Semantic Analysis in Sports Video

Yen-Ping Thseng Jenn-Jier James Lien
Robotics Lab.Department of Computer Science and Information Engineering
National Cheng Kung University
No.1, Ta-Hsueh Road, Tainan 701, Taiwan.
E-mail: yenping@csie.ncku.edu.tw E-mail:jjlien@csie.ncku.edu.tw

Abstract*-Sports video has been widely studied due to its tremendous commercial potentials. Therefore this work designs a high-level annotation system for tennis video to profit contents retrieval. First frames that contain tennis court-scenes are extracted and then the 13 important points on the court, defined by the proposed tennis court-point model, are detected and tracked, along with the locations of the two tennis players. The relative positions of the two players on the tennis court are analyzed using a high-level reasoning model to annotate video clips. Finally, the court of the image is mapped to the court model, using perspective projection to recover the shape of the tennis court. The results in VRML are shown in the experimental section.*

Keywords: Semantic analysis, Trajectory-based, Sports video.

1. INTORDUCTION

With the increasing of audio-visual information that is broadcasted or available in prerecorded media, users prefer to actively access information they are interested. Therefore, this research of semantic analysis of video is getting more popular. In [6], it presents a general mid-level representation framework for semantic sports video analysis. In [1,2,4,5,8,9], using domain knowledge to detect semantic scene and filter un-important scenes of sports video, so user can retrieve the video clip they are interested. In [3], this system analyzes the semantic meaning for tennis sports.

How can we know the players are near the net or baseline? This is an interesting thing user wanted to understand. For example, a player who wants to improve his serve-and-volley abilities or a tennis coach wants to give a visual demonstration of passing techniques. Therefore the annotations giving to the tennis video is needed by users. Therefore our goal is to produce useful and high-level annotation like baseline-rallies, net-games, passing-shots and serve-and-volley games, to pertinent segments of tennis video. Figure 1 shows the framework of our high-level annotation system. First we want to extract the frames containing tennis court-scene and then in the initial frame of the video clip, we detect the important thirteen court points and the two tennis players. In the following we introduce our tracking algorithm proposed to the court points and players. Then we use perspective projection model to project the locations of court points of the image to the court model defined by us. The high-level reasoning module analyzes the outputs of these two modules for inferencing different tennis-play events in the video segment. These analyses result in the meaningful high-level annotations for the segments of input video.

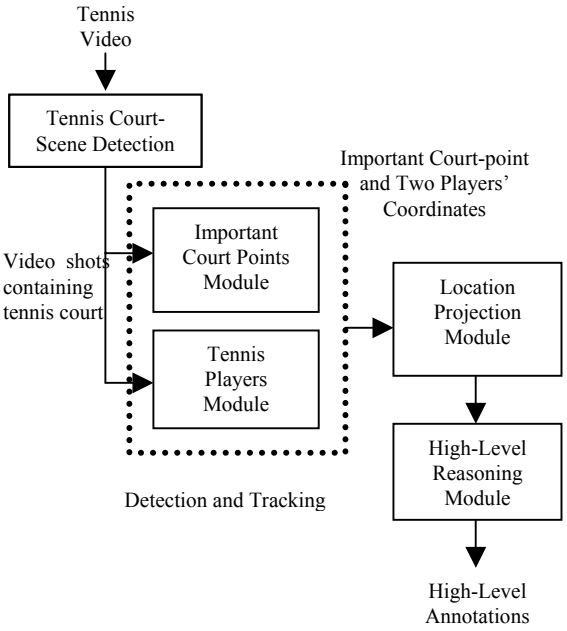


Figure 1. The high-level annotation system

2. COURT-SCENE DETECTION

In this paper, we want to select the video segments containing the tennis court in order to do some semantic analysis process. For this purpose, we first use IBM VideoAnnEx Annotation Tool [10], to execute the action of scene change detection. Then the video clearly segments to several different kinds of scenes, such as player close-up scene, whole court scene and so on [8]. Finally we only choose the video clips that containing the tennis court scene for our experiment.

