Semantic Concept Mining Based on Hierarchical Event Detection for Soccer Video Indexing

Maheshkumar H. Kolekar, Kannappan Palaniappan
Dept. of Computer Science, University of Missouri-Columbia, MO 65211-2060, USA
Email: {kolekarm, palaniappank}@missouri.edu
Somnath Sengupta
Dept. of Electronics and ECE, Indian Institute of Technology, Kharagpur-721302, INDIA
Email: ssg@ece.iitkgp.ernet.in
Gunasekaran Seetharaman
Air Force Research Lab, RITB, 525 Brooks Road, Rome, NY 13441, USA
Email: Gunasekaran.seetharaman@rl.af.mil

***Abstract*— In this paper, we present a novel automated indexing and semantic labeling for broadcast soccer video sequences. The proposed method automatically extracts silent events from the video and classifies each event sequence into a concept by sequential association mining.**

The paper makes three new contributions in multimodal sports video indexing and summarization. First, we propose a novel hierarchical framework for soccer (football) video event sequence detection and classification. Unlike most existing video classification approaches, which focus on shot detection followed by shot-clustering for classification, the proposed scheme perform a top-down video scene classification which avoids shot clustering. This improves the classification accuracy and also maintains the temporal order of shots.

Second, we compute the association for the events of each excitement clip using *a priori* mining algorithm. We propose a novel sequential association distance to classify the association of the excitement clip into semantic concepts. For soccer video, we have considered *goal scored by team-A, goal scored by team-B, goal saved by team-A, goal saved by team-B* as semantic concepts.

Third, the extracted excitement clips with semantic concept label helps us to summarize many hours of video to collection of soccer highlights such as goals, saves, corner kicks, etc. We show promising results, with correctly indexed soccer scenes, enabling structural and temporal analysis, such as video retrieval, highlight extraction, and video skimming.

***Index Terms*— content-based indexing, semantic concept, event detection, mining, sports, video classification, highlights, sequential association**

I. INTRODUCTION

Effective handling of videos such as browsing, retrieving [1], and editing requires semantic understanding of videos. Semantic video content analysis is quite a challenging problem due to a large variety of video content. However, there are certain similarities among certain types of videos, which can be cues to solve the problem. For example, a news video [2], [3] can be considered as a sequence of video segments which starts with anchor person followed by story units; a sports video [4], [5] as a repetitions of play and break scenes. As stated above, a video is often considered as a sequence of video segments, each of which can be considered as a unit to understand the semantic content or the story of the video. Therefore, structuring videos according to their semantic compositions, while understanding the semantic role of each video segment, is a step toward sematic understanding of the videos.

In recent years, sports video has emerged as important area of research due to its wide viewership, ease of digital archival and huge commercial potential [6], [7]. Moreover, the distribution of sports video across the Internet further increases the need for automatic video analysis, such as quick browsing, video summarization, detecting and recording interesting highlights for later review. Existing approaches of sports video analysis [8] can be broadly classified as genre-specific or genre-independent. Due to dramatically distinct broadcast styles for different sports genres, much of the prior art concerns genre specific approaches. Researchers have targeted the individual sports game such as soccer (football) [9], [10], [11], tennis [12], [13], cricket [4], basketball [5], baseball [14], [15], volleyball [16], etc. These works show that genre specific approaches typically yield successful results within the targeted domain. In comparison with the genre-specific research work, less work is observed for genre-independent studies [17], [16], [18], [19]. For a specific sports event detection task, it is not feasible to expect a general solution that will work successfully across all genres of sports video.

In the case of soccer sports video, Li et. al. [20] proposed rule based algorithm using low-level audio/video features for soccer video summarization. Lefevre et. al. [21] divided audio data into short sequences, which are classified them into three classes such as speaker, crowd, referee whistle. Babaguchi *et. al.* [22] proposed event detection by recognizing the textual overlays from soccer video. Wan *et. al.* [23] proposed dominant speech features to generate soccer highlights. Ding et. al. [24] proposed segmental Hidden Markov Model for view-based soccer analysis. Barnard *et. al.* proposed [25] HMM based framework to fuse audio and video features to recognize the play and break scenes in soccer video sequences. Ren and Jose [26] have proposed a HMM based framework

to extract 'attack' scene from football video. Huang *et. al.* introduced semantic analysis of soccer video system based on Bayesian network.

In [27], authors proposed Support Vector Machine based event detection for soccer video. In [28], authors proposed Finite State Machine based annotation of soccer video. Wang *et. al.* [29] proposed semantic notion of *offense* for event detection of soccer video. Yu. *et. al.* [30] proposed technique to retrieve the football video clips using its global motion information. Recently, researchers [31], [32], [33] have presented the use of tracking the positions of players or ball for soccer video analysis. Wang *et. al.* [9] proposed an automatic approach for personalized sports music video generation.

There have been many successful works in soccer video analysis as mentioned above. But most of these works fail to respond to action-based queries, such as "extract the goal clips out of this soccer sequence", or "extract the saves from this soccer video", "extract goals scored by player-A", "extract all the red card events from the collection of FIFA 2006 world cup matches", etc. Such queries may be always needed for editing and retrieval. To facilitate fast retrieval and highlight generation, automatic indexing of video clips is essential. Successful solution of this problem has to address event based classification. The challenge is that soccer game itself has loose structure. Also, there is a lot of motion and view changes in the soccer. Hence, we propose hierarchical framework for event based classification of soccer video and we extract high level concept label based on semantic concept mining technique.

Our proposed system shown in Figure 1 is composed of two components, i.e. hierarchical event detection and classification and semantic concept mining from the extracted events. First the recorded sports video is pre-processed using operations such as commercial removal, dividing game into slots. The excitement clips are extracted from these preprocessed slots. The pre-processed slots are fed to the level-1 of the hierarchical classifier. We have proposed the excitement clip extraction based on short-time audio energy, since the important activity in any sports corresponds to spectators' cheering and commentators' voice. We have observed that the clips extracted as above contain several events. The events within the extracted clips are detected using hierarchical classification tree based on low-level features of the excitement clip. The detected events $E_1, E_2, ..., E_m$ within the clip are arranged in their temporal order to form video event sequence. Since more number of events will be observed for important activity in the soccer sports video, we discard the clips with shorter video event sequence. The sequential association between the events is computed using *a priori* algorithm [34], [35]. Based on our proposed sequential association distance measure, a concept label is assigned to the clip.

