Semantic Annotation of Sports Videos

Jürgen Assfalg, Marco Bertini, Carlo Colombo, and Alberto Del Bimbo
University of Florence, Italy

Taking into consideration sports videos' unique qualities, we propose a system that semantically annotates them at different layers of semantic significance, using different elements of visual content. We decompose each shot into its visual and graphic content elements and, by combining several different low-level visual primitives, capture semantic content at a higher level of significance.

The quantity of videos generated by digital technologies has increased the need for automatically annotating video content and for techniques that support retrieving that content. Content-based video annotation and retrieval therefore has become an active research topic. Although we can successfully apply many of the results in content-based image retrieval to videos, additional techniques are necessary to address videos' unique qualities. For example, videos add the temporal dimension, requiring object dynamics. Furthermore, although people often think of a video as just a sequence of images, it's actually a compound medium, integrating diverse media such as realistic images, graphics, text, and audio.[1] Also, application contexts for videos are different than those for images and therefore call for different approaches to help users annotate, query, and exploit archived video data.

The huge amount of data that video streams deliver necessarily calls for higher levels of abstraction when we annotate content. This therefore requires us to investigate and model video semantics. Because of the type and volume of data, general-purpose approaches are likely to fail since semantics inherently depend on a specific application context. Many researchers have addressed semantic modeling of content in multimedia databases. Researchers have also reported on concrete video retrieval applications by high-level semantics in specific contexts such as movies, news, and commercials.[2,3]

Due to their enormous commercial appeal, sports videos represent an important application domain. However, most research efforts so far have been devoted to characterizing single, specific sports. (For example, Miyamori and Iisaku[4] proposed a method for annotating videos according to human behavior; Ariki and Sygiyama[5] proposed a method for classifying TV sports news videos using discrete cosine transform [DCT] features; and Zhou et al.[6] classified nine basketball events using color features, edges, and MPEG motion vectors.)

We propose an approach for semantic annotation of sports videos that include several different sports and even nonsports content. We automatically annotate videos according to elements of visual content at different layers of semantic significance. In fact, we primarily distinguish studio and interview shots from sports action shots and then further decompose the sports videos into their main visual and graphic content elements, including sport type, foreground versus background, text captions, and so on. We extract relevant semantic elements from videos by combining several low-level visual primitives such as image edges, corners, segments, curves, and color histograms, according to context-specific aggregation rules.

In this article, we illustrate three modules of our system, which performs semantic annotation of sports videos at different layers of semantic significance, using different elements of visual content.

The application context

The actual architecture of a system supporting video annotation and retrieval depends on the application context and, in particular, on end users and their tasks. Although all application contexts demand a reliable annotation of the video stream to effectively support selection of relevant video segments, it's evident that, for instance, service providers (such as broadcasters and editors) and consumers accessing a video-on-demand service have different needs.[7]

For both the old and new media, automatic annotation of video material opens the way for economically exploiting valuable assets. In particular, in the specific context of sports videos, two logging approaches exist, which let broadcasting companies reuse recorded material:

- *posterity logging*, where librarians add detailed and standardized annotation to archived material, and
- *production logging*, where (assistant) producers annotate live feeds or footage recorded a few

1070-986X/02/$17.00 © 2002 IEEE

Figure 1. Typical sequence of shots in a sports video.

hours before to edit a program that will be broadcast within a short time frame.

An example of reusing footage logged for posterity is selecting the best footage of a famous athlete to provide a historical context for a recent event. An example of production logging is selecting highlights, such as soccer goals or tennis-match points, to produce programs that contain one day's best sports actions.

In both scenarios, we should be able to automatically annotate video material, which is typically captured live, because detailed manual annotation is mostly impractical. The level of annotation should enable simple text-based queries. The annotation process includes such activities as segmenting the material into shots, grouping and classifying shots into semantic categories (such as type of sport), and supporting queries and retrieval of events that are significant to the particular sport. To achieve an effective annotation, we should have a clear insight into the current practice and established standards in the domain of professional sports videos, particularly concerning the nature and structure of their content.

