Automatic Annotation of Sport Video Content

Marco Bertini, Alberto Del Bimbo, and Walter Nunziati

Dipartimento di Sistemi e Informatica - Universitá degli Studi di Firenze
{bertini, delbimbo, nunziati}@dsi.unifi.it

Abstract. Automatic semantic annotation of video streams allows to extract significant clips for archiving and retrieval of video content. In this paper, we present a system that performs automatic annotation of soccer videos, detecting principal highlights, and recognizing identity of players. Highlight detection is carried out by means of finite state machines that encode domain knowledge, while player identification is based on face detection, and on the analysis of contextual information such as jersey's numbers and superimposed text captions. Results obtained on actual soccer videos shows overall highlight detection rates of about 90%. Lower, but still promising, accuracy is achieved on the very difficult player identification task.

1 Introduction and Background Work

To provide effective archiving and retrieval of video material, video streams must be annotated with respect to its semantic content, producing metadata that is attached to the video data and stored in databases. This will permit, for example, to produce special video summaries for a sport program such those that recollect the best actions occurred during a typical soccer turn, or those where are notable actions of a certain player. In this case the parts of the video containing important highlights must be selected and edited to create a new video sequence. One limitation to the diffusion of this practice is due to the fact that manually summarizing, annotating or tagging video is a cumbersome and expensive process. This has motivated recently the investigation of techniques to extract semantic information automatically from sports video sequences. At semantic level, video annotation regards the identification and recognition of meaningful entities and events represented in the video. Semantic video annotation is obtained combining observed features and patterns, like settings, text captions, people and objects, highlights and events, and domain knowledge. The latter is required in order to reduce the semantic gap between the observable features of the multimedia material and the interpretation that a user have. A good review of multimodal video annotation is provided in [16].

Due to their huge commercial appeal sports videos represent an important application domain for video automatic annotation [20]. Sport shots can be classified into the most common scene types, that are playfield, players' close-ups and crowd, using edges, segments and color information. From playfield shots it

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 1066–1078, 2005.

is possible to perform sport classification based on the characteristics of the playfield like ground color and lines. Solutions for recognition of specific highlights have been proposed for different sports like soccer, tennis, basketball, volleyball, baseball, American football. Usually these methods exploit low and mid level audio and visual cues, such as the detection of referee's whistle, excited speech, color and edge related features, playfield zone identification, players and ball tracking, motion indexes, etc. and relate them to a domain knowledge of the sports or of the video producers. In the first case knowledge of the sports rules and typical actions are used, in the second case the production rules employed by directors, like the presence of slow motion replays, are used. Good examples are reported in [17]for tennis, in [21] for basketball, in [11] and [13] for football and in [1], [6], [19] for soccer.

Several researchers have also focussed on the identification of people in the video for the purpose of video semantic annotation. Person recognition by means of association of interpreted textual content - extracted from text captions - to faces - automatically detected from skin tone analysis - has been investigated in the context of news video in [14], and more recently in [3] and [4]. Two important recent works for people identification are [7] and [15].

In this paper we present recent results of our research for providing rich annotations of highlights in soccer. The definition of highlights is based on formal methods (using finite state machines) and is detected through a model checking engine. Highlight detection is based on a limited set of visual cues, which are derived from low–level features such as edges, shapes and color. To provide a richer annotation, we add details related to the players who take part in a particular highlight occurrence using information extracted from faces, jersey numbers and superimposed text captions, which are usually present in the video stream.

The paper is organized as follows. In Sect. 2, we briefly introduce peculiarities that can be exploited for modeling highlights in soccer, providing details on estimation of visual cues, and on the model checking algorithm. Detection and recognition of the player is discussed in Sect.3. Players that are not identified in this process are then linked by similarity to one of the labeled faces (3.4). Examples of highlight recognition and superimposed caption extraction and face detection and recognition are shown in Sect.4 and 5, together with indications of results obtained. Conclusions and future work are discussed in Sect. 6.