These labeled clips are used for indexing and retrieval, video summarization and highlight generation purpose. The main contributions and the novelty of this paper are

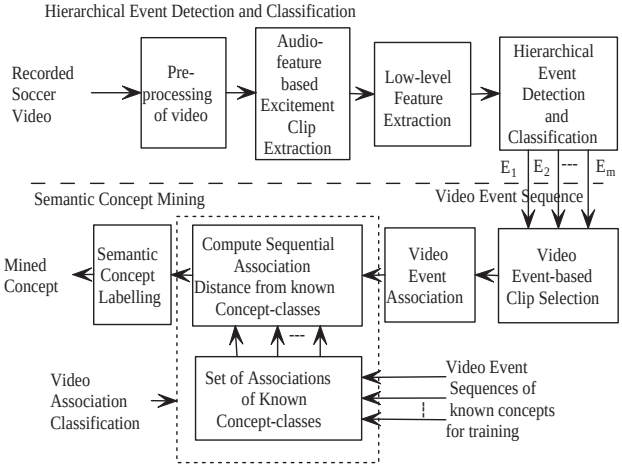

Figure 1. Overall block diagram for semantic concept mining using low-level feature extraction and video event extraction

summarized as follows: (1) We propose novel hierarchical framework for soccer video, (2) We propose novel close-up detection algorithm based on edge detection, (3) We propose novel domain specific close-up classification algorithm based on skin detection and jersey color comparison, (4) Our proposed sequential association distance-based concept mining framework generates retrieval-friendly semantic concept labels. The rest of the paper is organized as follows. Section-II presents proposed hierarchical classification tree. Section-III presents concept extraction based on sequential association rule. Section-IV presents experimental results. Section-V concludes the paper with direction for future work.

II. Hierarchical Classification

Most of the research in sports video processing [36], [37] assumes a temporal decomposition of video into its structural units such as scenes, shots and frames similar to other video domain including television and films. A shot refers to a group of sequential frames often based on single set of fixed or smoothly varying camera parameters (i.e. close-up, medium or long shots, dolly, pan, zoom, etc). A scene refers to a collection of related shots. A clip refers to a collection of scenes. In soccer game, there are moments of excitement, with relatively dull periods in between. Only excitement clips qualify for event detection and indexing.

In [9] the authors propose the integration of multi-modal features such as audio, video and text to detect the semantics of the events. Although the integration of multiple features improves the classification accuracy, it leads to other problems such as proper selection of features, proper fusion and synchronization of right modalities, critical choice of the weighting factor for the features and computational burden. To cope with these problems, we propose a novel hierarchical classification framework for the soccer videos as shown in Figure 2, which has the following advantages: (1) The approach avoids shot detection and clustering that are the necessary steps in most of video classification schemes, so that the

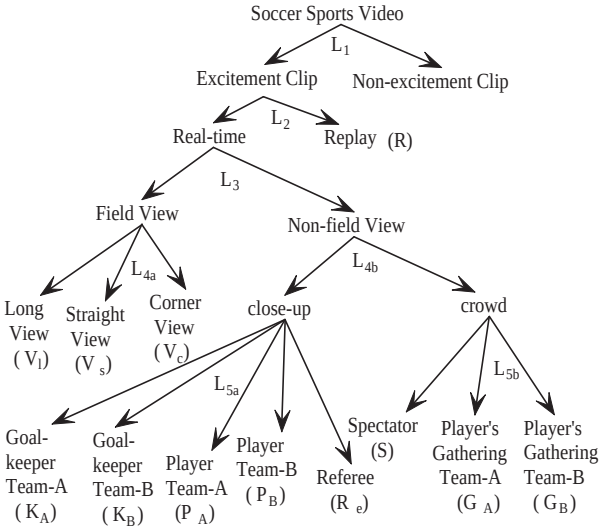

Figure 2. Hierarchical event detection and classification tree

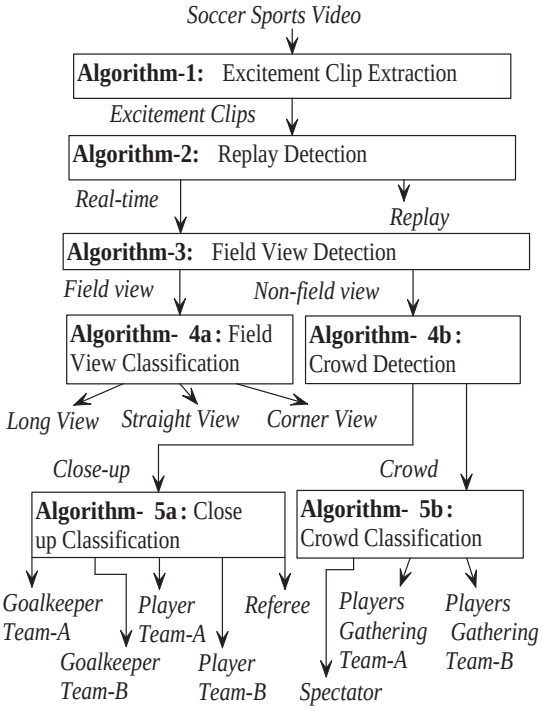

Figure 3. Algorithm data flow for event detection and classification to Figure 2

classification performance is improved. (2) The approach uses top-down five-level hierarchical method so that each level can use simple features to classify the videos. (3) This improves the computation speed, since the number of frames to be processed will remarkably reduce level by level. Figure 3 shows the system diagram for event detection and classification.

A. Definition of Event

We have defined the events as scenes in the video with some semantic meaning (i.e. labels from a semantic hierarchy) attached to it based on the leaf nodes shown in Figure 2. Events are extracted as the leaf nodes of the level-2 to level-5 of hierarchical tree. The events are *Replay, Long View, Straight View, Corner View, Spectator, Player Team-A, Player Team-B, Goalkeeper Team-A, Goalkeeper Team-B, Referee, Players Gathering Team-A, Players Gathering Team-B.*

The low-level feature extraction for the event detection and labeling is discussed in the following subsections.

B. Level-1:Excitement Detection

We have observed that during the exciting events spectator's cheer and commentator's speech becomes louder. Based on this observation, we have used short-time audio energy feature for extracting excitement clip. We are considering the short-time as the number of audio samples corresponding to one video frames. A particular video frame is considered as an excitement frame if its audio excitement exceeds a certain threshold. We propose following steps for excitement clip detection.

Algorithm-1: Excitement Clip Extraction

1. Short-time audio energy $E(n)$, is defined as

$$E(n) = \frac{1}{V}\sum_{m=0}^{V-1}[x(m)w(n-m)]^2 \tag{1}$$

where $x(m)$ is the discrete time audio signal,

$$w(m) = \begin{cases} 1 & \text{if} \quad 0 \le m \le V-1 \\ 0 & \text{otherwise} \end{cases}$$

$w(m)$ is a rectangular window and V is the number of audio samples corresponding to a *single video frame*.