3. THE IMPORTANT COURT POINTS MODULE

The tennis player's positions on the court are important in analyzing whether a player is near the net, the baseline or another location. Therefore the important tennis court-point model, which defines 13 points on the court, is proposed to determine the relative positions between the player and the court. Hence the 13 points on the court must be tracked in the video clip.

3.1. The Important Court-point Model

Figure 2 presents the important court-point model. These tennis court points are chosen because they are in the corners and can be associated with more textures easy to be detected and tracked. Choosing more points increases the robustness of the system because a court point may disappear or be occluded during tracking. Additionally, the middle points of (P_0 , P_1) and (P_{11} , P_{12}) are identified to determined whether a player near the center of baseline.

The 13 court-points in the first frame of the video clip usually are simultaneously present. Only one or two points on the baseline such as P_1, P_0, P_{12}, or P_{11} sometimes disappear because the camera takes a close-up shot of the motion of a player motion. This situation can be handled by considering the symmetry of the court.

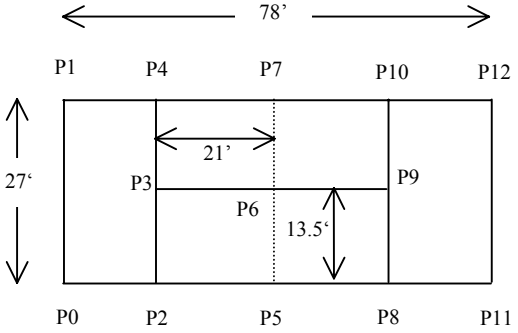


Figure 2. The important tennis Court-point model defined by thirteen points

3.2. Tennis Court points Tracking Algorithm

After the important tennis court-point model has been defined, the 13 tennis court points in the first frame of the video clip of the tennis court are manually located. They are then tracked in the following frames. The tracking algorithm is based on the template-matching method.

3.2.1. System framework. After the 13 court points have been manually located, 13 templates (default size: 17*17) are created. Then, the search window is centered in the coordinates of the court-points and the template is used to determine the region in the search window that is most similar to template using the formula given below. $\rho(x)$ is the correlation coefficient. F(r) is the gray-value of the template, and I(x + r) is the gray-value of the search window of the image.

$$\rho(x) = \frac{\sum_{r\in R}[F(r)-\bar{F}][I(x+r)-\bar{I}(x)]}{\sqrt{\sum_{r\in R}[F(r)-\bar{F}]^2}\sqrt{\sum_{r\in R}I(x+r)-\bar{I}(x)]^2}}$$

If correlation coefficient exceeds 0.9, similar court points in the template matching stage are found; otherwise they are not found. Court points are intersections; which fact is to filter non- intersections in the search window, to improve speed and accuracy. In template matching, similar court points maybe be incorrectly identified, so the tracking wrong points must be found. The translation of this point is compared to the average translation of the correct points. If the error is more than a threshold, this point is regarded as a wrong point, and the new average translation is computed without considering the translation of this point. A virtual point instead of, a wrong or a non-similar point, is generated based on the new translation. The camera motion causes the size of tennis court to vary, so the search window size, the template size and their corresponding thresholds must be varied accordingly. Finally the template of the correct court-point is updated; otherwise the template of virtually created court-point is not updated.

3.2.2. Restrict the search range. The 13 court points are all intersections between a horizontal court-line and a vertical court-line. Accordingly, Sobel edge detection was used to identify the horizontal edge and the results were combined to yield possible intersections.