2 Soccer Video Highlight Detection

Our solution for highlight modeling and detection employs finite state machines (FSMs). States represent main phases in which actions can be decomposed. Events indicate transitions from one state to the other: they capture the relevant steps in the progression of the play and are expressed in terms of a logical combination of a few visual cues, the camera motion, the playfield zone that is framed and the position of soccer players extracted from the video stream. Time constraints, for example a minimum temporal duration, can be applied to state

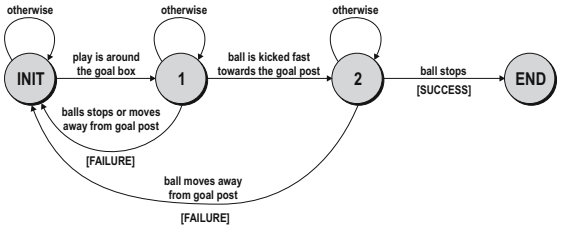

Fig. 1. The informal model of a shot on goal highlight in the soccer domain

transitions [1]. Fig. 1 shows how the essential phases of a shot on goal highlight have been represented as a FSM.

In the following we discuss in detail the solutions adopted for the estimation of the visual cues and the each separate detector and players' annotation.

2.1 Estimation of Visual Cues

Camera Motion. In soccer, ball instantaneous position and motion direction are important cues for the understanting of the play. A well known production rule of soccer videos is that the director uses the main camera to follow the ball and the play. For this reason, we rely on camera motion as a somewhat rough, but reliable estimate of the speed and the direction of the ball. As the main camera is observing the soccer playfield in a fixed position, a 3-parameter image motion estimation algorithm capturing horizontal and vertical translations and isotropic scaling is sufficient to get a reasonable estimate of camera pan, tilt and zoom. Motion estimation algorithm that has been used is an adaptation to the sports videos domain of the algorithm reported in [2], that is based on corner tracking and motion vector clustering. As it works with a selected number of salient image locations, the algorithm can cope with large displacements due to fast camera motion. The algorithm employs deterministic sample consensus to perform a statistical motion grouping. This is particularly effective to cluster multiple independent image motions, and is therefore suitable for the specific case of sports videos to separate camera motion from the motion of individual players.

Playfield Zone Estimation. To estimate playfield zone, the playfield is first partitioned in several, possibly overlapping zones, which are defined so that the change from one to the other indicates a change in the play. In general, a typical camera view is associated with each playfield zone, and we exploit common patterns of these views to recognize which zone is framed. Fig. 2 shows the partition of the playfield that we have used. Each zone is recognized using a dedicated Naìve Bayes classifier, which takes as input a (vector-quantized) descriptor. The descriptor itself is derived from low–level features such as edge, shape and color. Since classifiers have identical structure, large differences in

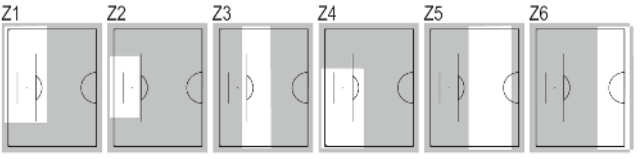

Fig. 2. Playfield partition for soccer

their outputs are likely to be significant, so we choose the classifier with the highest output probability as the one identifying the zone currently framed.

Player Position and Speed. Players' position is instrumental for the recognition of those highlights that are characterized by a typical deployment of players on the playfield, like all "free shots". For these highlights, both camera and players are usually still at the start of the action, allowing a robust estimation of the position of players, so that typical configurations of the deployment can be recognized. We exploited the knowledge of the actual planar model of the playfield to estimate automatically the homography [10] which maps any imaged playfield point (x, y) onto the real playfield point (X, Y). Players are first detected as "blobs" from each frame by color differencing to identify their position on the frame. Bottom-end point of each detected template is remapped onto the playfield model trough the estimated homography. Assuming the pinhole camera projection model, the image-to-playfield transformation has the form of a planar homography. Since we are provided with a set of line segments as the result of image segmentation, the homography is estimated using four line correspondences. If $[a\ \ b\ \ c]^T$ such that $ax + by + c = 0$ is the image of a playfield line $[A\ \ B\ \ C]^T$, it holds:

$$[a\ \ b\ \ c]^T = \mathrm{K}\,[A\ \ B\ \ C]^T\ \ , \tag{1}$$

with $\mathrm{K} = \mathrm{H}^T$, and H is the image–to–model homography.

