Anil Kokaram, Niall Rea, Rozenn Dahyot, A. Murat Tekalp,
Patrick Bouthemy, Patrick Gros, and Ibrahim Sezan

© DIGITALVISION, © ARTVILLE (CAMERAS, TV, AND CASSETTE TAPE) © STOCKBYTE (KEYBOARD)

Browsing Sports Video

Trends in sports-related indexing and retrieval work

The case for automated video information access systems has been made repeatedly by numerous researchers since the early 1990s. Finding important episodes in nontext media like video and audio is made difficult because the same material means different things to different people. Therefore, manually tagging data with textual keywords as a means of facilitating access is not necessarily reliable for this type of media. This is a crucial point. Information retrieval as a task is really only defined in the context of the user. This article considers one such context: sports media. In that context, there is a salient structure to the media as well as a series of definable user requirements. Sport broadcasting is certainly of high commercial importance [1]. All national and international news broadcasts contain specific regular segments devoted to sports. Consumers themselves are increasingly acting as generators of sport media as it becomes easier for parents and coaches to record school sports games, for instance.

From a broadcaster's point of view, efficient archiving of media facilitates later reuse in the creation of highlights packages or specialized DVDs. Manual keywording of events is feasible but time consuming and

1053-5888/06/$20.00©2006IEEE

also cumbersome for games archived long ago. The consumer, meanwhile, is faced with a large quantity of sports channels, and the usefulness of recording is limited in the sense that the whole event must be manually browsed for creating a semblance of a highlights package. Even more interesting, as broadcasters seek to cover more and more sports events to fill available program time, there is less manpower available to create highlights packages. Therefore, the consumer often has no option but to manually browse an event to create a virtual highlights effect. The longer the game, the more difficult manual browsing becomes.

These difficulties have resulted in growing research interest in automated content-based video analysis targeted at sports events. Since it is more natural for human operators to operate at a semantic level, algorithms that can understand how low-level features relate to semantic concepts must be created. Within the sports context, the semantic gap can be bridged since an understanding of the structure of the event and the user are both attainable. Hence, usable automated sports media access is feasible.

This article sets out to expose trends in sports-based indexing and retrieval work. In so doing, it discusses the essential building blocks for any semantic-level retrieval system and acts as a case study in content analysis system design. While there has been a huge amount of activity in sports retrieval, and this article cannot attempt to be exhaustive, the reader is given an organized snapshot of current work in the area.

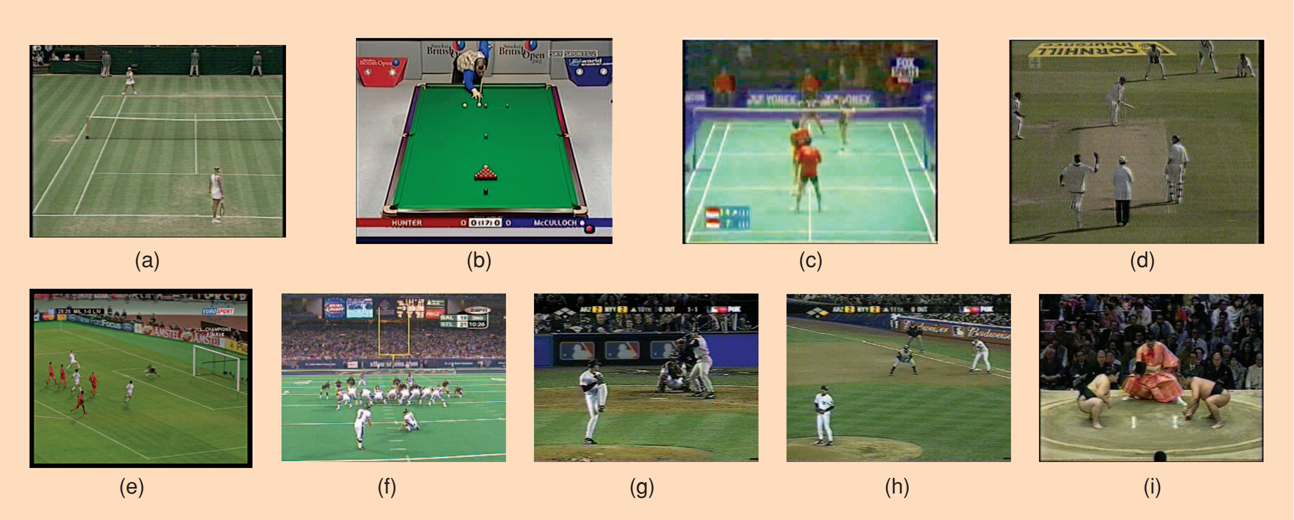
A STRUCTURE FOR SPORTS MEDIA

An overview of previous work and a consideration of the manner in which different sports are broadcasted show that two broad classes of sports can be identified: court/table sports (e.g., badminton, tennis, table tennis, snooker/pool, volleyball, and basketball) and field sports (e.g., soccer, football, baseball, cricket, and hockey). Surprisingly, this is not a useful classification in terms of content analysis.

In sports like tennis, badminton, cricket, and snooker, there is one camera view that dominates the broadcast (e.g., the full court in tennis) and contains almost all the semantically meaningful elements required to analyze the game. Yet in sports like football, soccer, baseball, hockey, basketball, and volleyball, there is no such well-defined camera view that dominates the course of play. In football, for instance, the full-field shot is almost never shown and camera views tend to flow from position to position showing portions of the field in which action is ongoing. Since it is the image representation that is subject to analysis, sports analysis systems tend to reflect these distinctions in their design. These classes can be defined as dominant semantic view (DSV) and multiple view semantics (MVS) sports. Figure 1 illustrates this idea with frames from different events.

An alternate viewpoint is generated by considering that games are generally time driven or point driven. Sports like football, basketball, and soccer are time driven in the sense that there is a loose structure in the progress of the game punctuated only by a few time periods e.g., four in basketball and two in soccer. In these sports, high-information events occur sparsely and randomly. In sports like tennis, snooker, and cricket, though, the progress of the game is point driven. In those sports, the structure of the game is highly formulaic and based around regular events and stylized actions that yield points. This implies that for point-driven games, it is likely that a high level of semantics can be extracted up to the very real possibility of generating a table of contents automatically. However, for time-driven games, it is more difficult to access semantics. The point-driven sports are typified by strong camera and domain models and are likely to be DSV, while the time-driven sports are MVS.

There is, of course, a wide variety of athletic sports (e.g., swimming, track and field, shooting, and skiing) that on first considera-



[FIG1] Typical content-rich views from a selection of sports, showing the playing area that is usually targeted for detection. (a) DSV sports: (a) tennis, (b) snooker, (c) badminton, and (d) cricket. MVS sports: (e) soccer, (f) American football, (g)–(h) baseball, and (i) sumo. Almost all the game semantics in DSV sports can be extracted via playing area detection and motion estimation in the views shown. Event detection for both types of sports can be made by analyzing edit sequences. A base-steal attempt in baseball (especially from the first base) is typically captured from a special camera angle, and thus an algorithm is able to distinguish it from a regular pitch.

tion appear to contain elements of both classes above. On a closer examination however, they are generally DSV sports and are typified by stylized movement as well as heavily structured broadcasting.

FEATURE EXTRACTION

Most games tend to employ rules about the play within a playing area. For instance, in snooker play takes place on a green table. In tennis, soccer, and football, the ball must be kept within set bounds. All sports analysis systems therefore tend to contain some element of playing area localization. Both camera motion and local player or object motion are also key information bearing features. Camera motion can be used to make inferences about the direction of play in soccer and cricket, while local motion of players and the ball are perhaps the most vital semantic features. The temporal broadcast format also contains information that can be exploited. For instance, replays tend to follow important events in the game, so detection of replays helps to infer those events. More sophisticated editing structure can also be exploited. Even though this kind of information tends to be very broadcaster, and even culturally, specific, it is still worth considering as it could shortcut much of the pain of object understanding. It is the extraction of this kind of information that motivates the choice of features for sports analysis systems.