The Figure 4 depicts the complete intersection detection process. A comparison of the results of two differently directional Sobel edge detections reveals that the vertical directional edge is clear and meets our requirements. Unfortunately, however, the

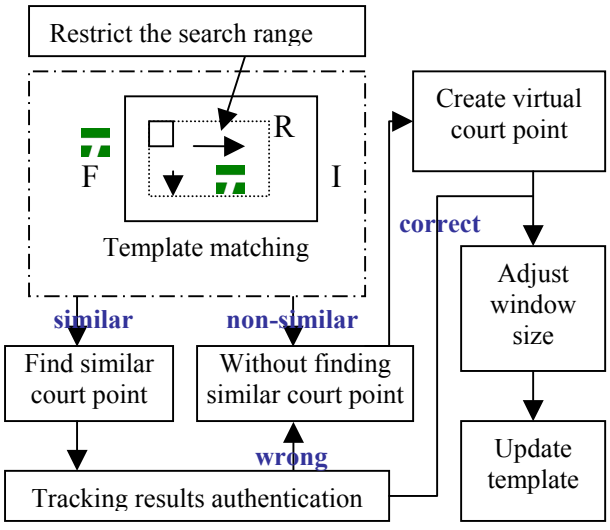


Figure 3. Tennis court points tracking framework

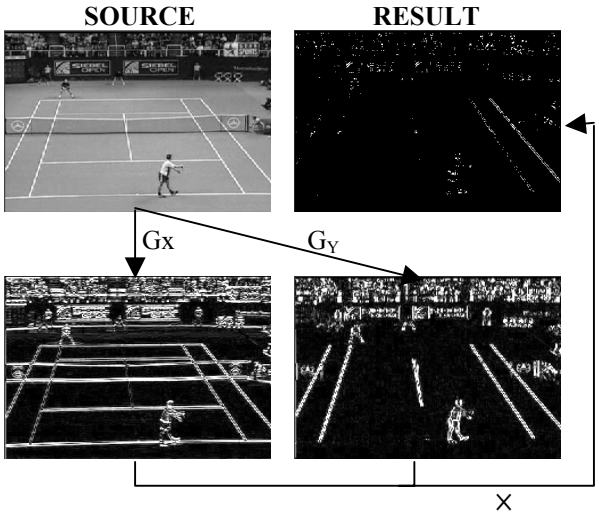
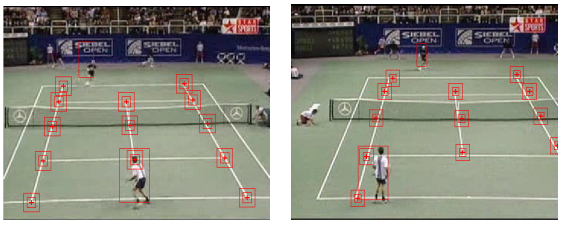


Figure 4. We use the Sobel edge detection in the horizontal direction (Gx) and the Sobel edge detection in the vertical direction (G_Y) to decide the intersection.

horizontal edge is more complicated because a vertical court-line actually has horizontal edge, because the tennis court is a trapezium governed by the depth of image. Therefore the right vertical court-line in the resulting image is detected as an intersection, but this fact does not create a problem for the proposed algorithm, because sufficient important court points are distinguished in the resulting image. Accordingly, the court-point tracking system can handle some important court points that have been wrongly or have temporarily disappeared.

3.2.3. Generate a virtual court-point. Observations of video clips of the tennis court indicate that, some important court points temporarily disappear because the camera moves horizontally or because the points are occluded by player. Figure 5 shows two examples of these situations, in which an appropriate court point cannot be identified. In another situation, a tracking error can be generated if the correlation coefficient exceeds 0.9, but the court-point is far from the required court-point. Therefore a virtual court-point, which is similar to the actual court-point, must be generated. The translation of the camera in the tennis video is slow and stable, and the 13 important court points are observed to be translated similarly



(a) (b)

Figure 5. (a) Occlusion situation: the court-point P_9 is covered by the bottom player's head. (b) Disappearance situation: the court-point P_{12} is disappeared in the image due to camera motion.

during tracking, so the average translation of (dx,dy) of the correctly tracked court points is taken as the new translation of the wrongly tracked court points. Then, a virtual court-point location is created by adding the new translation. If the court-point is virtually created, the template is not updated.