The output of the registration process is then used to build a compact representation of how the players are deployed, such as the histogram of players' occupation of playfield zones. The presence or absence of players in the areas contributes to discriminate between three classes of free kicks, namely penalty kicks, corner kicks and free kicks with wall.

Motion information of objects and/or athletes that are present in the scene is obtained from the same motion processing and clustering used for camera motion estimation. In fact, we cluster motion magnitude and direction of pixels that are moving independently. This measure is sufficient to detect characteristic acceleration and deceleration of groups of players when actions change somewhat.

2.2 Model Checking

To recognize highlights from the occurrence of the visual cues, operational models are used to perform model checking over the FSMs that model the highlights.

The combination of the measures of the visual cues that are estimated, are checked against each highlight model in parallel. The model checking algorithm works as follows: in the main loop, the visual features that are used to detect the occurrence of highlights (e.g. line segments, players' blobs, playfield color) are extracted from each frame. From these features, descriptors related to the three visual cues previously discussed are computed. Visual cues are discretized: for example, for soccer videos we have 12 possible values for the cue playfield zone, 5 values (both in horizontal and vertical direction) for the camera motion, and 3 different values for the player position descriptor. Hence, a 4-dimensional vector is input in all models at each instant, and the constraints associated with transitions from the current state are checked. If a constraint is verified, the current state is updated, leading either to an advancement in the model evolution, or to a rejection of the current segments (hence resetting the model). Whenever a model progresses from the initial state, the current frame number is stored, to mark the beginning of a possible highlight; if the model succeeds-i.e. a highlight is identified-the current frame number is also stored to mark the end of an actual highlight, otherwise the initial frame number information is discarded.

3 Player Identification

Player identification in sports videos is a complex task, mainly because of player's fast motion and frequent occlusions. Only close-up views are useful for recognition; however also in close–up views players may exhibit large variations in pose and expression, making them sometimes hard to recognize even for a human observer. On the positive side, close-up shots in sports videos have a strong visual appearance. In fact, players wear colored jerseys, usually decorated with some stripes or logos, and most important, showing the player's number. The player's jersey number is unique during an international tournament, and can be used to recognize players identity either analyzing a graphic screen, or checking an existing database, such as those available on the UEFA Euro 2004 website. Superimposed text captions are also shown to point out some interesting details about the player currently framed. They can be used as well to extract important information that is useful for the player identification.

These considerations lead to the fact that, for the purpose of player's identification, face recognition is not the only possible approach. We decided to exploit the information present in close-up shots with frontal faces and superimposed text captions and or the player's jersey number. After this first step, non-identified faces are in turn analyzed in order to understand if a face is similar to any of the faces already annotated using text or jersey's number.

In the following we discuss in detail the solutions adopted for each separate detector and players' annotation.

3.1 Face Detection

Detection of faces is achieved trough an implementation of the algorithm proposed by Viola and Jones [18]. We briefly outline here the algorithm, referring

the reader to the original paper by Viola and Jones, and their subsequent work. Basically, the algorithm employs several simple classifiers devoted to signal the presence of a particular feature of the face, like the alignment of the two eyes or the symmetry in frontal views. A large number of these simple features is initially selected. Then, a classifier is constructed by selecting a small number of important features using AdaBoost [8]. A feature is weighted combination of pixel sums of two rectangles, and can be computed for each pixel in constant time using auxiliary images like the *Summed Area Table* (SAT), which is defined as follows:

$$SAT(x, y) = \sum_{i \leq y} \sum_{j \leq x} I(i, j)$$

where I is the original image. Rotated version of the SAT are employed to compute rotated features. The current algorithm uses the templates of Fig. 3 to compute features: Computation of a single feature f at a given position (x, y) requires to subtract the sum of the intensity values of all the pixels lying under the white rectangle of a template (p_w) from the sum of the intensity values of all the pixels lying under the black rectangle (p_b) of the same template: $f(x, y) = \sum_i p_b(i) - \sum_i p_w(i)$.

Fig. 3. Rectangle features by the face and number detection algorithm

In the current implemementation, the algorithm has been trained to detect frontal and quasi-frontal faces. Training has been carried out with a few hundreds of positive examples taken from a standard face dataset, and another 100 examples manually cropped from soccer video sequences. To deal with the problem of false detection we defined a face verification procedure which is run within the bounding box of each hypothesized face. For each detected face, we produce a color histogram of the region immediately below the face bounding box. The histogram is applied to the Hue component in the HSV color space, normalized w.r.t. white, black, and 5 shades of gray. This histogram is clearly dominated by the principal color of the team jersey, which shows histograms for players belonging to different teams in the same game. For each detected faces, this context color histogram c is compared to a reference histograms r, using the χ^2 statistics. As a second verification step we perform eye detection directly using the intensity values. Pixels of the region of interest are first transformed into the *YCrCB* color space. Then, the eye map is obtained, combining two separate luminance and chrominance maps. Once the shapes present in the final map have been separated, roundness is checked to assess if they correspond to eye pupils value is greater than a threshold the shape is considered as possible eyes. After that, the position of the eye is considered and a region that has two eye-like regions in the appropriate positions is finally considered to be a face.

3.2 Detection and Recognition of Jersey's Numbers

Detection of numbers depicted on player's jerseys is achieved using the same approach as for faces. Official rules of most important soccer organization (like UEFA and FIFA) state that jerseys must be decorated with such numbers on their front, and that size of the numbers must be within a certain range. Moreover, numbers are in the range from 1 to 22, and remains assigned to each player for the entire tournament. We train a different detector for each number from 1 to 22. We found that this approach is far more reliable that having classifiers for digits 0-9, because two digits numbers are not always well separated, and so they tend to cause missed detections. Moreover, detecting each digit separately would force us to impose constraints on spatial arrangement of detected digits which are not easy to verify in the cases where numbers are not perfectly horizontal.

Each detector acts as a dichotomizer, allowing us to directly recognize which is the particular number that has been detected. Each classifier has been trained with 50 positive and 100 negative examples, the latter being randomly selected from images, while the former have been manually cropped. Other positive examples have been generated with graphic programs or obtained by small rotations of some selected images. Templates for numbers are such that bounding box side is about 30 pixels wide.

3.3 Superimposed Text Detection and Recognition

To locate text captions containing player's name and other useful information, we exploited typical production rules of sports videos. These are basically the fact that to enhance readability of characters, producers use luminance contrast (luminance is not spatially sub-sampled in the TV standards) and captions with names of athletes occupy a horizontal box. The algorithm for superimposed text detection we have developed is based on spatio-temporal analysis of image corners and has been described in detail in [4]. An image location is defined as a corner if the intensity gradient in a patch around is distributed along two preferred directions (non-isotropic condition). Therefore, in correspondence with corners the gradient auto–correlation matrix has large and distinct eigenvalues. Corners are detected from:

$$\mathbf{A} = \begin{pmatrix} \langle I_x^2 \rangle & \langle I_x I_y \rangle \\ \langle I_x I_y \rangle & \langle I_y^2 \rangle \end{pmatrix}$$

$$c(x, y) = \det \mathbf{A} - k \, \mathrm{tr}^2 \mathbf{A} \text{ with } k = 0.04$$

if $c(x, y)$ is above a predefined threshold, where subscripts denote partial differentiation with respect to the coordinate axis, and brackets indicate Gaussian smoothing. The first term of the equation has a significant value only if both eigenvalues are different from zero, while the second term inhibits false corners along the borders of the image.