2. Averaging using a sliding window in order to distinguish genuine audio excitement from audio noise. However, it helps for early detection of the events as well.

$$E_1(n) = \frac{1}{L}\sum_{l=0}^{L-1}E(n+l) \tag{2}$$

where, L is the length of sliding window.
The normalized values are as follows:

$$E_2(n) = \frac{E_1(n)}{\max_{1\le i\le N} E_1(i)} \tag{3}$$

where, N is the total number of video frames.

3. Excitement frame detection based on the $E_2(n)$, a video frame n will be finally labeled as $\psi(n) \in [0,1]$ as defined below:

$$\psi(n) = \begin{cases} 1 & (excitement) \quad E_2(n) \ge P_{audio} \\ 0 & (non-excitement) \quad otherwise \end{cases} \tag{4}$$

where, P_{audio} is the mean of $E_2(n)$.

4. We observed that the excitement clips of longer durations are important clips. If all frames in the duration of 20 seconds were all classified as excited frames, then the clip is regarded as excitement clip.

C. Level-2: Replay Detection

Replay events often represent interesting events. Wang *et. al.* [38] used motion and color based features to detect replays. Pan *et. al.* [39] detected slow-motion replays based on logo template. As shown in Figure 4 we observed that replays are generally broadcast with flying graphics (logo-transition) indicating the start and end of the replay. The flying graphics generally last for 10 to 20 frames. Replay segment is sandwiched by two logo-transitions. Since a replay shows many different viewpoints and thus contains many shots in a relatively short period, shot frequency in a replay segment is significantly higher than the average shot frequency in the excitement clip. The color of logo is unique and the size of the logo is big enough to affect the distribution of color histogram as shown in Figure 5. Based on these observations, we propose following algorithm:

Algorithm-2: Replay Detection

1. Convert the input RGB image into HSV image format.
2. Compute 256-bin hue-histogram H_n of the input image.
3. Compute average shot frequency f_c for the excitement clip.
4. Compute hue-histogram H_L of the logo-template.
5. Compute Hue-Histogram Difference (HHD) between frame n and the logo-template using the following formula:

$$HHD(n) = \frac{1}{N_1 N_2} \sum_{i=1}^{256} |H_n(i) - H_L(i)| \qquad (5)$$

where $N_1 \times N_2$ is the image size.
6. Select the logo-transition frames using threshold $P_{HHD} = 0.5 * mean$.
if ($HHD < P_{HHD}$) **then**
frame is logo-transition frame
7. Select the segment between two successive logo-transitions and compute shot frequency f_r.
8. The selected segment is classified as replay segment using following condition.
if ($f_r > f_c$) **then**
selected segment belongs to replay class
else *selected segment belongs to real-time class*

D. Level-3: Field View Detection

We are using a green color pixel ratio similar to [40] to classify the real-time video clips into field view and non-field view. We used 120 representative field view frames from various video (see Table III) to determine the parameters for the *DGPR* method. The combined 256-bin histogram of the hue component of the representative images is used to determine the size of the green window, which was found to be $[G_{min}, G_{max}] = [58, 68]$. Let G_{peak} be the peak in hue histogram with the green

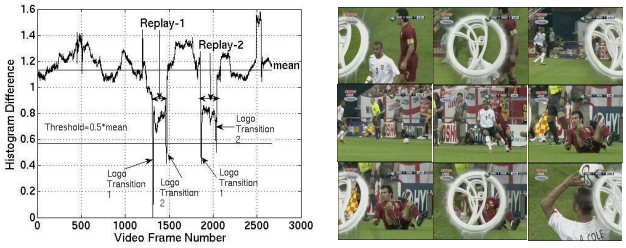

Figure 4. (a) Graph of Hue-histogram Difference with logo template vs. video frame number, (b) Representative frames of replay-2

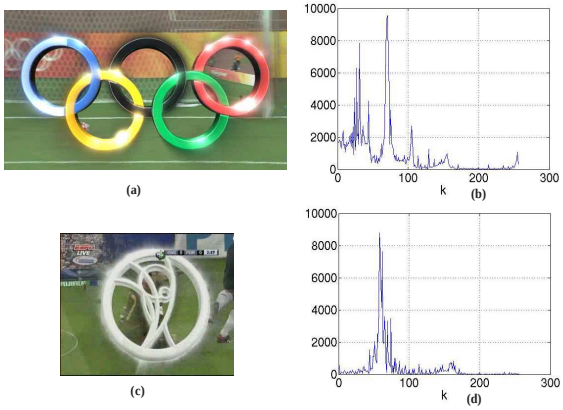

Figure 5. (a) Logo template of Olympic 2008 matches, (b) Hue-Histogram of (a), (c) Logo-template of FIFA 2006 matches, (d) Hue-histogram of (c)

window, $G_{peak} \in [G_{min}, G_{max}]$. The dominant green pixel ratio ($DGPR$) is defined as:

$$DGPR = \frac{H(G_{peak})}{N_1 N_2} * 100\% \qquad (6)$$

where, $H(G_{peak})$ is the number of pixels with value G_{peak} in the histogram for a given frame, and $N_1 \times N_2$ is the total number of pixels in the image. For the field view image shown in Figure 6 (a), we observed $DGPR = 32.75$ % and for non-field view image in Figure 6 (c), we observed $DGPR = 0.3$ %. We observed $DGPR$ values for field view images are greater than 16 %.

Algorithm-3: Field View Detection

1. Convert the input RGB image into HSV image format.
2. Plot histogram of the hue component of the image.
3. Select the G_{peak} window as $G_{peak} \in [G_{min}, G_{max}]$
4. Compute $DGPR$.
5. Classify the image using $DGPR$ threshold $P_{DGPR} = 16\%$
if ($DGPR > P_{DGPR}$) **then**
frame belongs to class field view
else *frame belongs to class non-field view*

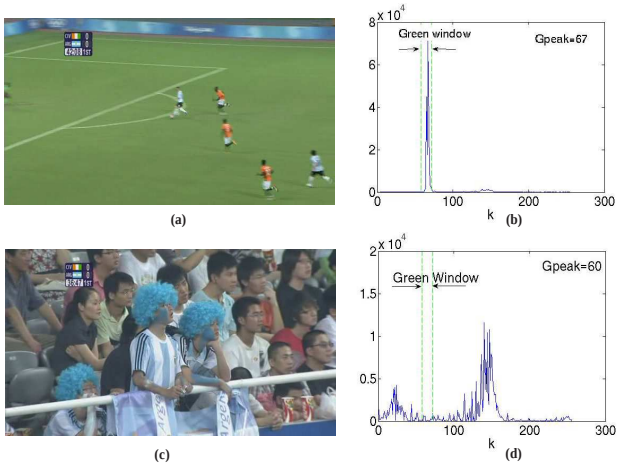

Figure 6. Typical greenness index distributions: (a) Field view image, (b) Hue-histogram of (a), (c) Non-field view image, (d) Hue-histogram of (c)