VIEW CHARACTERIZATION

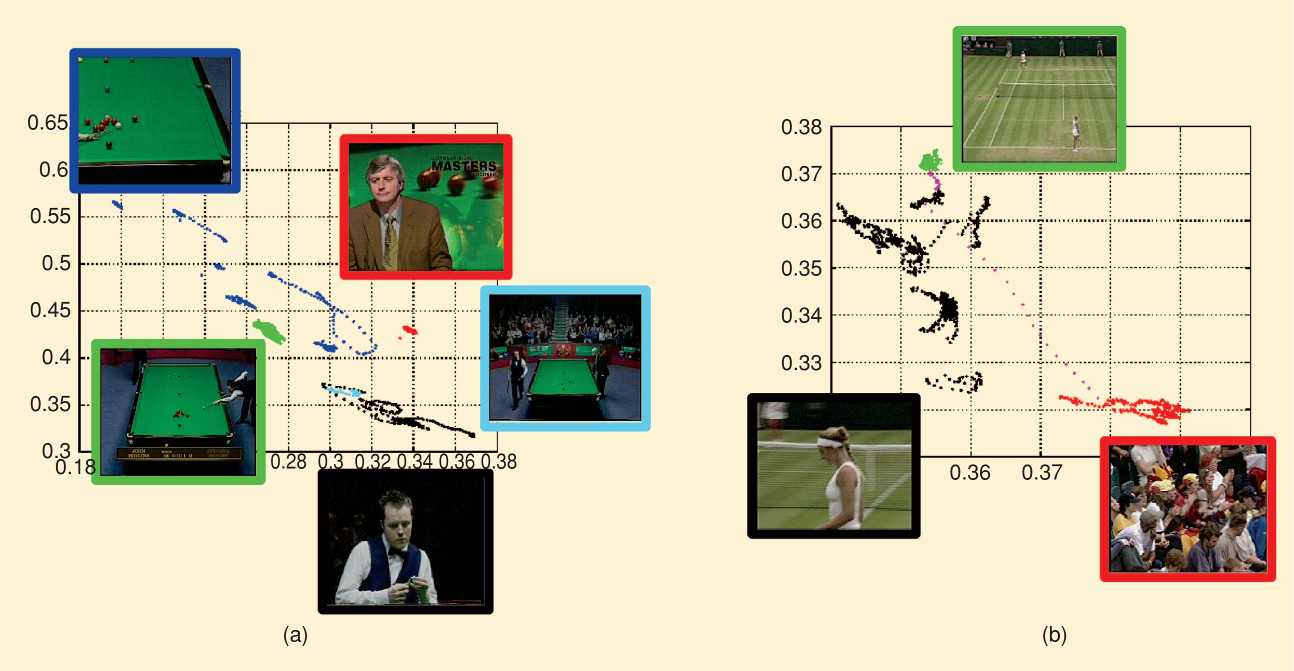
All sports analysis system design begins with the characterization of the views available and the identification of those views that contain the most information [2]–[8]. Classifying views allows information to be extracted more efficiently by allowing different processes to analyze each view. For instance, a close-up view would yield player identity information more readily than a long-field view, which in turn may yield game-related player motion more readily.

In court sports, views in which the entire playing area is visible contain almost all the semantic information of the game. Therefore, those analyzing court sports tend to attempt to parse the sequence in terms of those views. These kinds of shots in tennis and snooker have been called global views [6]. Field sports contain different kinds of coverage. Xu et al. [9] label shots in soccer as close up, medium, or long shot according to ratio of grass-colored pixels in the shot.

Classifying the shot type in sports can be achieved in court sports simply by locating the playing area. It is also feasible to model the temporal evolution of low-level image features (e.g., dominant color or shape moments [4]) across a shot. Denman et al. [3] use the hidden Markov model (HMM) to parameterize the temporal evolution of shape descriptors for each frame across a shot (see Figure 4). Babaguchi et al. [10] detect live and replay views by comparing dominant color distribution of the key frames and by calculating the count of field lines.

THE PLAYING AREA

Figure 1 shows a single frame of the playing area from a selection of sports as they would appear in broadcast footage. In most sports, the color content of frames containing the playing surface differs significantly from other views. Figure 2 illustrates



[FIG2] A scatter plot of the average color of each frame from snooker and tennis broadcasts. The vertical axis is normalized red $r/(r+g+b)$ and the horizontal is normalized green $g/(r+g+b)$. An example frame from each type of shot is given a colored border that corresponds to the color used for plotting the points corresponding to that shot. In general, the clusters show a marked correlation with view type, but the principal content rich frames in each sport lie within a well organized (Gaussian), compact cluster indicated in green. This justifies the use of color as a key component in many sports analysis systems.

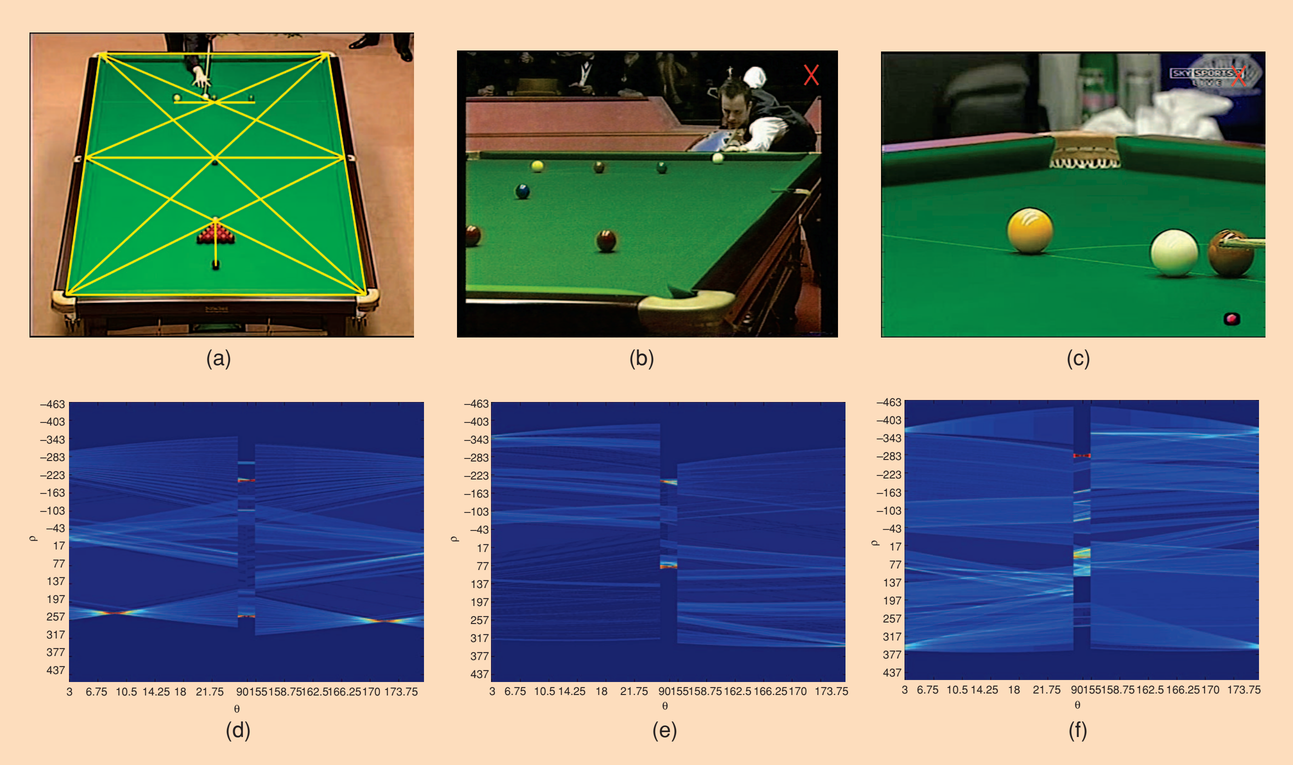
this point. Detection of a dominant color therefore leads to a simple mechanism for isolating shots of the playing area. Ekin et al., for instance, use the modes of the color distribution [8] to find shots containing the playing area in soccer, while Dahyot et al. [4], [11] employ other statistics like mean and variance for snooker and tennis.