3.2.4. Adaptive search window size and template size. The search window size is initial given 31*31 and the template size is initial given 17*17 in this experiment. The motion of the camera causes the apparent changes in the size of the tennis court. Hence the size of the search window and of the template must be adaptively adjusted. Figure 6 is an example. In Fig. 6, the region to be found is scaled by a factor of three in the horizontal and the vertical directions, because of motion of the camera. In such a case, if the template is not adaptively adjusted, a similar region cannot be found. The line (P_6, P_9) is used to determine scale. After several frames, (ten in this experiment), the scale is computed from the distance (P_6, P_9) in the current frame divided by that in the frame that was ten frames before the current frame.

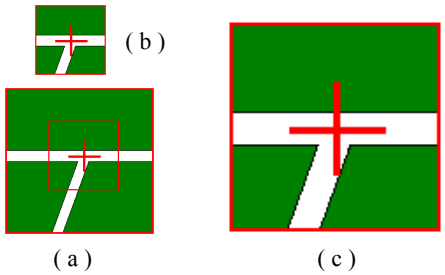


Figure 6. (a) The outer red square is searching window and the inner red square is template. (b) The template of the left bottom (c) Region we wanted to find scaled by three times due to camera motion.

4. TENNIS PLAYERS MODULE

In order to track the players over the video clip, the initial locations of two players in the image have to be located. For this, we use color-based player detection algorithm. Then we track the location of the players in the following frames in the video clip.

4.1. Tennis Player Detection Algorithm

We use the simple region segmentation based on color information to detect the two players. [7] Because the tennis court is almost uniform color, so we can use HS information to distinguish the court color and non-court color. Figure 7 shows a binary image where 0-pixel represent court color and 1-pixel represent non-court color. Because of reduction computation time, the process is performed on the down-sampling images so that the court-lines are nearly no preserved due to down-sampled of frame size. The down-sampling rate in our experiment is 4,

both horizontally and vertically, which results in images wit h size 88*60. And the bottom player closer to camera is on the image bottom location, so we use location information to limit the search range in the bottom image. The top player with long distance from the camera uses the same method. The template size of top player is 20*40, and the template size of bottom player is 40*60 from our observation to the video clip.

4.2. Tennis Player Tracking Algorithm

After the two players in the initial frame of the video clip have been detected, their initial coordinates are obtained, so that the search window in the next frame is restricted to near the player coordinates. The search is restricted to near the player coordinates. The search window is larger than the player template, and the tennis player detection method is used in the search window. The search window in the tracking process is smaller than that used in the detection stage, so the speed is higher. The template matching method is not used, because the template is larger than the space covered by the resulting court-points, the execution speed of template-matching method is therefore lower than that of the presented method.

5. LOCATION PROJECTION MODULE

In all frames of the video clip, the perspective projection model is applied to project the original image having the thirteen important court-point coordinates (Fig. 8(a)) onto the court model defined in Section 4.1.(Fig. 2) The process generates the virtual court that is unaffected by the depth of the image helping us to prevent confusion with the depth of the image. The governed by the depth of the image. Therefore the top baseline appears shorted than the bottom baseline resulting in confusion that for a particular same motion trajectory, the top player's appears to move less far than the bottom player. A

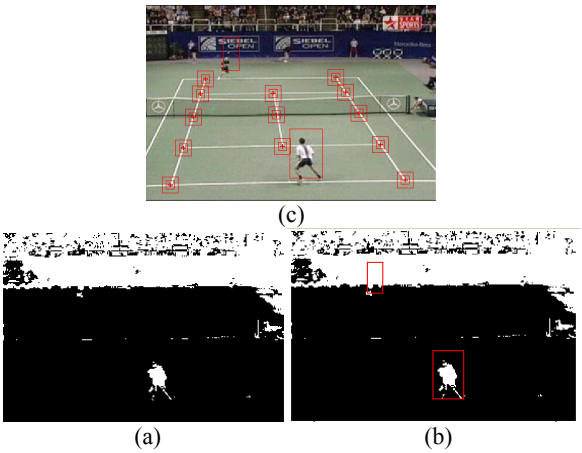


Figure 7. (a) Binarize by court color and non-court color (b) Use template search to find the player region. (c) Map to the source image.