E. Level-4a: Field View Classification

Under the constant illumination model, the optic-flow equation of a spatiotemporal image volume $\mathbf{I}(\mathbf{x})$ centered at location $\mathbf{x} = [x, y, t]$ is given by Eq. 7 [41] where, $\mathbf{v}(\mathbf{x}) = [v_x, v_y, v_t]$ is the optic-flow vector at $\mathbf{x}$,

$$\begin{aligned} \frac{d\mathbf{I}(\mathbf{x})}{dt} &= \frac{\partial \mathbf{I}(\mathbf{x})}{\partial x}\, v_x + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial y}\, v_y + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial t}\, v_t \\ &= \nabla \mathbf{I}^T(\mathbf{x})\, \mathbf{v}(\mathbf{x}) = 0 \end{aligned} \tag{7}$$

and $\mathbf{v}(\mathbf{x})$ is estimated by minimizing Eq. 7 over a local 3D image patch $\mathbf{\Omega}(\mathbf{x}, \mathbf{y})$, centered at $\mathbf{x}$.

In order to reliably detect only the moving structures *without* performing expensive eigenvalue decompositions, the concept of the *flux tensor* is proposed [41]. Flux tensor is the temporal variations of the optical flow field within the local 3D spatiotemporal volume. Computing the second derivative of Eq. 7 with respect to t, Eq. 8 is obtained where, $\mathbf{a}(\mathbf{x}) = [a_x, a_y, a_t]$ is the acceleration of the image brightness located at $\mathbf{x}$.

$$\begin{aligned} \frac{\partial}{\partial t}\left(\frac{d\mathbf{I}(\mathbf{x})}{dt}\right) &= \frac{\partial^2 \mathbf{I}(\mathbf{x})}{\partial x \partial t}\, v_x + \frac{\partial^2 \mathbf{I}(\mathbf{x})}{\partial y \partial t}\, v_y + \frac{\partial^2 \mathbf{I}(\mathbf{x})}{\partial t^2}\, v_t \\ &+ \frac{\partial \mathbf{I}(\mathbf{x})}{\partial x}\, a_x + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial y}\, a_y + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial t}\, a_t \end{aligned} \tag{8}$$

which can be written in vector notation as,

$$\frac{\partial}{\partial t}(\nabla \mathbf{I}^T(x)\mathbf{v}(\mathbf{x})) = \frac{\partial \nabla \mathbf{I}^T(\mathbf{x})}{\partial t}\, \mathbf{v}(\mathbf{x}) + \nabla \mathbf{I}^T(\mathbf{x})\, \mathbf{a}(\mathbf{x}) \tag{9}$$

Using the same approach for deriving the classic 3D structure, minimizing Eq. 8 assuming a constant velocity model and subject to the normalization constraint $||\mathbf{v}(\mathbf{x})|| = 1$ leads to Eq. 10,

$$\begin{aligned} e_{ls}^F(\mathbf{x}) = \int_{\mathbf{\Omega}(\mathbf{x},\mathbf{y})} &\left(\frac{\partial(\nabla \mathbf{I}^T(\mathbf{y})}{\partial t}\, \mathbf{v}(\mathbf{x})\right)^2 W(\mathbf{x}, \mathbf{y})\, d\mathbf{y} \\ &+ \lambda\left(1 - \mathbf{v}(\mathbf{x})^T \mathbf{v}(\mathbf{x})\right) \end{aligned} \tag{10}$$

Assuming a constant velocity model in the neighborhood $\mathbf{\Omega}(\mathbf{x}, \mathbf{y})$, results in the acceleration experienced by the brightness pattern in the neighborhood $\mathbf{\Omega}(\mathbf{x}, \mathbf{y})$ to be zero at every pixel. The 3D flux tensor $\mathbf{J_F}$ using Eq. 10 can be written as

$$\mathbf{J_F}(\mathbf{x}, \mathbf{W}) = \int_{\mathbf{\Omega}} W(\mathbf{x}, \mathbf{y}) \frac{\partial}{\partial t} \nabla \mathbf{I}(\mathbf{x}) \cdot \frac{\partial}{\partial t} \nabla \mathbf{I^T}(\mathbf{x}) d\mathbf{y} \tag{11}$$

and in expanded matrix form as Eq. 12.

$$\mathbf{J_F} = \begin{bmatrix} \int_{\mathbf{\Omega}} \left\{\frac{\partial^2 \mathbf{I}}{\partial x \partial t}\right\}^2 d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial x \partial t} \frac{\partial^2 \mathbf{I}}{\partial y \partial t} d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial x \partial t} \frac{\partial^2 \mathbf{I}}{\partial t^2} d\mathbf{y} \\ \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial y \partial t} \frac{\partial^2 \mathbf{I}}{\partial x \partial t}\, d\mathbf{y} & \int_{\mathbf{\Omega}} \left\{\frac{\partial^2 \mathbf{I}}{\partial y \partial t}\right\}^2 d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial y \partial t} \frac{\partial^2 \mathbf{I}}{\partial t^2} d\mathbf{y} \\ \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial t^2} \frac{\partial^2 \mathbf{I}}{\partial x \partial t} d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2 \mathbf{I}}{\partial t^2} \frac{\partial^2 \mathbf{I}}{\partial y \partial t} d\mathbf{y} & \int_{\mathbf{\Omega}} \left\{\frac{\partial^2 \mathbf{I}}{\partial t^2}\right\}^2 d\mathbf{y} \end{bmatrix} \tag{12}$$

As seen from Eq. 12, the elements of the flux tensor incorporate information about temporal gradient changes which leads to efficient discrimination between stationary and moving image features. Thus the trace of the flux tensor matrix which can be compactly written and computed as,

$$\mathbf{trace}(\mathbf{J_F}) = \int_{\mathbf{\Omega}} ||\frac{\partial}{\partial t} \nabla \mathbf{I}||^2 d\mathbf{y} \tag{13}$$

and can be directly used to classify moving and non-moving regions without the need for expensive eigenvalue decompositions. Motion-mask is obtained by thresholding and post-processing averaged flux tensor trace. Post-processing include morphological operations to join fragmented objects and to fill holes.