The videos that comprise the data set we used for the experiments in this article include a variety of typologies. We drew most of the videos from the BBC Sports Library, which in turn collected them from other BBC departments and other broadcasters. The data set comprises more than 15 video tapes, each lasting from 30 to 120 minutes. The videos were mostly captured from digital video tapes and, in some cases, from S-VHS tapes. Digital video is the lowest acceptable standard for broadcasting professionals, and it provides digital quality at full PAL resolution.

Many of the videos in this sample collection are from the 1992 Barcelona Olympics while some contain soccer games from other events. Thus, we used various types of sports to perform our experiments. The videos differ from each other in terms of types of sports (outdoor and indoor sports) and the number of athletes (single player or teams). Also, the videos differ in terms of editing—some are live feeds from a single camera for a complete event, some include different feeds of one event edited into a single stream, and others only feature highlights of minor sports assembled in a summary. We weren't able to make assumptions on the presence of a spoken commentary or superimposed text because that depends on a number of factors, including the technical facilities available on location and the agreements between the hosting broadcaster and other broadcasters. Figure 1, which includes sport sequences interwoven with studio scenes, shows the typical structure of a sports video; some videos might include superimposed graphics (such as captions or logos).

The computational approach

We organized the annotation task into three distinct subtasks:

1. preclassify shots (to extract the actual sports actions from the video stream),
2. classify graphic features (which, in sports videos, are mainly text captions that aren't synchronized with shot changes), and
3. classify visual shot features.

Next, we'll expound on the contextual analysis of the application domain. We analyze specificity of data and provide an overview on the rationale underlying how we selected relevant features.

Preclassifying sports shots

Our anchorman/interview shot classification module provides a simple preliminary classification of shot content, which subsequent modules can also exploit and enhance. This type of classification is necessary because some video feeds contain interviews and studio scenes featuring an anchorman and athletes. One example of this is the Olympic Games, where the material to be logged is often pre-edited by the hosting broadcaster. The purpose of this module is to roughly separate shots that contain possible sport scenes from shots that do not. To this end, we can follow a statistical approach to analyze visual con-

Figure 2. Studio scene with alternating anchorman shots.

tent similarity and motion features of the anchorman shots, without requiring us to use any predefined shot content model as a reference. In fact, our system doesn't require this latter constraint to be able to correctly manage interviews, which don't feature a standard studio set-up, because athletes are usually interviewed near the playing field and each interview has a different background and location. Also, detecting studio scenes requires such independence of a shot content model because each program's style is unique and often changes. This would require us to create and maintain a database of shot content models.

Studio scenes have a well-defined syntax: shot location is consistent within the video, the number of cameras and their view field is limited, and the sequence of shot content is often a repeating pattern. An example of such a structure is in Figure 2, which displays the first key frames of shots comprising a studio scene.

Identifying graphic features

In sports videos, graphic objects may appear anywhere within the frame, even if most of the time they're in the lower third or quarter of the image. Also the vertical and horizontal ratio of the graphic object zones varies—for example, a team roster might occupy a vertical box and one athlete's name might occupy a horizontal box (see Figure 3). For text graphics, character fonts can vary in size and typeface and may be superimposed either on an opaque background or directly on the image captured by the camera. Graphic objects often appear and disappear gradually, with dissolve or fade effects. These properties call for automatic graphic object localization algorithms with the least amount of heuristics and possibly no training.

Past research has used several features such as edges and textures as cues of superimposed graphic objects.[8,9] Such features represent global image properties and require the analysis of large frame patches. Moreover, natural objects such as woods and leafs or man-made objects such as buildings and cars can present a local combination of such features that the algorithms might wrongly classify as a graphic objects.[10]

To reduce visual information to a minimum and preserve local saliency, we've elected to work with image corners, extracted from the images' luminance information. This is appealing for the purpose of graphic-object detection and localization because it prevents many misclassification problems arising with color-based approaches. This is particularly important when considering the characteristics of TV standards, which enforce a spatial subsampling of the chromatic information, causing the borders of captions to be affected by color aliasing. Therefore, to enhance the readability of characters, producers typically exploit luminance contrast because luminance isn't spatially subsampled and human vision is more sensitive to it than to color contrast. Another distinguishing feature of our approach is that it doesn't require any knowledge or training on superimposed captions features.