Following the text localization several steps for text recognition are performed. They involve binarization, temporal integration and image enhancement.

Fig. 4. Left - examples labeled by means of text or number. Right - an unlabeled example to be assigned to one of the labels. Lines represent possible correct pairings.

In our experiments, we have used a freely available OCR software [9]. The tool provides a good separation between different words, while making some mistakes in character recognition. In order to recognize player's names, we deal with this problem using an approximate string matching algorithm to perform query on a database of players' names.

3.4 Face Matching

To assign every non-identified face to one of the player classes we exploit the fact that players are a fixed and somewhat limited population. More in detail, we considered each annotated example as an individual, avoiding to merge clusters relative to the same player.

An example of this situation is given in Fig.4, where the unlabeled face in the center must be assigned to one of the labeled faces, which are (a subset of) faces annotated by means of number or text caption. The faces on the first row of the left side, and of the first and third row of the right side represent the same player. Considering each row as a distinct individual, we built a compact representation based on local facial features. This has the effects of increasing inter-class distances in our classification task, but at the cost of having an increased number of classes. Hence, for the unlabeled example there are three possible correct pairings. In practice, to label an unknown face, we require to find a face of the same player with a similar pose and expression.

To cope with the large variation of poses and expressions we followed a part-based approach to recognition, similarly to [15], using the SIFT descriptors [12]. We experimented with several part–based representation schemes, and obtained the most satisfying results using three SIFT descriptor centered on the two eyes (20×20 pixel, with the face size normalized to be 80 pixels wide), and on the midpoint of the eyes (15×30 pixels). This choice is motivated by the facts that a) these are the most robust facial features to detect and track troughout the shots and b) the lower part of the face is often characterized by appearance changes due to variation in expression, that exceed those due to identity changes. The basic SIFT descriptor has been modified to avoid to include in the descriptor non-face part of the image. In particular, we rely on skin–maps to adaptively compute the weights of the components of the SIFT descriptor. For each pixel

of the patch, its weight in the descriptor is cut to zero as the pixel falls off the region defined by the skin–map.

The matching process begins by obtain a single face track for each of the frontal faces found in a shot. A face track is a set of consecutive faces of the same player in the same shot. The detected face is used as a starting point to initialize the track. First, a skin-tone model is built for the face. This is done by collecting an histogram in the C_bC_r space of the bounding box, and then using the dominant color as a skin tone. Then, eyes are tracked throughout the shot, using a simple correlation based tracker that uses eyes-centroids as measure. To avoid false track, the eye search is performed within a limited region (10 pixels) centered on the last observation, and completely included in the region delimited by the skin map. As the tracker looses the eye-tracks, the face track is closed and a compact representation of the whole track is produced.

Similarity between face tracks is computed using the minimum distance between the two sets. If U is a face track corresponding to a non–labeled player, and L is a labeled face track, their distance is computed as follows:

$$d(U, L) = \min_{i,j} \|U_i - L_j\|,$$

where U_i and L_j are two 384–length vector, and their distance is measured using the l_1 norm. A single track may be labeled with several labels. A threshold has been set such that no more than three labels are assigned to the same player. In the worst case, correction of multiple annotations must be done manually with little effort.

4 Highlights Detection Results

Experiments have been carried out with videos of soccer games provided by BBC Sports Library and recorded off-air from other broadcasters. Videos recorded at full PAL resolution and 25 fps. The overall test set includes over 100 sequences with typical soccer highlights, of duration from 15 seconds to 1.5 minutes. The relative frequency of the different types of highlights reflects that of a typical soccer game. Tables 1 and 2 show precision, misclassification and miss rates for the principal soccer highlights. It can be noticed that, for most of the highlights, correct detection is close to 90%.