In field view, players and crowd are moving objects and field is non-moving object. Hence, we used motion-mask to classify the frames of the field view as long view, straight view and corner view. Our approach is summarized as follows:

Algorithm-4a: Field View Classification

1. Generate motion-mask for the input field-view frame (see second column of the Figure 7).

2. Apply connected component technique to remove noisy objects from the image (see third column of Figure 7).

3. In the connected component image, background color is the color of object 'field'. Divide the frame into three regions 11, 12, and 2 (see first column of Figure 7).

4. Let FP_2, FP_{11}, FP_{12} be the percentages of Field Pixels in the region 2, 11, 12 of the connected component image respectively. The field-view frame is classified into *long view*, *corner view*, and *straight view* using the thresholds P_1, P_2, P_3 as follows:

if $(FP_2 > P_1) \wedge ((FP_{11} + FP_{12}) > P_2)$ **then**
  frame belongs to class long-view
else if $|FP_{11} - FP_{12}| > P_3$
  frame belongs to class corner-view
else
  frame belongs to class straight-view

F. Level-4b: Close-up detection

We observed that non-field view generally contains only close-up and crowd frames. The percentage of edge

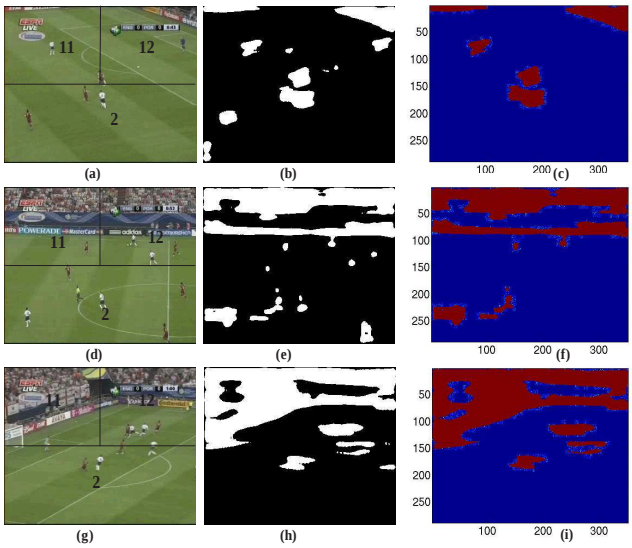

Figure 7. Row-1 shows long view: (a) Image (b) motion-mask (c) connected component image, Row-2 shows straight view: (d) Image (e) motion-mask (f) connected component image, Row-3 shows corner view: (g) Image (h) motion-mask (i) connected component image

pixels (PEP) is used to classify the frame as crowd or close-up, since we typically observe more edge pixels for crowd frames as shown in Figure 8. Any robust edge detector can be used. In our case, we applied Canny edge detector and use the following ratio as the close-up detection parameter:

$$PEP = \frac{Total\ number\ of\ edge\ pixels}{Total\ number\ of\ pixels\ in\ the\ frame} \times 100\% \quad (14)$$

Our approach is summarized in Algorithm-4b.

Algorithm-4b: Crowd Detection

1. Convert the input RGB image into YC_bC_r model.
2. Apply *Canny* operator to detect the edge pixels.
3. Compute Percentage Edge Pixel (PEP) for the image.
4. Classify the image using following condition:
if *($PEP > P_{PEP}$)* **then**
frame belongs to class crowd
else *frame belongs to class close-up*

G. Level-5a:Close up Classification

In sports video, jersey colors are typically differentiate players of different teams, as well as the umpire or referee. At level-5a, the close-up images are classified as *Player Team-A, Player Team-B, Goalkeeper Team-A, Goalkeeper Team-B* or *Referee*. The location of the face of the player in the close-up frame is segmented using skin color information as shown in Figure 9 (b). The connected component technique is applied to remove noisy skin detected objects. The face position will generally occur in the block B_6, B_7, B_{10}, B_{11} depending on the size of close-up as shown in Figure 9 (a). If the number of skin pixels in the block is greater than

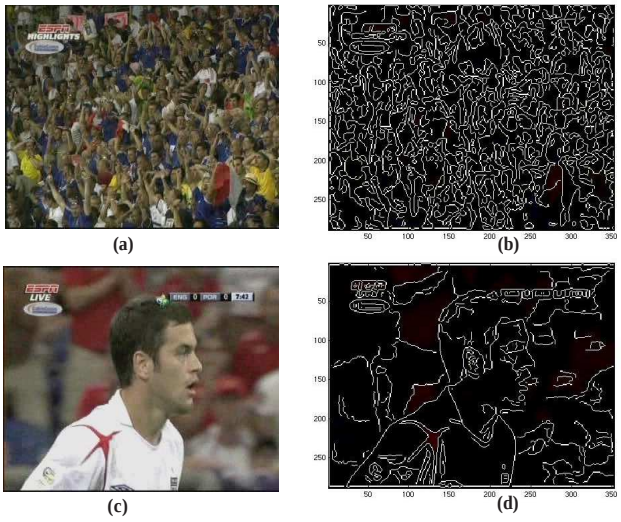

Figure 8. Row-1: (a) Crowd Image, (b) Edge detection results of image (a), Row-2: (c) Close-up Image, (d) Edge Detection results of image (c)

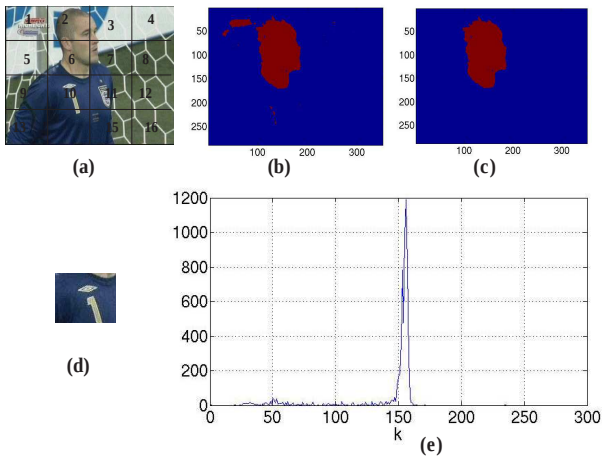

Figure 9. (a) Image of goalkeeper (b) Image showing skin detection, (c) Connected component image, (d) selected block-10, (e) Hue-histogram of (d)

the threshold P_{skin}, then block is considered as skin block. The skin block (face block) locations are used to locate and identify the player's jersey color. For example, if block B_6, B_7, B_{10}, B_{11} are skin blocks, that is $B_6 > P_{skin}$, $B_7 > P_{skin}$, $B_{10} > P_{skin}$, $B_{11} > P_{skin}$, then the skin block binary pattern is 1111 as shown in Table I. Corresponding to these skin block locations, we check B_{14} and B_{15} for the jersey color of the player, which is the binary pattern 0011.