Classifying visual shot features

Generic sports videos often feature numerous different scene types intertwined with each other in a live video feed reporting on a single event or edited into a segment summarizing highlights of different events. A preliminary analysis of our data set—covering more than 10 hours of sports events—revealed that three types of scenes are most prevalent: the playing field, player, and audience (see Figure 4). Most of the action of a sport game takes place on the playing field—hence, the relevance of playing field scenes, showing mutual interactions among subjects (players, referees, and so on) and objects (ball, goal, hurdles, and so on). However, along with playing field scenes, a number of scenes often appear, such as player close-ups and audience shots. The former typically show a player who had a relevant role in the most recent action—for example, the athlete who just failed

throwing the javelin or the player who shot a penalty. Audience shots occur at the beginning and at the end of an event, when nothing is happening on the playing field, or just after a highlight. For example, in soccer, when a player shoots a goal, an audience shot showing the crowd's reaction is common.

In a sample of 1,267 key frames extracted randomly from our data set, approximately 9 percent were audience scenes. Player scenes represented up to 28 percent of the video material. To address the specificity of such a variety of content types, we devised a hierarchical classification scheme. The first stage performs a classification in terms of the categories of playing field, player, and audience, with a twofold aim. On the one hand, this provides an annotation of video material that is meaningful for users' tasks. On the other hand, it's instrumental for further processing, such as identifying sports type and detecting highlights.

Inspecting the video material revealed that

- playing field shots typically feature large, homogeneous color regions and distinct long lines;
- in player shots, the player's shape appears distinctly in the foreground and the image's background tends to be homogeneous or blurred (either because of camera motion or lens effects); and
- in audience shots, individuals in the audience aren't always clear but the audience as a whole appears as a texture.

These observations suggest that basic edge and shape features could significantly help differentiate between playing field, player, and audience scenes. It's also worth pointing out that models for these classes don't vary significantly across different sports, events, and sources.

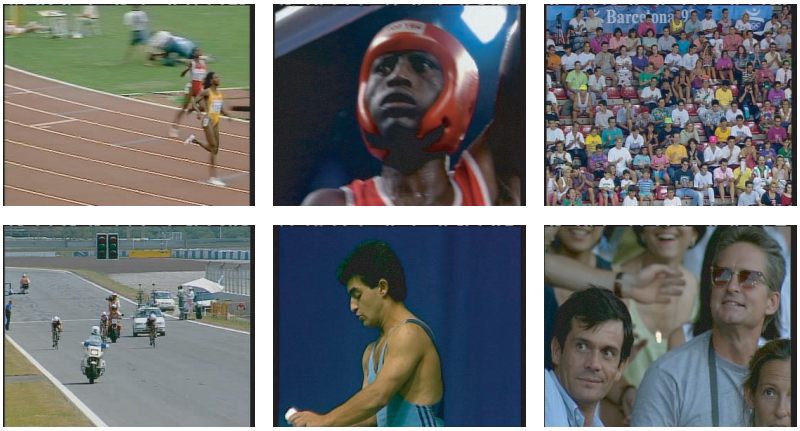
We propose that identifying the type of sport in a shot relies on playing field detection. In fact, we can observe that

- most sports events take place on a playing field, with each sport having its own playing field;
- each playing field has several distinguishing features, the most relevant of which is color; and
- the playing field appears in many frames of a video shot and often covers a large part of the camera frame—that is, a large area of single images comprising the video.

Hence, the playing field and the objects that populate it can effectively support identification of sports types. Therefore, in our approach, we apply sports type identification to playing field shots output by the previous classification stage.

(a) (b)

Figure 3. Examples of superimposed graphic objects. (a) A team roster might occupy a vertical box or (b) one athlete's name might occupy a horizontal box.



(a) (b) (c)

Figure 4. Although sports event footage can vary significantly, they share distinguishing features. For example, (a) playing field lines are explicitly present in some outdoor and indoor sports (such as soccer and swimming) but can also be extended to other sports (such as cycling on public roads). Similarly, (b) player and (c) audience scenes appear in most sports videos.