5 Player Identification Results

The player identification method of Sect.3 has been tested on the same material of the experiments on highlight detection. On average, the system selected about 6000 frames for each game, providing 4 minutes of close-up shots with name/face association. The average number of players identified is 12 for game, without repetition. Figure 5 shows key–frames taken from shots where the either a face and a number have been found, or a face and text have been found. Table 3 reports performance of the number, face and text detectors, averaged on the

Table 1. Precision and misclassification rates of soccer highlight automatic annotation

HIGHLIGHT DETECTED	CLIP EVENT					
	Forward launch	Shot on goal	Placed kick	Attack act.	Counter att.	No highlight
Forward launch	**89.75%**	1.67%	0.00%	0.0%	0.00%	8.58%
Shot on goal	1.52%	**93.90%**	0.00%	0.00%	0.00%	4.58%
Placed kick	0.00%	0.00%	**89.75%**	0.00%	0.00%	10.25%
Attack action	1.50%	1.10%	0.00%	**96.40%**	1.00%	0.00%
Counter attack	0.00%	0.00%	0.00%	8.33%	**83.34%**	8.33%

Table 2. Miss rates of soccer highlight automatic annotation

HIGHLIGHT MISSES				
Forward launch	Shot on goal	Placed kick	Attack action	Counter attack
5.12%	13.05%	7.05%	25.00%	20.10%

Fig. 5. Examples of key frames selected by the system from a Euro 2004 game. The face in the last frame was not detected, but the player was correctly labeled using its number.

duration of a game. Reported ground truth (column "present"), is referred to single player close–up shots, those that are of interest for desired annotation. Table 4 reports average results obtained running the system on the duration of a game. Not surprisingly, detection of face-caption shots is most reliable than detection of face-number shots. This is mainly due to misdetections of the face and number detectors, while the closed caption detector correctly detects nearly all the shots where a caption was present. Moreover, the number of close-up shots detected is fairly low if compared with the total number of close-up shots, where identification is not performed because neither jersey's number nor text caption was present. However, it must be noticed that player's close up occurring during the most interesting moments of the game (after a goal for instance) are usually detected by the system.

Table 3. Face, number and text detector performances

Detector	**Present**	**Detected**	**Correct**	**False**	**Missed**
Face	112	98	90	8	22
Numbers	36	24	20	16	4
Text	12	11	11	1	0

Table 4. Detailed results of the annotation of a single game

		Correct
Total number of close-up shots	112	
Face-number shots present	36	
Numbers of distinct players present	18	
Face-number shots detected	24	20
Face-caption shots present	12	
Face-caption shots detected	11	11
Number of annotated shots	31	27
Number of distinct players identified	13	10

Fig. 6. Face matching experiment. Left: ground truth data. 9 players annotated with their name, plus a "null" class, comprised of unlabeled players. Right: results based on face matching. Some examples were not labeled, since they were not found similar to any labeled example. Crosses indicate incorrect assignments.

5.1 Face Matching Results

To test the face matching scheme of Sec. 3.4, we picked 10 correctly identified faces from the various games present in our testbed, and 30 non–labeled face tracks, for which ground truth was manually obtained. Of these, 25 tracks had a matching face in the annotated set, while the other 5 were completely new to the system. Results are shown in Fig. 6. To keep the result more readable, we avoid the multiple labeling scheme described in Sec. 3.4, and we simply assign a face track to the closest face in the labeled dataset. Also, in this experiment the goal is to test the performance of the face matching module, hence we deliberately avoid to use context information, such as the color of the player's jersey, to rule out obvious false matches (e.g., assigning a player to the wrong team).

6 Conclusions and Future Work

We presented solutions to perform automatic annotation of soccer video for the principal highlights and active players by exploiting a limited set of visual cues and a-priori knowledge of the rules and development of the soccer play. Improvements in the performance of player identification might require more discriminative face representation schemes and new and more effective solutions for jersey number detection.