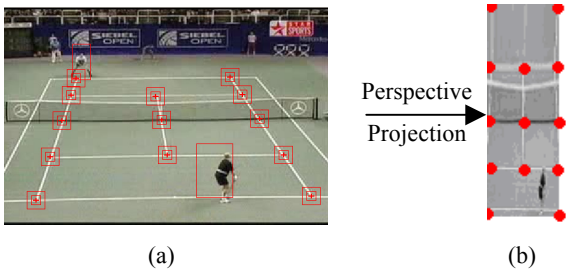


Figure 8. perspective projection from original image (a) to the court model (b)

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} 1 & u & v & 0 & 0 & 0 & u^2 & uv \\ 0 & 0 & 0 & 1 & u & v & uv & v^2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \end{bmatrix}$$

tennis game is best watched from different perspectives. In the experiment herein, the results correspond to a bird's-eye view of the court.

6. HIGH-LEVEL REASONING MODULE

Users are interested in the different tennis video clips. Therefore giving annotation to video clips is more needed. So we propose an idea on the basis of the relative position of the player locations with respect to the tennis court-point. The baselines are composed of the two points (P0,P1) and (P11,P12), the service lines are composed of the two points (P2,P4) and (P8,P10),the net line is composed of the two points (P5,P7) and the center of baseline is the middle point of baseline.

In our experiment, our system could track the important thirteen court points and two players, so we can use this location information to determine the high-level interpretation relevant to the tennis-play event in the video clip. For an instance, the two players are near the baseline during the video clip, and then we annotate the video clip "Baseline-rallies". Table 1 defines other annotations.

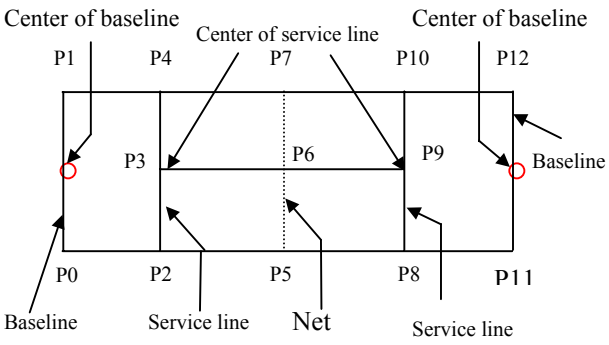


Figure 9. The name of important court-line (reference [2])

Table 1. Mapping low-level positional information to high-level tennis-play events (reference [2]) BL: Baseline, NN: Near the Net, BLC: Center of Baseline, SLC: Center of Service Line

Top Player Location		Bottom Player Location		High-level Annotation
Initial	Final	Initial	Final	
BL	BL	BL	BL	Baseline-rallies
BL	NN	BL	BL	Passing-shot
BL	BL	BL	NN	Passing-shot
BLC	SLC	BL	BL	Serve-and-Volley
BL	BL	BLC	SLC	Serve-and-Volley
SL	NN	SL	NN	Net-game

7. EXPERIMENTAL RESULT

Our test-sets consist of 3 full-games continuously recorded with the MPEG-1 WinTV-USB card, each encoded at 1.15Mbps, CIF-352×240,29.97fps. Here the sources of videos are described below.