Implementation

In this section, we describe the implementations of modules supporting each subtask. We show how to compute the features and how we implemented the feature combination rules.

The most basic property of graphic objects is that they must remain stable for a certain amount of time so people can read and understand them.

Preclassifying sports shots

Studio and interview shots repeat at intervals of variable length throughout a video sequence. The first step for classifying these shots stems from this assumption and is based on the computation, for each video shot S_k, of a shot's lifetime $L(S_k)$. The shot lifetime measures the shortest temporal interval that includes all the occurrences of shots with similar visual content within the video. Given a generic shot S_k, we compute its lifetime by considering the set $T_k = \{t_i \mid \sigma(S_k, S_i) < \tau_s\}$, where $\sigma(S_k, S_i)$ is a similarity measure applied to key frames of shots S_k and S_i, τ_s a similarity threshold, and t_i is the value of the time variable corresponding to the occurrence of the key frame of shot S_i. We define the lifetime of shot S_k as $L(S_k) = \max(T_k) - \min(T_k)$. We base shot classification on fitting values of $L(S_k)$ for all the video shots in a bimodal distribution. This lets us determine a threshold value t_l that we use to classify shots into the sport and studio/interview categories. Particularly, we classify all the shots S_k so that $L(S_k) > t_l$ are studio/interview shots, where t_l was determined according to the statistics of the test database and set to 5 seconds. We classify the remaining shots as sport shots.

Typical videos in the target domain don't contain complete studio shows, and in feeds produced on location, interviews have a limited time and shot length. This lets us reduce false detections caused by the repetition of similar sport scenes (for example, in the case of edited programs or summaries) by limiting the search of similar shots to a window of shots whose width we set experimentally to six shots. The adopted similarity metric is a histogram intersection of the mean color histogram of shots (H and S components of the HSI color space). Using the mean histogram takes into account the dynamics of sport scenes. In fact, even if some scenes take place in the same location, and thus the color histogram of their first frame may be similar, the following actions yield a different color histogram. When applied to studio and interview shots, where the dynamics of the scene's lighting changes are much more compressed and the reduced camera and objects movement don't introduce new objects, we get a stable histogram.

Although the mean color histogram accounts for minor variations due to camera and objects movement, it doesn't account for spatial information. We refined the results from the first classification step by considering motion features of the studio/interview shots. This develops on the assumption that in an anchorman shot both the camera and the anchorman are almost steady. In contrast, for sport shots, background objects and camera movements—people, free-hand shots, camera panning and zooming, changes in scene lighting—cause relevant motion components throughout the shot. We performed classification refinement by computing an index of the quantity of motion Q_M for each possible anchorman shot. The algorithm for the analysis of this index considers the frame-to-frame difference between the shot key-frame f_1 and subsequent frames f_i in the shot according to a pixel-wise comparison. To enhance sensitivity to motion, the algorithm subsamples the shot in time and compares the frames to the first key-frame f_1. Only those shots whose Q_M doesn't exceed a threshold τ_M are definitely classified as studio/interview shots.

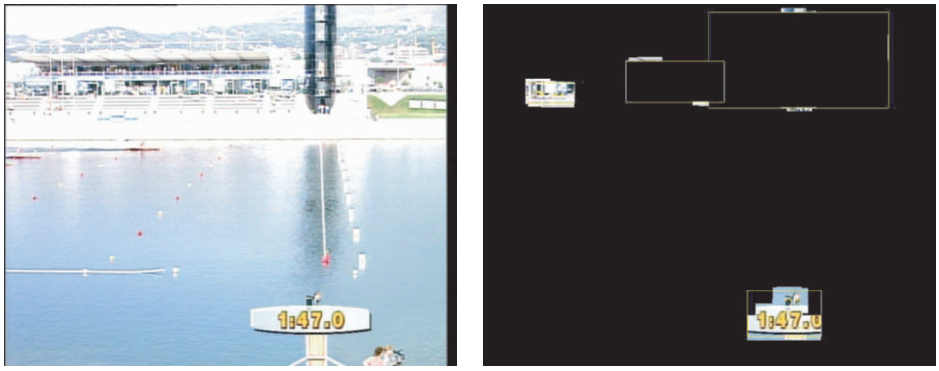
We used a subset of three videos to test the algorithm. There were 28 shots comprising short interviews of athletes and studio scenes. The algorithm identified 31 studio/interview shots, with five false detections and two missed detections. Using the movement feature reduced the number of false detections from five to three. The remaining false detections were due to slow-motion replays.