[THE CASE FOR AUTOMATED VIDEO INFORMATION ACCESS SYSTEMS HAS BEEN MADE REPEATEDLY BY NUMEROUS RESEARCHERS SINCE THE EARLY 1990S.]

In most sports, the playing area needs to be calibrated in some way, i.e., by finding court markings. For DSV sports, this is key for semantic content extraction since many events are correlated to player position. Sudhir et al. [12] were perhaps the first to detect/analyze the exact court geometry in each frame of tennis using edge and corner detection combined with a three-dimensional (3-D) camera model. A much simpler approach was applied by Kijak et al. [6] in which the Hough transform was used to locate lines in tennis. A generic mechanism for summarizing shape was developed by Denman et al. [3] and used in tennis and snooker. The idea was that, given a simple dominant color-based segmentation of each frame, the shape of that segmentation mask completely characterizes the view in sports. Figure 3 shows the significant difference between the Hough transform of different views of the segmented table in snooker. The geometric moment of the Hough/Radon transform therefore changes with the view. Figure 4 shows how the eighth order moment changes with time in a sequence. It is clear that different shots, and particularly the table view, can be characterized this way.

Given the detection of frames in which the playing area is in view and the lines delineating the playing area, it is then possible to calibrate any rectangular area in the view plane. This can be done without the need for 3-D models and is based on the knowledge that the diagonals of a rectangle intersect at the center regardless of the view angle [3]. Figure 5 shows the successful detection of the playing area for several types of games.

In soccer, Ekin [8] identifies three classes of playing area geometry: 1) long shots, 2) in-field medium shots, and 3) out-of-field and close-up shots. The ratio of grass-colored pixels (hue values between 65 and 85°) to all pixels in a frame is used to characterize each type (see Figure 5). Because only spatial features are used, a shot type can ideally be determined from a single key frame. In some applications, only detection of long shots may be of interest, so the problem reduces to two-class pattern recognition. Although it is usually correct to assume that a long shot contains more grass

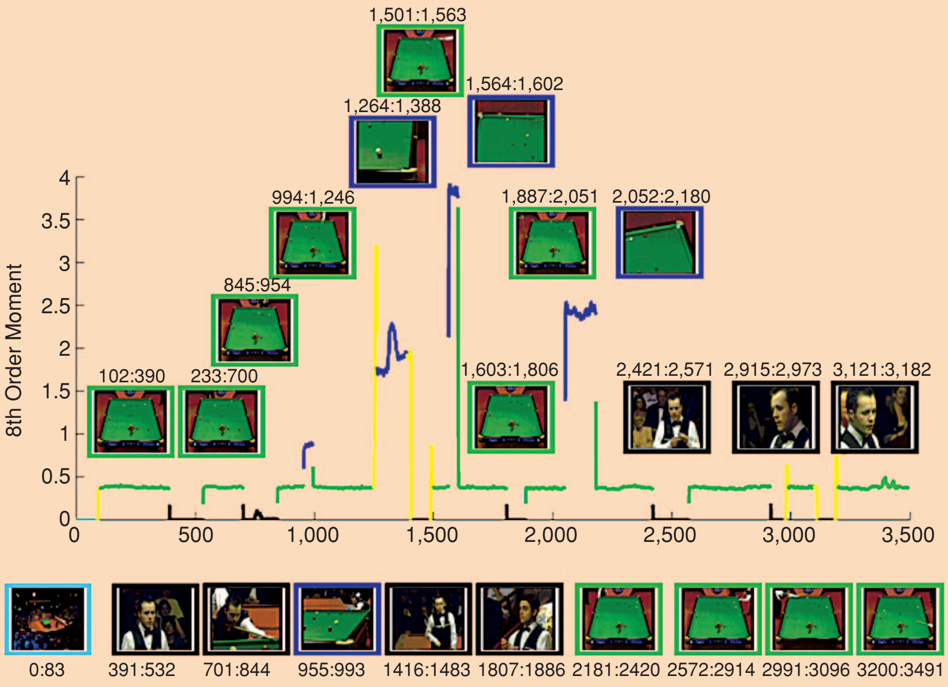


[FIG3] **View detection by implicit scene geometry. (a) Different views of the snooker table during play. (b) Hough transform of the segmentation mask (indicating all green areas) of respective views showing selected areas of the transform corresponding to the a priori knowledge of rough table geometry. Each Hough transform is distinctive and so can be used to classify views (see Figure 4).**

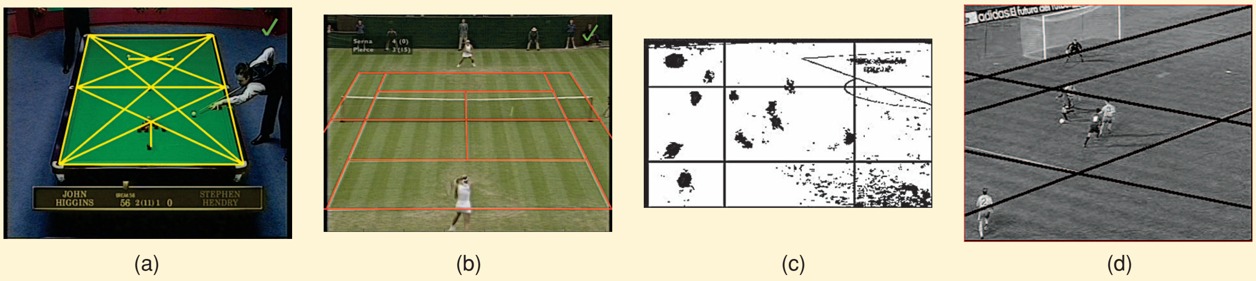
pixels than a medium shot, it is generally difficult to find a clear threshold value to separate medium and long shots. Therefore, other distinguishing features are considered. One such feature is the object size and count in each shot. In long shots, many players cover only a small portion of the field, while in a medium shot, only a couple of players appear in the shot and they occupy most of the frame area.

MOTION

It is through the motion of the camera that the essential game elements are kept in view. It therefore stands to reason that the motion of the camera contains semantic information. In cricket, Kokaram et al. [13] show that the motion of the camera can be used to detect when a play is about to start, the duration of the play, and the direction of the ball after it is hit. Action in soccer [14] can also be characterized in this manner. Local motion information contains the motion of the players and, hence, is directly relevant to the play. Local/global motion segmentation can be crudely achieved by citing areas with large motion-compensated frame difference (after global motion estimation) as local motion. Local motion, and in particular motion activity, can then also be used as a powerful feature for sequence classification [15].



[FIG4] 8th order geometric moment for snooker footage showing keyframes corresponding to each view episode. The moment varies with view, and can be modeled with two state HMMs. Different HMMs characterize different kinds of shots (e.g., player close up, corner pocket view and so on). The HMM that best describes the temporal feature evolution therefore identifies the shot type. Classifying shots in this way is made feasible only through the use of shot detection processes that separate shots using shot cut or special effect detectors.



[FIG5] Calibration of playing area in snooker, tennis, and soccer. (a) and (b) show superimposed lines on a snooker table and a tennis court after correct calibration in the field of view without 3-D information (after Denman et al.). (c) and (d) show different views in soccer indicating playing area detection (grass pixel detection) and delineation of the field (after Ekin et al. [3]). The view in soccer is first classified as long, medium, or close-up using the grass color ratio in the quadrants indicated.

The motion of players and the ball is crucial for an understanding of any game. Object/player identification is typically achieved using some contextual color-based segmentation [8], [12], [16]. Subsequent tracking/position information is then possible either simply by identifying the same object in successive frames (e.g., [12]) or explicit tracking (e.g., [8], [16], [17]).