Siebel Open San Jose, USA （Aggasi vs. Fish）
Siebel Open San Jose, USA (Roddick vs. Kendrick)
Kroger St. Jude Memphis, USA (Johansson vs.Kiefer)

Figure 10 shows the tracking result of the court-point model. This video clip contains the movement of the camera, hence the location of the court is translated and the size of the court is scaling. Our court-port tracking algorithm is excellent in handling this situation. In Fig. 10 (a), this is the first frame of video clip and the court points are connected as a white court-line according to the standard tennis court. In Fig. 10 (b), the court-point model is re-drawn after each five frames. The video clip has thirty frames, so there are the six shapes of the tennis court in Fig. 10 (b). The result shows that the tracking results of the court-points are correct, so the shape of the court is not deformable.

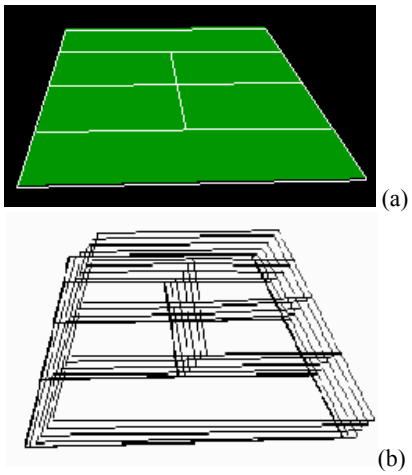


Figure 10. (a) The first frame of the video clip (b) Tennis court tracking result

Implementation of the player tracking, there are some problems in the segmentation of the top player. Because the region of the top player is too small, the background of the top of image is complicated, and the result of the segmentation is sensitive to the background color, we have to modify the tracking result of the top player slightly. The tracking result of the bottom player is successful and useful. Then there are four different kind of semantic events discussed below. For example, in Fig. 11(a), the video clip has 180 frames and the court-point model is re-drawn after each 10 frames, so there are 18 points in the tennis court. The location of the player is represented as an orange sphere. The initial position of the player is represented as a blue sphere and the final position of the player is represented as a red sphere. Finally the red line is represented as the trajectory. Then from Fig.11 (a), Table 1 is proved. Fig. 11(b) shows the result that the court of image is mapped to the court of model. From Fig.11 (b), the court is recovered to a standard rectangle, helping us unaffected by the depth of the image. Figure 12,13,and 14 present another kind of tennis play. We believe these results demonstrate the properties of the high-level annotation system.

Finally by the property of VRML, the tennis court can be rotated to watch this tennis game in the different view. This is a nice choice for users to watch tennis game in the different view.

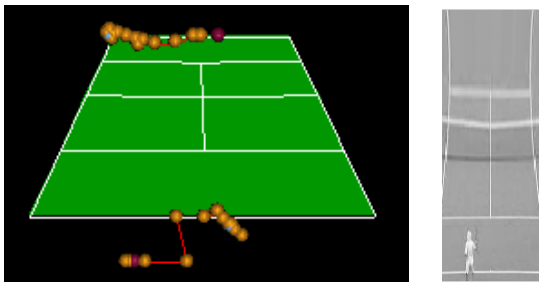


Figure 11. A Baseline-rallies Example: 180 frames (a) Players trajectory (b) the court of image map to the court model using perspective model

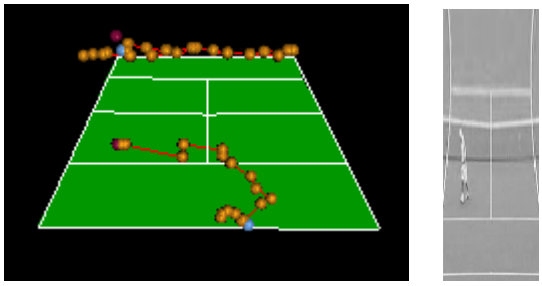


Figure 12. A Passing-shot Example: 220 frames (a) Players trajectory (b) the court of image map to the court model using perspective model