Identifying graphic features

We extracted the salient points of the frames, which we analyze in the following steps, using the Harris algorithm, from the luminance map of each frame. Corner extraction greatly reduces the amount of spatial data to be processed by the graphic-object detection and localization system. The most basic property of graphic objects is that they must remain stable for a certain amount of time so people can read and understand them.

We used this property in the first step of graphic-object detection. We checked each corner to determine if it's still in the same position for at least two more frames within a sliding window of four frames.

We marked each corner that complied with this property as persistent and kept it for further analysis. We then discarded all the others. Every eighth frame is processed to extract its corners, thus further reducing the computational resources needed to process a whole video. This choice is based on the assumption that, in order for the viewer to perceive and understand it, a graphic object must be stable on the screen for 1 second. We inspected the patch surrounding each corner (20 × 14 pixels), and in the absence of enough neighboring corners (corners whose patches don't intersect), we didn't consider the corner in further processing. This process, which we repeated in order to eliminate corners that got isolated after the first processing, assured that isolated high-contrast background objects contained within static scenes weren't recognized as possible graphic-object zones.



(a) (b)

Figure 5. Determining graphic objects' bounding boxes. (a) Source frame. (b) Detected captions with noise removal.

(a) (b)

Figure 6. The results from detecting the bounding boxes in Figures 3a and 3b.

An unsupervised clustering is performed on the corners that comply with the temporal and spatial features we described earlier. This is aimed at determining bounding boxes for graphic objects (see Figures 5 and 6). For each bounding box, we calculate the percentage of pixels that belong to the corner patches, and if it's below a predefined threshold, we discarded the corners. This strategy reduces the noise due to high-contrast background during static scenes that typically produce small scattered zones of corners that the spatial feature analysis can't eliminate.

The test set we used was composed of 19 sports videos acquired from PAL digital video tapes at full PAL resolution and frame rate (720 × 576 pixels, 25 fps) resulting in 47,014 frames (31 minutes 20 seconds).

Figure 6 provides an example of graphic-object detection. To test the robustness with respect to text size and video resolution, we also digitized two S-VHS videos. One video was acquired at full PAL resolution (720 × 576 pixels, 25 fps) resulting in 4,853 frames (3 minutes 14 seconds), and the second one was acquired at half PAL resolution (368 × 288, 25 fps) and contained 425 frames (totaling 17 seconds). With this latter resolution, some text captions in the videos become only 9 pixels high.

Table 1. Text event detection and text box localization.

	Occurrences	Misdetections	False Detections
Graphic-object event	63	5	9
Graphic-object boxes	100	5	37

When we evaluate our results, we consider graphic-object detection (whether the algorithm correctly detected a graphic object's appearance) and correct detection of the graphic-object's bounding box (see Table 1). Graphic-object detection has a precision of 80.6 percent and recall of 92 percent. These figures are due to missed detections in the VHS video and to only