Algorithm-5a: Close-up Classification

1. Convert input RGB image into YC_bC_r image format. Use the following condition for detecting skin pixels.
if $(105 < Y < 117) \wedge (110 < C_b < 113) \wedge (C_r > 128)$,
then *pixel belongs to skin color*
else *pixel does not belong to skin color*
2. Apply connected component technique to remove noisy skin detected objects.

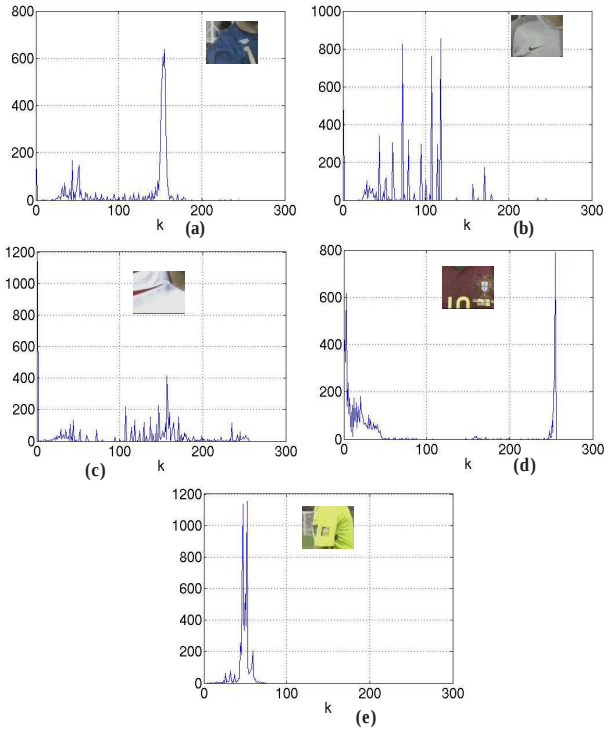

Figure 10. (a) Hue-histogram of Goalkeeper Team-A (class-1), (b) Hue-histogram of Goalkeeper Team-B (class-2), (c) Hue-histogram of Player Team-A (class-3), (d) Hue-histogram of Player Team-B (class-4), (e) Hue-histogram of Referee (class-5)

3. Divide the image into 16 blocks and compute the percentage of skin color pixels in each block.
4. Table I shows the correspondence between the location of skin block in the frame and the location of the associated jersey block.
5. Compute 256-bin hue-histogram H_{JB} of the selected jersey block.
6. Compute the average 256-bin hue-histogram H_k for each close-up class (see Figure 10).
7. Compute the histogram distance of jersey block JB for frame n from the class k using following formula.

$$min_k d_k(JB), d_k(JB) = (\sum_{i=1}^{256}[H_{JB}(i) - H_k(i)]^2)^{1/2} \quad (15)$$

The *Goalkeeper Team-A, Goalkeeper Team-B, Player Team-A, Player Team-B* and *Referee* classes have class labels $k = \{1, 2, 3, 4, 5\}$ respectively.

H. Level-5b:Crowd Classification

At this level, we classify the crowd using jersey color information into *Players Gathering Team-A, Players Gathering Team-B*, and *Spectator* as shown in Figure 11 and Algorithm-5b. We observed that players Gathering generally have field background. Hence, if we set green bins of hue histogram to zero, the error due to background can be removed. Our approach is summarized as follows:

TABLE I.
CORRESPONDENCE BETWEEN SKIN BLOCKS AND JERSEY BLOCKS

Skin Blocks $B_6B_7B_{10}B_{11}$	Jersey Blocks $B_{10}B_{11}B_{14}B_{15}$
1 1 1 1	0 0 1 1
0 0 1 1	0 0 1 1
1 1 0 0	1 1 0 0
1 0 1 0	0 0 1 0
0 1 0 1	0 0 0 1
1 0 0 0	1 0 0 0
0 1 0 0	0 1 0 0
0 0 1 0	0 0 1 0
0 0 0 1	0 0 0 1

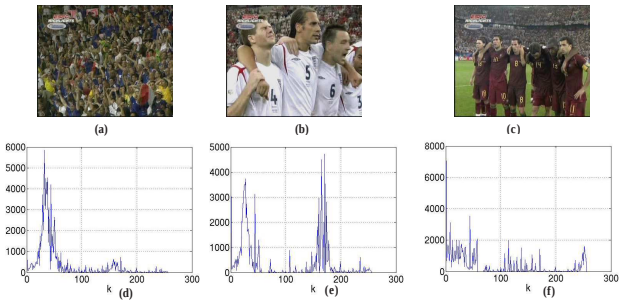

Figure 11. (a) Spectator, (b) Players Gathering Team-A (England), (c) Players Gathering Team-B (Portugal), (d) Hue-histogram of (a) (class-1), (e) Hue-histogram of (b) (class-2), (f) Hue-histogram of (c) (class-3)

Algorithm-5b: Crowd Classification
1. Convert input RGB image n into HSV format.
2. Compute 256-bin Hue-histogram for the input image n.
3. Zero the bin $[G_{min}, G_{max}]$ to remove green background effects.
4. Compute average 256-bin hue-histogram H_k for each crowd class.
5. Compute histogram distance of image n from the class k using following formula.

$$min_k d_k(n), \quad d_k(n) = (\sum_{i=1}^{256}[H_n(i) - H_k(i)]^2)^{1/2} \quad (16)$$

The *Spectator Players Gathering Team-A* and *Players Gathering Team-B* classes have class labels $k = \{1, 2, 3\}$ respectively.

III. SEMANTIC CONCEPT MINING

For semantic concept extraction application, the objective is to capture information about how events in the extracted clip are related to one another. In comparison with the classification rule [42], [43], [44], the association-based technique doesn't need to define the models in advance. Instead, the association mining help us to explore models from video. The leafnodes of the level 2 to 5 of the classification tree of Figure 2 are arranged in their temporal order. The labels are attached to these events to form video event sequence D

for particular excitement clip. The labels assigned to the events are shown in the brackets: *Replay (R), Long View (V_l), Straight View (V_s), Corner View (V_c), Goalkeeper Team-A (K_A), Goalkeeper Team-B (K_B), Player Team-A (P_A), Player Team-B (P_B), Referee (R_e), Spectator (S), Players Gathering Team-A (G_A), Players Gathering Team-B (G_B)*. Semantic Concepts are mined from the video event sequence through a sequential association rule-base.