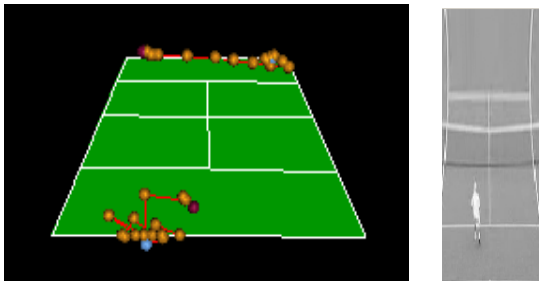


Figure 13. A Serve-and-volley Example: 160 frames (a) Players trajectory (b) the court of image map to the court model using perspective model

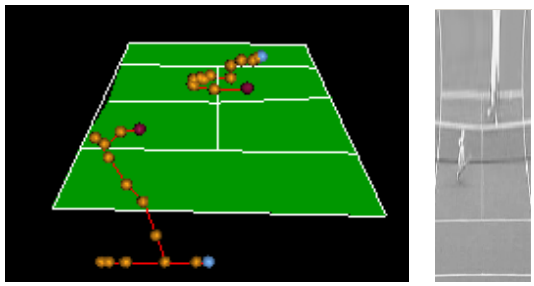


Figure 14. A Net-game Example: 140 frames (a) Players trajectory (b) the court of image map to the court model using perspective model

8. CONCLUSION

In this paper, we describe a high-level annotation system that can detect and track the important thirteen court points defined by the proposed model and then detect and track two tennis players. Then we use perspective projection model to project the location of court-point of the image to the model with standard size of the tennis court. That can help us to analyze the player trajectory without deformed by depth of image due to camera. Then using domain knowledge to analyze tennis-play events in the video segment. We consider relative location of player in the image to determine annotation. In our experiment, we show four examples of the player trajectory to prove our system.

REFERENCE

[1] D. Zhong and S.-F. Chang, "Structure Analysis of Sports Video Using Domain Models," IEEE Conference on Multimedia and Exhibition, Japan, Aug. 2001

[2] G. Sudhir, J.C.M. Lee and A.K. Jain," Automatic classification of tennis video for high-level content-based retrieval, Proc. of the 1998 In Workshop on Content-based Access of Image and Video Databases, January 3, pp.81-99 1998, Bombay, India.

[3] Gopal Pingali, Agata Opalach, Yves Jean and Ingrid Carlbom **"**Instantly Indexed Multimedia Databases of Real World Events**,"** *IEEE Transactions on Multimedia: Special Issue on Multimedia Databases*, vol. 4, no. 2, pp. 269-282, June 2002

[4] Jürgen Assfalg, Marco Bertini, Carlo Colombo, Alberto Del Bimbo, "Semantic Annotation of Sports Videos," IEEE MultiMedia, vol.9 n.2, p.52-60, April 2002

[5] Kongwah Wan, Xin Yan, Xinguo Yu, Changsheng Xu, "Real-time goal-mouth detection in MPEG soccer video," In Proc. of ACM Multimedia 2003: 311-314

[6] L-Y Duan, M Xu, T-S Chua, Q Tian, C-S Xu, "A Mid-level Representation Frame-work for Semantic Sports Video Analysis," In Proc. of ACM Multimedia' 03, ISBN:1-58113-722-2, Pages: 33-44 . Nov. 2003

[7] M. j Swain, D.H. Ballard, "Color Indexing," International Journal of Computer Vision 7(1): 11-32,1991

[8] Soo-Chang Pei and Fan Chen, "Semantic Detection abd Classification in Sports Videos," In Conference on Computer Vision, Graphics and Image Processing '03, VP-05,

[9] S. F. Chang, D. Zhong, Raj Kumar, "Real-time content-based adaptive streaming of sports video," In IEEE workshop on Content-based Access of Image and Video Libraries, Kauai, HI, Dec. 2001.

[10] http://www.research.ibm.com/VideoAnnEx/