In court sports, explicit player and object tracking is instigated after playing area localization. Color segmentation is typically used to detect the object once the relevant view is confirmed. Having located the object, tracking can begin. Both Ekin [18] and Rea et al. [16], [17] have proposed schemes based on tracking color histograms. Ekin employed a deterministic strategy that operated on a grid of blocks in each frame, tracking being performed by connecting blocks yielding close matches to the prototype histograms of color for the player and referee. Rea et al. employed a framework based on the particle filter. They were also able to incorporate information about perspective (extracted from the previous step of playing area delineation) to improve tracking robustness (see Figure 6). Pitié et al. [19] have recently proposed a new scheme based on presegmentation of candidate locations for objects.

AN ALTERNATE VIEWPOINT IS GENERATED BY CONSIDERING THAT GAMES ARE GENERALLY TIME DRIVEN OR POINT DRIVEN. SINCE IT IS MORE NATURAL FOR HUMAN OPERATORS TO OPERATE AT A SEMANTIC LEVEL, ALGORITHMS THAT CAN UNDERSTAND HOW LOW-LEVEL FEATURES RELATE TO SEMANTIC CONCEPTS MUST BE CREATED.

SOUND

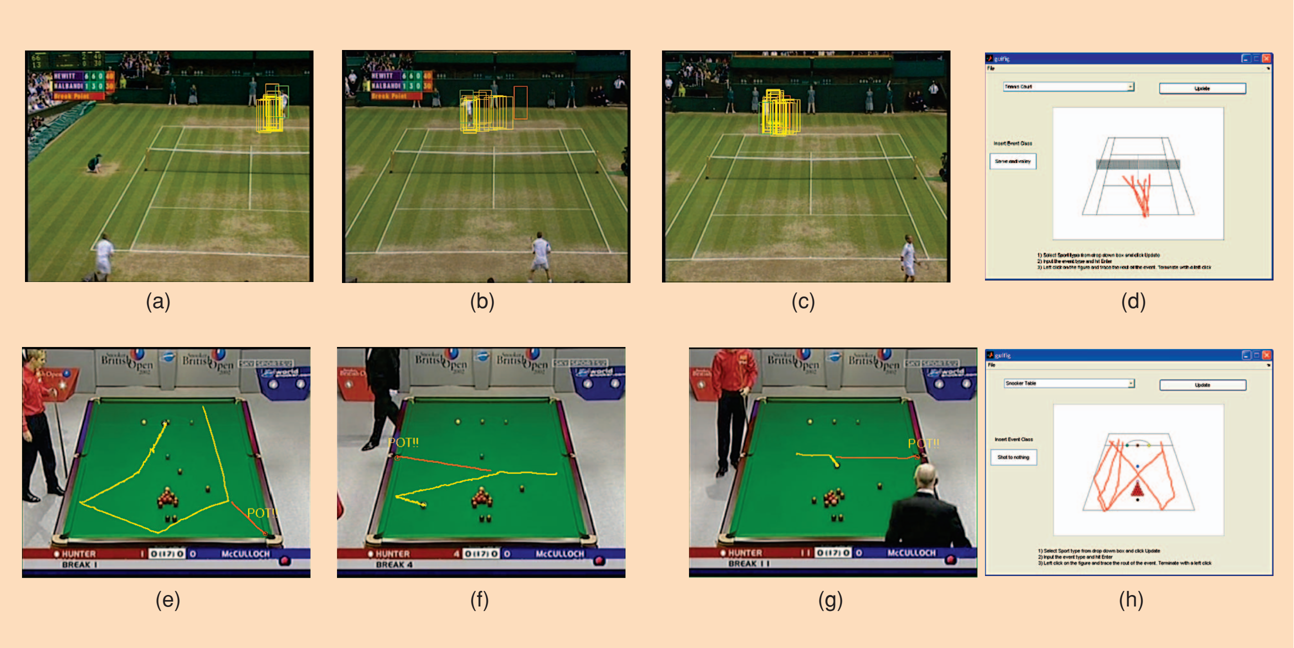
Audio features extracted from sports events can be connected with the excitement level of the audience (e.g., whistling [20]), characteristics of the energy envelope, and loudness [6], [14]. More sophisticated processing employs cepstral coefficients [21] and the normalized power spectrum [22].

TEXT

Text is available from closed captions [10] during broadcasting and can be used for extracting score and player information. Of more interest is the rich source of live textual commentary that is associated with many sports, including cricket, tennis, and baseball. Cricket in particular is often broadcasted over the Internet in textual form, detailing all player actions. This kind of commentary could be exploited at the semantic level of game understanding but has yet to be fully explored.

LOW-LEVEL EVENT DETECTION

The next stage of analysis in most proposed sports systems combines extracted features to detect events or objects that are semantically relevant. Camera views themselves are often considered to be events since a particular temporal sequence of views is indicative of certain semantics, e.g., goals or replays.



[FIG6] (a) Tennis player. (b) Snooker ball tracking. In both cases, a particle filter tracker is instantiated after court calibration and player detection. It uses the color distribution of the object for tracking. The images in (a), (b), and (c) show the detected bounding boxes of the player as he moves through the court. A red box is the first position, and a green box is the last. The end picture in both rows shows a GUI used to train an HMM to detect the play that is occurring by modeling the tracks. In tennis, for instance, aces and baseline rallies can be detected while in snooker, "snookers" can be detected.

Direct low-level event detection may take the form of detection of object disappearance (e.g., in snooker/pool [3]) or detection of impacts by fusing audio and video data [11], [22]. Impact detection (Figure 7) can be immediately connected with semantic events in tennis (racket hits) and cricket (batsman shots) for example.

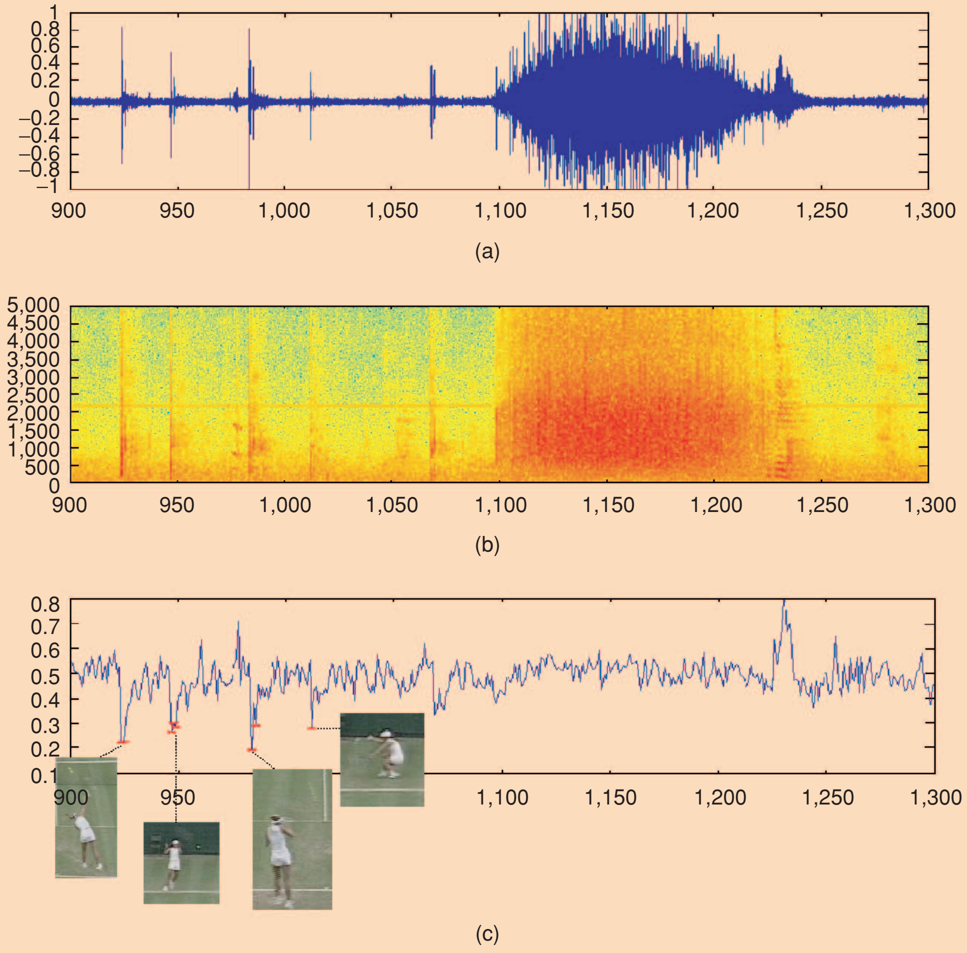
Replays in sports broadcasts are indicators of important events in the game. Replay detection techniques exploit editing effects [23], detection of frame repetitions (via interframe differences) [24], and compressed domain manipulation. Detection in the compressed domain uses the fact that when frame repeats occur during slow motion, the direction of frame prediction changes. Furthermore, the bit rate increases rapidly when a frame with new content follows repeated frames.