Acknowledgments

This work has been partially funded by the European VI FP, Network of Excellence DELOS (2004-06).

References

1. J. Assfalg; M. Bertini; C. Colombo; A. Del Bimbo, A. and W. Nunziati; "Semantic annotation of soccer videos: automatic highlights identification", *Computer Vision and Image Understanding*, Volume 92, November–December 2003.
2. G. Baldi, C. Colombo, and A. Del Bimbo. "A compact and retrieval-oriented video representation using mosaics." *Proc. 3rd ICVS*, Amsterdam, 1999.
3. Tamara L. Berg, Alexander C. Berg, Jaety Edwards, Michael Maire, Ryan White, Yee-Whye Teh, Erik Learned-Miller, D. A. Forsyth. "Names and Faces in the News", *in Proc. of CVPR*, 2004.
4. M. Bertini, A. Del Bimbo, and P. Pala. "Content-based indexing and retrieval of TV news", *Pattern Recognition Letters*, 22(5), 2001.
5. Datong Chen, Jean-Marc Odobez and Herv Bourlard. "Text detection and recognition in images and video frames", *Pattern Recognition*, Volume 37, March 2004.
6. Ekin, A.; Tekalp, A.M.; Mehrotra, R.; "Automatic soccer video analysis and summarization", *IEEE Transcations on Image Processing*, July 2003.
7. M. Everingham, and A. Zisserman. "Automated Person Identification in Video"., *Proc. of CIVR*, 2004.
8. Y. Freund and R. E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting", In *Proc. of Eurocolt 95*, Springer-Verlag, 1995.

9. GOCR: Open Source Character Recognition. *http://jocr.sourceforge.net/screenshots.html*
10. R. Hartley and A. Zisserman. "Multiple View Geometry in Computer Vision." Cambridge University Press, 2000.
11. S.S. Intille and A.F. Bobick. "Recognizing planned, multi-person action." *Computer Vision and Image Understanding* (1077-3142) 81(3):414-445, 2001.
12. D. Lowe, "Distinctive image features from scale–invariant keypoints", *International Journal of Computer Vision*, 60, 2, 2004.
13. M.Mottaleb and G.Ravitz. "Detection of Plays and Breaks in Football Games Using Audiovisual Features and HMM." In *Proc. of Ninth Int'l Conf. on Distributed Multimedia Systems*, pp. 154-160, 2003.
14. Shin'ichi Satoh, Yuichi Nakamura, and Takeo Kanade, "Name-It: Naming and Detecting Faces in News Videos," *IEEE MultiMedia*, Vol. 6, No. 1, January-March, 1999.
15. J. Sivic, M. Everingham, and A. Zissermann, "Person spotting: video shot retrieval for face sets", *Proceedings of CIVR*, July 2005.
16. C.G.M. Snoek and M. Worring. "Multimodal video indexing: a review of the state-of-the-art", Multimedia, Tools and Applications, Volume 25, January 2005.
17. G. Sudhir, J.C.M. Lee and A.K. Jain, "Automatic Classification of Tennis Video for High-level Content-based Retrieval." *Proc. of CAIVD'98*, pp. 81-90, 1998.
18. P. Viola and M. Jones. *"Rapid object detection using a boosted cascade of simple features"*, In Proc. CVPR, pages 511-518, 2001.
19. Xinguo Yu , Changsheng Xu , Hon Wai Leong , Qi Tian , Qing Tang , Kong Wah Wan. "Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video." In *Proc. of ACM Multimedia*, November 02-08, 2003.
20. Xinguo Yu, Dirk Farin. "Current and Emerging Topics in Sports Video processing". *Proc. of IEEE ICME*, 2005.
21. W. Zhou, A. Vellaikal, and C.C.J. Kuo, "Rule-based video classification system for basketball video indexing." *Proc. of ACM Multimedia 2000 workshop*, 2000.