A. Definition of Concept

Semantic concepts are the collection of a temporally ordered set of events. It can be expressed as $C_i = \cup_{j=1}^{m} E_j$, for $i = 1, 2, ..., z$ where, z is the number of concepts, $E_1, E_2, ..., E_m$ correspond to the events and m is the number of events associated with the i^{th} concept. $C = \{C_i\}_{i=1,2,...,z}$ represents the set of all extracted concepts. We have considered four types of concept-classes *Goal Scored by team-A* ($Goal_A$), *Goal Scored by team-B* ($Goal_B$), *Goal saved by team-A* ($Save_A$), *Goal saved by team-B* ($Save_B$), because these are the most important concepts from the spectator's point of view.

B. Video Event-based Clip Selection

In soccer video, we will observe more number of events during the importance activity. For example, during the goal, we will observe the close-up of the player/goalkeeper who has contributed, the close-up of referee, celebration of the players by gathering, slow-motion replays, etc. Generally, the length of video event sequence $(LVES)$ is longer for exciting activity such as goal or save in soccer video. Hence, we are selecting the clips based on the following criteria:

if $(LVES > P_{LVES})$ **then**

Select the Clip

As shown in Figure 12, we have selected the threshold $P_{LVES} = 4$. Hence, clip-3 has been rejected because it contains only 4 video events.

C. Video Event Association

For a better understanding of video mining from the database point of view, let's assume that D is a transaction database of one consumer, and that each event in D is one transaction. The sequential correlation among the events of D would reflect the association among video events. Table II shows a typical example for computing the sequential association of the video events from the sequence $D = "P_B V_c P_B G_B S G_B R S"$. The terminologies used are as follows:

1: *Video Event* is a database *item* that denotes an events with semantic meaning attached to it. In our example, P_B, G_B, V_c, R, S are video events.

2: *Transactional database D* typically includes a list of sequential patterns. In our example, $D = "P_B V_c P_B G_B S G_B R S"$

3: *L-EventAssociation* is a sequential association that consists of L sequential events. In our example, P_B, G_B, S are *1-EventAssociation*, $P_B G_B$ is *2-EventAssociation*, and $P_B G_B S$ is *3-EventAssociation*.

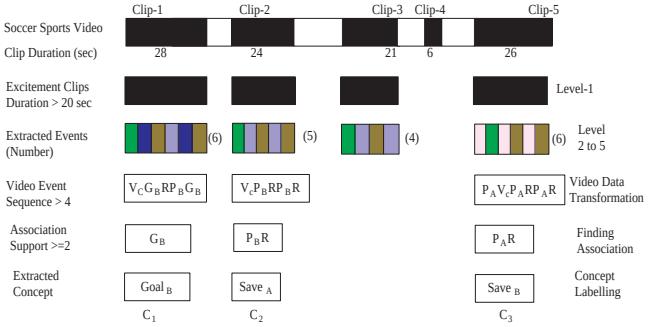

Figure 12. Typical example of concept extraction. Clip-4 is rejected by hierarchical classifier since it has duration smaller than the threshold. Clip-3 is rejected since it has video event sequence not greater than threshold (=4).

TABLE II.
VIDEO EVENT ASSOCIATION FOR $D = "P_B V_c P_B G_B S G_B R S"$

1-L Event Set	*2-Event Set*	*2-LEvent Set*	*3-LEvent Set* (after join)	*3-LEvent Set* (after pruning)	*3-LEvent Set*
$\{P_B\}_2$ $\{G_B\}_2$ $\{S\}_2$	$\{P_B G_B\}_2$ $\{P_B S\}_2$ $\{G_B P_B\}_0$ $\{G_B S\}_2$ $\{S\ P_B\}_0$ $\{S\ G_B\}_1$	$\{P_B G_B\}$ $\{P_B S\}$ $\{G_B S\}$	$\{P_B G_B\ S\}_2$ $\{P_B S\ G_B\}_2$	$\{P_B G_B\ S\}$	$\{P_B G_B S\}$

4: The *Support* of an association is the number of times particular association appears sequentially in the clip. A minimum support threshold requires to be set to extract important association and we have set this as 1. The associations having support larger than 1 are qualified for next level. In our example, P_B, G_B, S have support 2 and V_c, R have support 1 at level-1.

5: *L-EventSet* is an aggregation of all *L-EventAssociation* with each of its member being an *L-EventAssociation*. In our example, *2-EventSet* is $\{P_B G_B, P_B S, G_B P_B, G_B S, S P_B, S G_B\}$.

6: *L-LEventSet* is an aggregation of all *L-EventAssociation* whose *support* is greater than a given threshold. In our case, *2-LEventSet* is $\{P_B G_B, P_B S, G_B S\}$, since these associations have support equal to 2.

We have used *a priori* algorithm to extract the association between the video events. It employs an iterative approach known as level-wise search, where *L-LEventSet* is used to explore *(L+1)-EventSet*. The *a priori* algorithm from the database field [34], [35] is described in Algorithm-6 for ready reference.

As shown in Table II, in the first level, we sequentially scan the given sequence *D* and find the events with their support larger than a threshold. The aggregation of these events form *1-LEventSet* (see column-1 of Table-II). We use these *1-LEventSet* as input to generate the candidates of *2-EventSet* by using the candidate generation algorithm (see column-2). We then scan *D*

again to calculate the support of each association in *2-EventSet*. The associations with their support larger than the threshold are collected to form *2-LEventSet* (see column-3) and generate the candidates of *3-EventSet*. We will iteratively execute this phase until no more non-empty *EventSet* can be found.

The candidate generation function deletes the members in I_k whose subsequences are not in L_{k-1}. Take the *2-LEventSet* in the third column of Table II as an example. If L_2 is given as the input, we will get the *3-EventSet* shown in the fourth column after the join. After pruning out sequences whose subsequences are not in L_2, the sequences shown in the fifth column will be left. For example, the sequence P_BSG_B is pruned out because its subsequence SG_B is not present in L_2 (column-3).