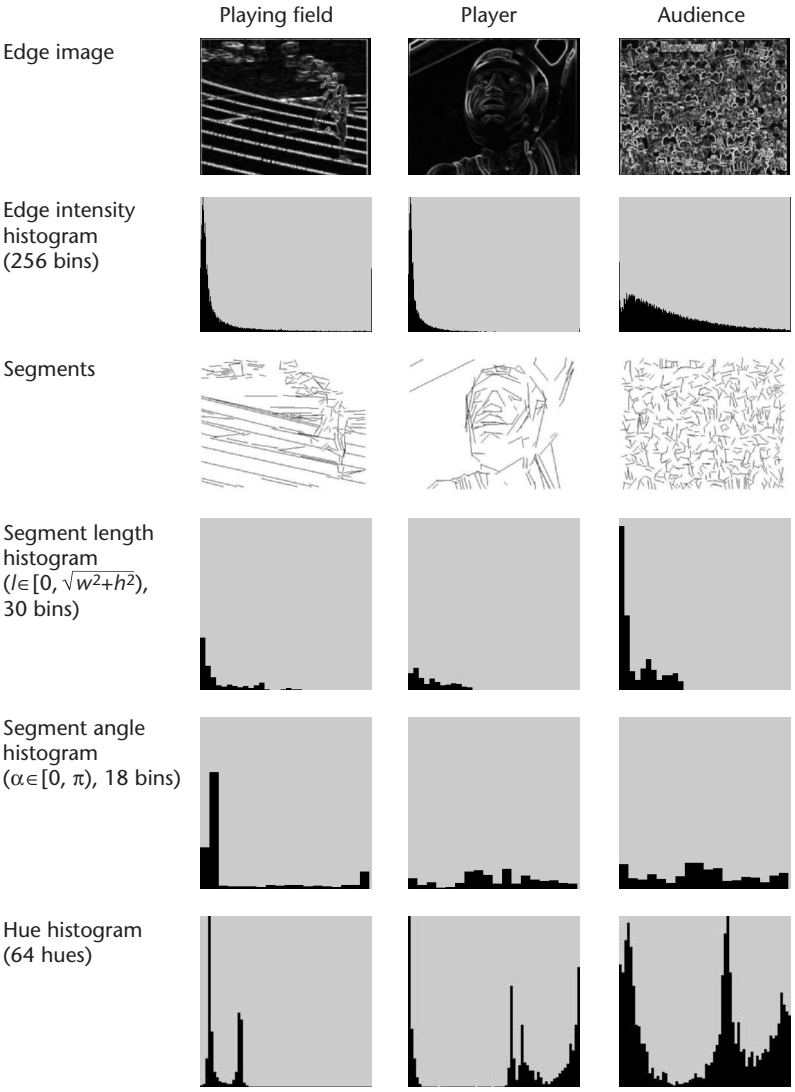


Figure 7. Edge, segment length and orientation, and hue distribution for the three representative sample images from Figure 4. Synthetic indices derived from these distributions let us differentiate between the three classes playing field, player, and audience. (Note that we scaled hue histograms to the maximum value.)

one missed detection in the digital videos. The graphic-object bounding box miss rate is 5 percent. Results also included six false detections, which were due to the presence of scene text.

Classifying visual shot features

We devised a feature vector comprising edge, segment, and color features. Figure 7 shows some of the selected features for representatives of the three classes playing field, player, and audience, respectively.

First, we performed edge detection and applied a recursive growing algorithm to the edges to identify the image's segments.[11] We analyzed the distribution of edge intensities to evaluate the degree of uniformity. We also analyzed distributions of lengths and orientations of the segments to extract the maximum length of segments in an image, as well as to detect whether peaks exist in the distribution of orientations. Our choice was driven by these observations:

- playing field lines are characteristic segments in playing field scenes,[12] because these lines determine peaks in the orientation histogram and longer segments than in other types of scenes;
- audience scenes are typically characterized by more or less uniform distributions for edge intensities, segments orientation, and hue; and
- player scenes typically feature fewer edges, a uniform segment orientation distribution, and short segments.

We also expect color features to increase robustness to the first classification stage (for example, audience scenes display more uniform color distributions than playing field or player scenes) and to support sports type identification. In fact, the playing field each sport usually takes place on typically features a few dominant colors—one or two, in most cases. This is particularly the case in long and mid-range camera shots, where the frame area occupied by players is only a fraction of the whole area. Furthermore, for each sports type, the color of the playing field is either fixed or it varies in a small set of possibilities. For example, a soccer field is always green and a swimming pool is always blue. We describe color content through color histograms. We selected the HSI color space and quantized it into 64 levels for hue, three levels for saturation, and three levels for intensity. We also derived indices describing the distribution—that is, degree of uniformity and number of peaks—from these distributions.

