. She is currently an Assistant Professor with the Politecnico of Milano. Her research interests include identification and adaptive control of stochastic systems, statistical learning theory for system analysis and design, coordination and control of multi-agent systems, air traffic management, and verification of hybrid systems.