BRIDGING THE SEMANTIC GAP

It is in the last stage of content analysis that events from the previous level can be articulated to convey semantic meaning at higher levels of the information pyramid. Figure 8 illustrates the hierarchy of information in a tennis match as an example. Positions, player/ball motion, or particular sequences of events can be used to infer high-level semantics [16], [25], [26]. Sports is a unique domain in the sense that many low-level events correspond immediately with elements of game semantics. For instance, the racket hit detector discussed above can be used immediately to detect rallies.

Motion is a key feature here. In [13], a shot in cricket was detected by a strong horizontal pan in the camera motion, while in [12] and [16], the motion of objects themselves is used to classify an entire play sequence in tennis and snooker. Statistical modeling of motion content has also been proposed for classification of sequences in different sport videos, such as skating and athletics [15].

Scene sequencing, however, depends crucially on the broadcasting edit style and has been exploited in [5], [7], [8],



[FIG7] (a) An audio signal from a tennis rally of five racket hits followed by crowd noise. (b) The spectrogram of the signal is computed on 40-ms windows with 20-ms overlap. PCA analysis on the spectrum of prototype racket hits yields a cloud of points in frequency space indicating the class of importance. (c) The distance of each spectral window from that space. Four racket hits are detected by thresholding that distance. The corresponding images (zoom) are presented below the distance plot. The ball and the racket in rapid motion are not always visible. On the fifth racket hit, the player lost the point because the ball could not be returned properly; hence, the sound was atypical and it was not detected.

and [27]. In baseball, for instance, based on whether or not a pitch is immediately followed by a camera break that results in a field scene, it is possible to distinguish a pitch with field action (a hit in most cases) from a pitch without field action (a ball or a strike in most cases) [28]. It is sensible to presume that analyzing intrinsic motion of objects is a more robust approach to the problem. However, given the computational cost of motion estimation in general, scene sequencing can be used effectively for lower-cost practical systems.

ALL SPORTS ANALYSIS SYSTEM DESIGN BEGINS WITH THE CHARACTERIZATION OF THE VIEWS AVAILABLE AND THE IDENTIFICATION OF THOSE VIEWS THAT CONTAIN THE MOST INFORMATION.

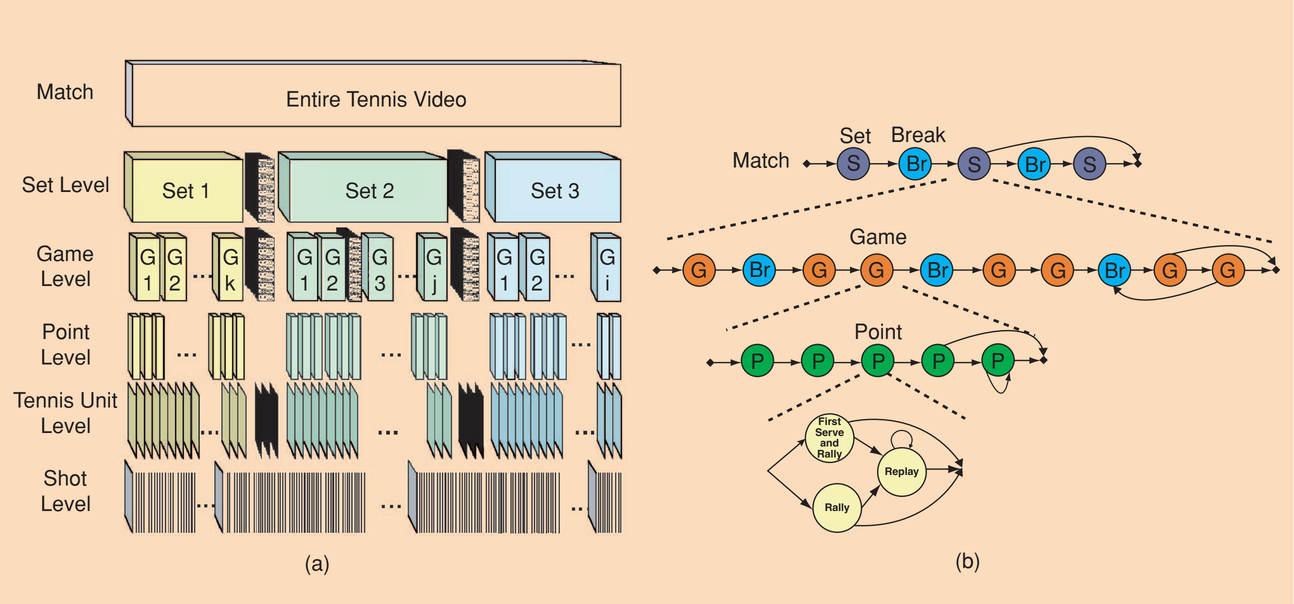
Affective content analysis (measuring the emotional reaction of the audience) of these semantic-level events can bring another dimension to content-based access [29], [30]. The article by Hanjelic et al. gives a good overview of this topic.

In general, though, the use of machine learning frameworks is the only feasible approach to extracting game semantics. The HMM has been the most popular tool adopted in this area. It has been used both for modeling the temporal evolution of low-level features for view classification (e.g., Dahyot/Denman et al. [3], [4], [22] (see Figure 4), Li et al. [31], Xie et al. [32], [33]) and for modeling the temporal succession of events for different types of game semantics, e.g., Kijak et al. [6] (shot sequence modeling) and Rea et al. [16], [34] (object motion modeling) (see Figure 8). In essence, the technique requires that an HMM be created and trained for each possible game semantic, e.g., a point in tennis or a play in baseball. The appropriate semantic is extracted by choosing the model that maximizes the likelihood of an observation sequence among the set of all models available. Figure 6 shows a GUI used to train an HMM in tennis and snooker for object-level motion analysis.

Independently generated rich textual game data may be synchronized with video event segments to enable semantic access. SportsTicker (www.sportsticker.com) is a textual service covering professional and college sports in the United States. Li et al. [28] consider automatic synchronization of the textual play-by-play data for football and baseball. The algorithm uses dynamic programming to articulate text and video features to achieve synchronization.

SPORTS VIDEO SUMMARIZATION

The goal of most of the work cited thus far was to create a summary of a sports event useful for streaming media to low bandwidth devices [3], [14], [18], [20]. Each low-level event detected (e.g., racket hit in tennis, ball pot in snooker, goal in soccer) can yield an automated index for a sports game. Summaries can either be condensed or selective representations of a game. The condensed summary attempts to give a well-balanced summarized view of the whole video content. It requires a hierarchical temporal segmentation of the video and usually applies to sports exhibiting a clear structure, regarding both their own rules (typically, point-driven sports like tennis)



[FIG8] (a) The information pyramid for tennis. Feature extraction and low-level event detection yield information in the bottom two levels. Inference at the higher semantic level is typically attained by temporal modeling of the evolution of features and events, generally with an HMM. (b) A typical HMM hierarchy for tennis, modeling evolution of shot events (rally, service) at the bottom and higher-level events at the top (points, breaks, etc).

and broadcasting style. In contrast, the selective approach targets sequences that convey by themselves the highest interest (to be specified according to the application). A selective summary of a soccer match will contain only the goals for instance, while for a cricket game it would contain only the wickets. The article by Xiong et al. [54] considers generic problems in summarization.