We used two neural network classifiers to perform the classification tasks. To evaluate their performance, we extracted more than 600 frames from a range of video shots and manually annotated them to define a ground truth. We then subdivided the frames into three sets to train, test, and evaluate the classifiers. We computed the edge, segment, and color features for all the frames. Table 2 summarizes results for the scene-type classification. Extending this classification scheme to shots, rather than just limiting it to key

frames, will yield an even better performance, because integrating results for key frames belonging to the same shot reduces error rates. For instance, some key frames of a playing field shot might not contain playing field lines (for example, because of zoom), but others will. Hence, we can classify the whole shot as a playing field shot.

Table 3 shows the results of the sports-type identification. The first column refers to an experiment we carried out on a data set including playing field, player, and audience scenes. The second column summarizes an experiment we carried out after we used a filtering process that kept only playing field frames. As expected, in the former case, we obtained lower success rates. By comparing results in the two columns, we can observe that introducing the playing field, player, and audience classes is instrumental to improving identification rates for sports types. On average, the rates improve by 16 percent, with a maximum of 26 percent. The highest improvement rates are for those sports where the playing field is shown only for small time intervals (such as high diving) or in sports where only one athlete competes, videos which frequently show close-ups of the athlete (such as the javelin).

Table 2. Results for the classification of key frames in terms of playing field, player, and audience classes.

Class	Correct (Percent)	Missed (Percent)	False (Percent)
Playing field	80.4	19.6	9.8
Player	84.8	15.2	15.1
Audience	92.5	7.5	9.8

Table 3. Sports-type identification results. The evaluation set in the first experiment included playing field, player, and audience scenes. The second set included only playing field scenes.

Sports Type	All Frames (Percent)	Playing Field Only (Percent)
High diving	56.9	83.2
Gymnastics floor exercises	78.7	97.4
Field hockey	85.0	95.1
Long horse	53.4	64.3
Javelin	37.8	58.8
Judo	80.6	96.9
Soccer	80.3	93.2
Swimming	77.4	96.1
Tennis	69.1	94.5
Track	88.2	92.7

Future work

Besides refining the techniques we describe in this article, our future work includes introducing new semantic elements for a given semantic layer, such as motion and framing terms (for example, close-up versus long shots); increasing the overall level of semantic description (for example, by adding descriptors for events and relevant highlights); and transitioning from elements of visual content to relationships among elements (spatio-temporal relations). We're also implementing shape-analysis techniques to support highlight detection (for example, by identifying playing field lines, goals, hurdles, and so forth).

Eventually, our work should yield an exhaustive annotation of sports videos, letting us select highlights in a sports event to enhance production logging or extract metadata to achieve posterity logging. For example, we expect our system to detect (missed) goals, penalties, or corner shots in a soccer game by combining (camera) motion patterns, information on the location of players with regards to playing field lines, and other relevant markers. When accessing historical archives, users could benefit from richly annotated files, helping them retrieve specific shots (or types of shots) on demand. **MM**

Acknowledgments

This work was partially supported by the ASSAVID EU Project (Automatic Segmentation and Semantic Annotation of Sports Videos, http://www.bpe-rnd.co.uk/assavid/) under contract IST-13082. The consortium comprises ACS SpA, Italy; BBC R&D, UK; Institut Dalle Molle D'Intelligence Artificielle Perceptive (Dalle Molle Institute for Perceptual Artifical Intelligence), Switzerland; Sony BPE, UK; the University of Florence, Italy; and the University of Surrey, UK.

References

1. R.S. Heller and C.D. Martin, "A Media Taxonomy," *IEEE MultiMedia*, vol. 2, no. 4, Winter 1995, pp. 36-45.
2. C. Colombo, A. Del Bimbo, and P. Pala, "Semantics in Visual Information Retrieval," *IEEE MultiMedia*, vol. 6, no. 3, July–Sept. 1999, pp. 38-53.
3. S. Eickeler and S. Muller, "Content-Based Video Indexing of TV Broadcast News Using Hidden Markov Models," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP), IEEE Press, Piscataway, N.J., 1999, pp. 2997-3000.