FROM A BROADCASTER'S POINT OF VIEW, EFFICIENT ARCHIVING OF MEDIA FACILITATES LATER REUSE IN THE CREATION OF HIGHLIGHTS PACKAGES OR SPECIALIZED DVDS.

The creation of a condensed summary requires two stages: an off-line learning stage and a supervised recognition step. The creation of a selective summary mainly relies on an unsupervised detection of events. Condensed summaries tend to achieve inference with HMMs [6], [14], [31], [35], [36]. Much of this has been discussed in the previous section. In contrast, in time-driven team sports like soccer or rugby, interesting events (e.g., goal scores or tries) are usually rare compared to the duration of the game. Also, they may occur with very different visual aspects. Thus, it is difficult to obtain a sufficient number of diverse examples for reliable training of the HMMs attached to each considered event, and unsupervised techniques tend to be more dominant here.

Leonardi et al. have designed a cross-modal (audio/video) method to detect goals in soccer games [14]. It exploits HMMs to classify each pair of successive shots as "goals," "corner kicks," and four other classes. Audio loudness is used to rank the detected goal shots. Petkovic et al. [37] include pause rate in the audio features combined with word spotting and video analysis (text, motion, and color analysis) to detect highlights in Formula 1 TV programs. Their method is based on a dynamic Bayesian network (DBN). Audio/visual fusion allows Coldefy et al. [38] to select events for summarization based on a combination of detection of increase in audio pitch and volume, with image color and large camera motion detection. In contrast to methods applied only to game sequences, this method is able to select the relevant highlights within the complete broadcast TV program including studio shots, interviews, and commercials. In their work, over 20 hours of footage was analyzed (World Cup'98 and French League Cup), having four different commentator teams. Out of a possible 17 goals, 15 were detected. A missed goal was due to a dramatic drop of the audio gain control. The summary duration created for each 2 h, 45-min long program was about 3 min and 45 s long, i.e., 4.5% of the total game duration and about 2.25% of the overall TV program.

EMERGING AREAS

RETRIEVAL IN SPORTS VIDEO

The identification of semantic-level events enables direct retrieval of those events [22]. However, by constructing a kind of domain-specific language for the game description, it is possible to give the user more flexible access to the system, which would possibly allow the retrieval of events that were not specifically marked [18]. The issue of describing the events or features extracted from sports remains open. The database community has acknowledged for some time the importance of a description language that allows the description of a video to be manipulated in a formalized way. In current database systems, the only video access is through a binary large object (blob), a black box where bits can be stored without any interpretation or processing of these bits. More appropriate "multimedia algebras"' that acknowledge the temporal nature of video streams have been derived from the work of Allen et al. [39], [40]. XML databases are promising, as they allow encapsulation of more flexible data than relational DBMS. But despite the fact that XML allows the encapsulation of numerical descriptors and signal data itself, the high dimensionality of that kind of description remains a problem.

In contrast, the languages defined by the media processing community are more description oriented. MPEG'7 (http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm) contains a long list of descriptor definitions, description schemes to combine these descriptors, and a language (DDL) to define these elements. Unfortunately it is too general to provide an efficient description for particular programs and languages like TV anytime (http://www.tv-anytime.org/) have been specifically designed for TV program descriptions. As yet, there is no such standardized description for sports.

A final issue in the design of language descriptors is coping with multimodal description arising from multimodal analysis of the stream as discussed in previous sections [41]. The two key issues are 1) coping with contradictory information from different streams and 2) coping with the different sampling rate of audio (44 KHz), video (25 Hz), and text. There are three different approaches to a solution. Audio and video analysis can be cascaded, so for instance audio is used to detect interesting segments, then video is used to identify the events occurring [42], [43]. Secondly, the audio and video data can be combined into a single feature vector, which implies a synchronization requirement for both data streams [44]. Finally, the results of separate analysis can be fused [45].

The collaboration between description languages and signal analysis is just emerging in the sports analysis domain. This is because the signal analysis tools are reaching maturity. It remains to be seen whether generic sports description is feasible or even desirable.

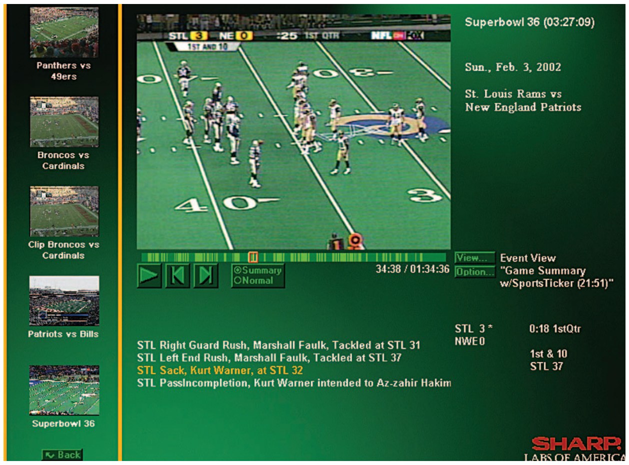
CONTENT-BASED SPORTS VIDEO COMPRESSION

Streaming of visual information over low-bit-rate networks is a challenging problem. For instance, when a soccer video is encoded at low bit rates with uniform quality, the quality degradation may be so severe that the ball and players are not visible and pitch lines are lost in the most important scenes (e.g., goals). The goal of content-adaptive sports video compression is to ensure acceptable visual quality in semantically important shots, while offering best effort quality in the other shots [46], [47]. Coding in this manner poses some new technical challenges. Aside from the problem of event detection in the video itself, one key problem is to quantify the relationship between the target bit rate and assigned importance measures in each segment. Details are outside the scope of this article, but the reader is directed to [48] for more information.

[INDEPENDENTLY GENERATED RICH TEXTUAL GAME DATA MAY BE SYNCHRONIZED WITH VIDEO EVENT SEGMENTS TO ENABLE SEMANTIC ACCESS.]

COMMERCIAL SYSTEMS

Adding value to sports broadcasts through innovative use of graphics is a well established exercise. Orad (www.orad.tv) has been responsible for in-screen advertising and object-attached labels for some time, while Dartfish (www.dartfish.com) is superimposing player/athletes with video histories in the same view. Many in the mobile phone industry have recognized the potential for sports to improve the market for new video-enabled mobile phones. Nokia (www.nokiausa.com/sports/nba.html) and 3 (www.three.co.uk/indexcompany.omp) offer sports packages for mobile phones. These packages include highlights and video updates. However, it is unclear whether these packages are offered by exploiting automatic sports analysis technology. Certainly, at least one company—IBIS (Integrated Broadcast Information Systems Ltd.) (www.ibistv.com)—is offering a product that facilitates manual compilation of sports highlights packages.



Superbowl 36 (03:27:09)

Sun., Feb. 3, 2002

St. Louis Rams vs
New England Patriots

34:38 / 01:34:36

Event View
"Game Summary
w/SportsTicker (21:51)"

STL 3 *
NWE 0

0:18 1stQtr

1st & 10
STL 37

STL Right Guard Rush, Marshall Faulk, Tackled at STL 31
STL Left End Rush, Marshall Faulk, Tackled at STL 37
STL Sack, Kurt Warner, at STL 32
STL PassIncompletion, Kurt Warner intended to Az-zahir Hakim

[FIG9] A screen shot from a Sharp prototype system for American football. The game statistics extracted from the SportsTicker data feed are shown to the right and lower-right corner of the main window. Directly below the window is a scrolling textural description of the game, with each line corresponding to a play and the line in golden font corresponding to the current play being viewed.

Two exceptions are Sharp Labs, United States, and Hitachi, Japan. Hitachi has introduced Prius Navistation3 software for its Prius line of personal computers. Navistation3 is claimed to summarize baseball and soccer broadcast videos. The user chooses one of the prespecified percentages for time reduction factor of the summary. It is reported that the Hitachi software uses crowd cheer as a feature. Navistation3 does not attempt to detect plays or exciting events.

Sharp Laboratories of America has created HiMPACT Sports and HiMPACT Coach software packages. HiMPACT Sports detects, in real-time, key events in baseball and football, and replays are labeled. The technology also supports sumo wrestling, where each bout is detected, and soccer. Automatic synchronization with SportsTicker (or any other structured data) is also implemented.

A screen shot from the Sharp system is shown in Figure 9 for American football. Users are able to retrieve event clips from different games stored in the database according to their queries, e.g., using a favorite player's name. The system can generate event-based summaries of recorded games and it allows nonlinear browsing. HiMPACT Coach uses a real-time automatic segmentation algorithm (exploiting HMMs [49]) that segments football coaching videos.

FINAL COMMENTS

This article set out to review work in sports content retrieval and unify the various approaches under a simple framework. The interest in this area is widespread, as evidenced not only by the large research activity but also by reports cited in the popular media [50]–[53]. It is interesting that one of the major benefits of digital media and digital television in particular has been that the user will be provided with more choices and a more interactive viewing experience. However, with the vast amount of data provided to viewers, particularly via the many available digital channels, the freedom to choose has in fact manifested as the freedom to choose from the options the broadcaster provides. It is only through the use of automated content-based analysis that sports viewers will be given a chance to manipulate content at a much deeper level than that intended by broadcasters, and hence put true meaning into interactivity.

AUTHORS

Anil Kokaram received the Ph.D from The University of Cambridge, United Kingdom, in 1993. From 1993–1998, he was a research associate and then fellow of Churchill College, Cambridge, while with the Signal Processing Group at the Engineering Department at Cambridge University. In 1998, he joined the University of Dublin, Trinity College, where he founded the Sigmedia group (www.sigmedia.tv). He is now a senior lecturer and fellow at that university and an associate editor of the *IEEE Transactions in Image Processing*. Much of his work has involved digital video processing of some kind, including digital motion picture restoration (publishing a book on the subject in 1998), digital cinema, and content-based analysis. His main research interests include practical Bayesian inference for signal processing, motion estimation, signal restoration, and data fusion for content-aware analysis.

Niall Rea received the Ph.D. from the University of Dublin, Trinity College, Ireland, in 2005. He is currently a postdoctoral researcher working between the Sigmedia and CTVR groups in Trinity College on content and network-aware sports multimedia streaming, adaptation, and summarization. His research interests include semantic event detection, content-based analysis, and tracking.

Rozenn Dahyot received the Ph.D. in 2001 from University Louis Pasteur Strasbourg, France, and has been a research fellow in Trinity College, Ireland, and Cambridge University, United Kingdom, from 2002–2005, working on video indexing and restoration. She is now a lecturer in the School of Computer Science and Statistics at Trinity College, Dublin, Ireland. Her research interests include multimedia understanding and restoration, object or event detection and recognition, tracking, and statistical learning and modeling.

A. Murat Tekalp received the Ph.D. from Rensselaer Polytechnic Institute, Troy, New York, in 1984. After 18 years with the University of Rochester, New York, where he was promoted to distinguished university professor, he is currently a professor at Koc University, Istanbul, Turkey. His research interests are in the area of digital image and video processing, including video compression and streaming, motion-compensated video filtering for high-resolution, content-based video analysis and summarization, multicamera video processing, and protection of digital content. He has chaired the IEEE Signal Processing Society Technical Committee on Image and Multidimensional Signal Processing (Jan. 1996–Dec. 1997). He was the general chair of the IEEE International Conference on Image Processing (ICIP) in 2002. At present, he is the editor-in-chief of the EURASIP journal *Signal Processing: Image Communication* (Elsevier). He authored the book *Digital Video Processing* (1995). He holds seven US patents. He is a Fellow of IEEE and was named a Distinguished Lecturer by IEEE Signal Processing Society.

Patrick Bouthemy received the Ph.D and the dipoma Habilitation a Diriger des Recherches from the University of Rennes, France, in 1982 and 1989. From 1982–1984, he was employed by INRS-Telecommunications, Montreal, Canada, in the Department of Visual Communications. Since 1984, he has been with INRIA at IRISA in Rennes. He is currently directeur de recherche Inria and head of the Vista group. His main research interests are: statistical approaches for image sequence processing, motion detection, motion computation, tracking, motion learning and recognition, video processing, content-based video indexing, and video microscopy. He has been involved in several European projects or networks. He was head of the Scientific Committee of Irisa from 1998–2002. He has served as member of the program committees of major conferences in image processing and computer vision. He was an associate editor of the *IEEE Transactions on Image Processing* from 1999–2003.

Patrick Gros has been involved in research in the field of computer vision for 14 years. In 1990, he joined the LIFIA-IMAG laboratory in Grenoble, achieving a Ph.D. Since 1993, he has had a research position at CNRS. In 1999, he moved from Grenoble to IRISA in Rennes. In 2002, he founded TEXMEX, a new research group devoted to multimedia document analysis and management, with a special emphasis on the problems raised by the management of very large volumes of documents. His research interests are image indexing and recognition in large databases and multimedia documents description. He teaches graduate courses in computer science and computer vision. He is associate editor of the *Traitement du Signal*. He participates in numerous national and European projects on multimedia description and indexing, with applications to television archiving and repurposing, copyright protection for photo agencies, and personal picture management of set-top boxes.

Ibrahim Sezan received the Ph. D. in 1984 from Rensselaer Polytechnic Institute, Troy, New York. He is the director of Information Systems Technologies at Sharp Laboratories of America Inc., Camas, Washington. He is currently leading R&D programs in visual-science-based display algorithms for LCD displays and future audiovisual flat-panel displays for super-realistic visual experience and with power and cost efficiency; video coding and video streaming over wireless networks; and intelligent systems for enabling smarter audiovisual devices and for creating new ways of consuming video content on new networked audiovisual display devices. Between 1990–1996, he was the founding leader of the Motion and Video Processing Technology Area at Eastman Kodak Research Laboratories, where he joined in 1984. He is a Fellow of the IEEE.

REFERENCES

[1] Informal EU Ministerial Conference on Broadcasting, "The impact of transfrontier broadcasting services on television markets in individual member states," Drogheda, Ireland, Feb. 2004.

[2] D. Zhong and S. Chang, "Structure analysis of sports video using domain models," in *Proc. Int. Conf. Multimedia and Expo*, Tokyo, Aug. 2001.

[3] H. Denman, N. Rea, and A. Kokaram, "Content-based analysis for video from snooker broadcasts," *J. Comput. Vision Image Understand., Special Issue on Video Retrieval and Summarization*, vol. 92, pp. 141–306, Nov./Dec. 2003.

[4] R. Dahyot, N. Rea, and A. Kokaram, "Sport video shot segmentation and classification," in *Proc. PIE Int. Conf. Visual Communication and Image Processing*, July 2003, pp. 404–413.

[5] Y. Gong, L.T. Sin, C.H. Chuan, H. Zhang, and M. Sakauchi, "Automatic parsing of TV soccer programs," in *Proc. Int. Conf. Multimedia Computing and Systems*, May 1995, vol. 7, pp. 167–174.

[6] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot, "HMM based structuring of tennis videos using visual and audio cues," in *Proc. IEEE Int. Conf. Multimedia and Expo*, July 2003, vol. 3, pp. 309–312.

[7] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden Markov models," in *Proc. IEEE Int. Conf. Image Processing*, Sept. 2002, pp. 609–612.

[8] A. Ekin, A.M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Processing*, vol. 12, no. 7, pp. 796–807, July 2003.

[9] P. Xu, L. Xie, S.-F. Chang, A. Divarakan, A. Vetro, and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Aug. 2001, pp. 721–724.

[10] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 68–75, Mar. 2002.

[11] R. Dahyot, N. Rea, A. Kokaram, and N. Kingsbury, "Inlier modeling for multimedia data analysis," in *Proc. IEEE Int. Workshop Multimedia Signal Processing*, Sep. 29–Oct. 1 2004, pp. 482–485.