4. H. Miyamori and S.-I. Iisaku, "Video Annotation for Content-Based Retrieval Using Human Behavior Analysis and Domain Knowledge," *Proc. Int'l Conf. Automatic Face and Gesture Recognition 2000*, IEEE CS Press, Los Alamitos, Calif., 2000.
5. Y. Ariki and Y. Sugiyama, "Classification of TV Sports News by DCT Features Using Multiple Subspace Method," *Proc. 14th Int'l Conf. Pattern Recognition* (ICPR 98), IEEE CS Press, Los Alamitos, Calif., 1998, pp. 1488-1491.
6. W. Zhou, A. Vellaikal, and C.C.J. Kuo, "Rule-Based Video Classification System for Basketball Video Indexing," *Proc. ACM Multimedia 2000 Workshop*, ACM Press, New York, 2000, pp. 213-216.
7. N. Dimitrova et al., "Entry into the Content Forest: The Role of Multimedia Portals," *IEEE MultiMedia*, vol. 7, no. 3, July–Sept. 2000, pp. 14-20.
8. T. Sato et al., "Video OCR for Digital News Archive," *Proc. IEEE Int'l Workshop Content-Based Access of Image and Video Databases* (CAIVD 98), IEEE CS Press, Los Alamitos, Calif., 1998, pp. 52-60.
9. Y. Zhong, H. Zangh, and A.K. Jain, "Automatic Caption Localization in Compressed Video," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, Apr. 2000, pp. 385-392.
10. H. Li and D. Doermann, "Automatic Identification of Text in Digital Video Key Frames," *Proc. Int'l Conf. Pattern Recognition* (ICPR 98), IEEE CS Press, Los Alamitos, Calif., 1998, pp. 129-132.
11. R.C. Nelson, "Finding Line Segments by Stick Growing," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, May 1994, pp. 519-523.
12. Y. Gong et al., "Automatic Parsing of TV Soccer Programs," *Proc. Int'l Conf. Multimedia Computing and Systems* (ICMCS 95), IEEE CS Press, Los Alamitos, Calif., 1995, pp. 167-172.

Jürgen Assfalg is a research associate at the Department of Information Systems at the University of Florence, Italy. His research activity addresses multimedia databases, advanced user interfaces, usability engineering, virtual reality, and 3D graphics. He graduated with a degree in electronics engineering and has a PhD in information and telecommunications technology from the University of Florence, Italy. He is an IEEE and ACM member.

Readers may contact Jürgen Assfalg at the Dipartimento di Sistemi e Informatica, Via S. Marta 3, 50139 Firenze, Italy, email assfalg@dsi.unifi.it.

Marco Bertini is a PhD candidate in information and telecommunications technology at the University of Florence, Italy. His research interests include content-based video annotation and retrieval and Web user interfaces. He has a Laurea degree in electronics engineering from the University of Florence, Italy.

Carlo Colombo is a senior assistant professor at the University of Florence, Italy. His current research activities focus on theoretical and applied computer vision for semi-autonomous robotics, advanced human–machine interaction, and multimedia systems technology. He has an MS in electronic engineering from the University of Florence, Italy, and a PhD in robotics from the Sant'Anna School of University Studies and Doctoral Research, Pisa, Italy. He is an editorial board member of the *Journal of Robotics and Autonomous Systems*.

Alberto Del Bimbo is a full professor of computer engineering and the Director of the Master in Multimedia of the University of Florence, Italy. He is also the Deputy Rector of the University of Florence, in charge of research and innovation transfer. His research interests include pattern recognition, image databases, multimedia, and human–computer interaction. He graduated with honors in electronic engineering from the University of Florence, Italy. He is the author of the *Visual Information Retrieval* monograph on content-based retrieval from image and video databases. He is also associate editor of *Pattern Recognition*, the *Journal of Visual Languages and Computing*, the *Multimedia Tools and Applications Journal*, *Pattern Analysis and Applications*, the *IEEE Transactions on Multimedia*, and the *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He is a member of the IEEE and the International Association for Pattern Recognition (IAPR).