[12] G. Sudhir, J.C.M. Lee, and A.K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in *Proc. EEE Int. Workshop Content-Based Access of Image and Video Databases*, Jan. 1998, pp. 81–90.

[13] A. Kokaram and P. Delacourt, "A new global estimation algorithm and its application to retrieval in sport events," in *Proc. IEEE Int. Workshop Multimedia Signal Processing,* Oct. 2001, pp. 251–256.

[14] R. Leonardi, P. Migliorati, and M. Prandini, "Semantic indexing of soccer audio-visual sequences: A multimodal approach based on controlled Markov chains," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 634–643, May 2004.

[15] N. Peyrard and P. Bouthemy, "Detection of meaningful events in videos based on a supervised classification approach," in *Proc. IEEE Int. Conf. Image Processing*, 2003, pp. III-621–624.

[16] N. Rea, R. Dahyot, and A. Kokaram, "Modeling high level structure in sports with motion driven HMMs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 2004, vol. 3, pp. 621–624.

[17] N. Rea, R. Dahyot, and A. Kokaram, "Classification and representation of semantic content in broadcast tennis videos," in *Proc. IEEE Int. Conf. Image Processing*, Genoa, Italy, 2005, pp. 1204–1207.

[18] A. Ekin and A.M. Tekalp, "Automatic soccer video analysis and summarization," in *Proc. SPIE Int. Conf. Electronic Imaging: Storage and Retrieval for Media Databases*, Jan. 2003, pp. 339–350.

[19] F. Pitié, S.-A. Berrani, R. Dahyot, and A. Kokaram, "Off-line multiple object tracking using candidate selection and the Viterbi algorithm," in *Proc. IEEE Int. Conf. Image Processing*, Genoa, Italy, 2005, pp. 109–112.

[20] D. Tjondronegoro, Y.-P. Phoebe Chen, and B. Pham, "Integrating highlights for more complete sports video summarization," *IEEE Multimedia*, vol. 11, no. 4, pp. 22–37, Oct.–Dec. 2004.

[21] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. ACM Conf. Multimedia*, Marina del Rey, Nov. 2000, pp. 105–115.

[22] R. Dahyot, A.C. Kokaram, N. Rea, and H. Denman, "Joint audio visual retrieval for tennis broadcasts," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2003, pp. III-561–564.

[23] B. Li, H. Pan, and M.I. Sezan, "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, May 2002, pp. IV-3385–3388.

[24] P. van Beek, H. Pan, and M.I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, May 2001, pp. 1649–1652.

[25] G. Piriou, P. Bouthemy, and J.-F. Yao, "Extraction of semantic dynamic content from videos with probabilistic motion models," in *Proc. European Conf. Computer Vision ECCV'04*, Prague, May 2004, pp. 145–157.

[26] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using HMMs," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Aug. 2002, pp. 825–828.

[27] D. Tjondronegoro, Y-P P. Chen, and B. Pham, "The power of play-break for automatic detection and browsing of self consumable sport video highlight," in *Proc. ACM Int. Workshop Multimedia and Information Retrieval MIR'04*, New York, Oct. 2004, pp. 267–274.

[28] B. Li, J. Errico, H. Pan, and M.I. Sezan, "Bridging the semantic gap in sports video retrieval and summarization," *J. Vis. Commun. Image Represent.*, vol. 15, pp. 393–424, 2004.

[29] A. Hanjalic, "Multimodal approach to measuring excitement in video," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME 2003)*, July 2003, vol. 2, pp. 289–292.

[30] A. Hanjalic and L.Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, pp. 143–154, Feb. 2005.

[31] B. Li and M.I. Sezan, "Event detection and summarization in sports video," in *Proc. IEEE Workshop Content-based Access of Image and Video Libraries CAIVL01*, 2001, pp. 132–138.

[32] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden Markov models," *Pattern Recognit. Lett.*, vol. 25, no. 7, pp. 767–775, May 2004.

[33] L. Xie and S.-F. Chang, "Unsupervised Mining of Statistical Temporal Structures in Videos," *Video Mining*. Norwell, MA: Kluwer, 2003.

[34] N. Rea, R. Dahyot, and A. Kokaram, "Semantic event detection in sports through motion understanding," in *Proc. 3rd Int. Conf. Image and Video Retrieval (CIVR 04)*, July 2004, pp. 88–97.

[35] J. Assfalg, M. Bertini, C. Colombo, A. del Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: Automatic highlights identification," *Comput. Vis. Image Understand.*, vol. 92, no. 2–3, pp. 285–305, Nov. 2003.

[36] L. Xie, S. Chang, A. Divakaran, and H. Sun, "Learning hierarchical hidden Markov models for video structure discovery," ADVENT Group, Columbia Univ., New York, Tech. Rep. 2002-006, Dec. 2002.

[37] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan, "Multi-modal extraction of highlights from TV Formula 1 programs," in *Proc. Int. Conf. Multimedia and Expo, ICME'2002*, Lausanne, 2002, pp. 817–820.

[38] F. Coldefy and P. Bouthemy, "Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis," in *Proc. ACM Multimedia*, Oct. 2004, pp. 268–271.

[39] K.L. Liu, A.P. Sistla, C.T. Yu, and N. Rische, "Query processing in a video retrieval system," in *Proc. Int. Conf. Data Engineering*, Orlando, FL, Feb. 1998, pp. 276–283.

[40] C. Decleir, M.S. Hacid, and J. Kouloumdjan, "A database approach for modeling and querying video data," in *Proc. IEEE Int. Conf. Data Engineering,* Sidney, Australia, Mar. 1999, pp. 6–13.

[41] Y. Wang, Z. Liu, and J.C. Huang, "Multimedia content analysis using both audio and visual cues," *IEEE Signal Processing Mag.*, vol. 17, pp. 12–36, Nov. 2000.

[42] K. Kim, J. Choi, N. Kim, and P. Kim, "Extracting semantic information from basket-ball video based on audio-visual features," in *Proc. Int. Conf. Image and Video Retrieval*, London, England, July 2002, vol. 2383, pp. 278–288.

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

[44] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan, "Multimodal extraction of highlights for TV Formula 1 programs," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Lausanne, Switzerland, Aug. 2002, vol. 1, pp. 817–820.

[45] N. Adami, A. Bugatti, R. Leonardi, and P. Migliorati, "Low-level processing of audio and video information for extracting the semantic content," in *Proc. 4th IEEE Workshop on Multimedia Signal Processing*, Cannes, France, Oct. 2001, pp. 607–612.

[46] S.F. Chang, D. Zhong, R. Kumar, and A. Jaines, "Method and system for indexing and content based adaptive streaming of digital content," B.S. Patent Application 20040125877, July 1 2004.

[47] S.F. Chang, D. Zhong, and R. Kumar, "Real-time content-based adaptive streaming of sports video," in *Proc. IEEE Workshop Content-Based Access to Video and Image Library*, Hawaii, Dec. 2001, pp. 139–146.

[48] T. Ozcelebi, A.M. Tekalp, and M.R. Civanlar, "Optimal rate and input format control for content and context adaptive video streaming," in *Proc. Int. Conf. Image Processing*, Singapore, Oct. 2004, pp. 2043–2046.

[49] B. Li and M.I. Sezan, "Semantic sports video analysis: Approaches and new applications," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2003, pp. I-17–20.

[50] J. Randerson, "Let software catch the game for you," *New Scientist*, July 3, 2004.

[51] S. Maguire, "Software detects sports highlights on its own," *Sunday Times*, July 11, 2004.

[52] The Guardian, "Software senses highlights to save time for busy football fans," *The Guardian*, Sept. 1, 2003.

[53] I. Austen, "For a squeeze play, software seeks out game highlights," *NY Times*, Apr. 29, 2004.

[54] Z. Xiong, X. Zhou, Q. Tian, Y. Rui, and T.S. Huang, "Semantic retrieval of video," *IEEE Signal Processing Mag.*, vol. 23, no. 2, pp. 18–27, 2006.