Algorithm-6: *A priori* Algorithm
$I_1 = \{$*1-EventSet*$\}$; $L_1 = \{$*1-EventSet*$\}$;
for $(k = 2; L_{k-1} \neq 0;\ k++)$
{
$I_k \leftarrow$ Candidate_Generation_Function (L_{k-1});
$L_k \leftarrow$ Candidate in I_k with the minimum support;
}

Candidate_Generation_Function (L_{k-1})
1. Join the events of association in L_{k-1}
2. Insert the join results into I_k
Select $\{p.event_1, ..., p.event_{k-1}, q.event_{k-1}\}$
from $\{L_{k-1\ .p},\ L_{k-1\ .q}, p \neq q\}$
where $\{p.event_1 = q.event_1, .., p.event_{k-2} = q.event_{k-2}\}$
3. Delete any member $x \in I_k$ such that some $\{$*(k-1)-EventAssociation*$\}$ of x is not in L_{k-1}

D. Video Association Classification

1) Sequential Association Distance: Let association of known concept as $A(wc_k) = \{X_1, X_2, .., X_P\}$ indicating a sequential association with P events, and association of concept under test as $A(C_T) = \{X_1, X_2, ..., X_Q\}$ with Q events. We propose a distance measure, described as the **SEQ**uential **A**ssociation **D**istance ($SEQAD$) between $A(wc_k)$ and $A(C_T)$ as given below:

$$SEQAD\{A(wc_k), A(C_T)\} = 1 - \frac{|LCS\{A(wc_i), A(C_T)\}|}{min(P,Q)} * \frac{|NCI\{A(wc_i), A(C_T)\}|}{min(P,Q)} \quad for \ \ k = \{1,2,3,4\} \tag{17}$$

where, LCS and NCI are the length of the *Longest Common Subsequence* and *Number of Common events* respectively. Let $A(wc_1)$, $A(wc_2)$, $A(wc_3)$, and $A(wc_4)$ are the associations for the concept $Goal_A$, $Goal_B$, $Save_A$, and $Save_B$ respectively. The $SEQAD$ is used to compute the distances between association of the concept under test and the association of the known concept-class. The label of that concept-class which has minimum $SEQAD$ is assigned to the concept under test.

2) Key Association for known concepts: First, we have computed the association of 180 excitement clips whose concepts are known to us. Then we have selected five frequently occurring associations for the particular concept-class and determined key association in such a way that the length of common subsequence between the key association and any member of five frequently occurring associations is at least 2. We observed the key associations as $A(\omega c_1) = \{P_AG_AS\}$, $A(\omega c_2) = \{P_BG_BS\}$, $A(\omega c_3) = \{P_BR\}$, $A(\omega c_4) = \{P_AR\}$.

3) Example: If $A(c_T) = \{P_BG_B\}$, then

$$SEQAD\{A(\omega c_1), A(c_T)\} = 1 - \frac{|LCS\{P_AG_AS, P_BG_B\}|}{min(3,2)} * \frac{|NCI\{P_AG_AS, P_BG_B\}|}{min(3,2)} = 1 \tag{18}$$

$$SEQAD\{A(\omega c_2), A(c_T)\} = 1 - \frac{|LCS\{P_BG_BS, P_BG_B\}|}{min(3,2)} * \frac{|NCI\{P_BG_BS, P_BG_B\}|}{min(3,2)} = 0 \tag{19}$$

$$SEQAD\{A(\omega c_3), A(c_T)\} = 1 - \frac{|LCS\{P_BR, P_BG_B\}|}{min(2,2)} * \frac{|NCI\{P_BR, P_BG_B\}|}{min(2,2)} = 0.75 \tag{20}$$

$$SEQAD\{A(\omega c_4), A(c_T)\} = 1 - \frac{|LCS\{P_AR, P_BG_B\}|}{min(2,2)} * \frac{|NCI\{P_AR, P_BG_B\}|}{min(2,2)} = 1 \tag{21}$$

Hence, the concept derived from the association $A(c_T)$ is *Goal scored by team-B*, since its association with $A(\omega c_2)$ gives the minimum $SEQAD$ score.

IV. Experimental Results

We have tested our proposed approach using 13 hours of video containing live recordings of FIFA world cup 2006, FIFA world cup 2002, Olympic 2008, Scottish cup 2002, and Champians League 2002 matches as shown in Table III. Since commercials may also be classified as a excitement based on audio excitement, we remove the commercials from the videos before applying hierarchical classifier. We first present results for proposed hierarchical classification tree and then we will present the results of semantic concept mining.

We are extracting excitement clips at level-1 and from level-2 to level-5, we are analyzing the clips to extract the event. This extracted event sequence is going to be used for mining semantic concept for the excitement clip. We are only interested to know whether the particular event is present or absent in the particular excitement clip. Hence, we are presenting clip-based performance instead of frame-based for classifiers of level-2 to level-5. The length of the clips decreases as the level of hierarchy

TABLE III.
SOCCER VIDEO SEQUENCES USED FOR TESTING

Video ID	Name of the Match	Total Duration	A vs B	Date
FIFA-1	FIFA World Cup 2006	94 min	Germany vs Argentina	30/06/ 2006
FIFA-2	FIFA World Cup 2006	95 min	England vs Portugal	01/07/ 2006
FIFA-3	FIFA World Cup 2006	96 min	Brazil vs France	01/07/ 2006
FIFA-4	FIFA World Cup 2002	98 min	Germany vs Korea	25/06/ 2002
Olympic-1	Olympic 2008	96 min	Argentina vs Ivory	07/08/ 2008
Olympic-2	Olympic 2008	96 min	Brazil vs Belgium	07/08/ 2008
Scott-1	Scottish Cup 2002	97 min	Celtic vs Rangers	04/05/ 2002
Champ-1	Champions League 2002	99 min	Real Madrid vs B. Leverkusen	15/05 2002

TABLE IV.
PARAMETERS USED FOR EXPERIMENTATION

Parameter	Definition	Value
P_{audio}	audio energy threshold	mean
P_{HHD}	hue-histogram difference threshold	0.5*mean
$[G_{min}, G_{max}]$	green intensity window size	[58, 68]
P_{DGPR}	dominant green pixel ratio threshold	16%
P_1, P_2	field pixels thresholds	65%,65%
P_3	field pixels difference threshold	10%
P_{PEP}	percentage of edge pixels threshold	8%
P_{skin}	percentage of skin pixels in the block threshold	60%
P_{LVES}	length of video event sequence threshold	4
$Support$	association support	1

increases, because at each level, we segment the clips into sub-clips. The parameters used for the experimentation are given in the Table IV. For measuring the performance of classifiers, we use following parameters: $Recall = \frac{N_c}{N_c+N_m} * 100\%$ and $Precision = \frac{N_c}{N_c+N_f} * 100\%$, where, N_c, N_m, N_f represents the number of clips correctly detected, missed and false positive, respectively.

A. Hierarchical Event Detection

1) Level-1: Excitement Detection: We present here the results of typical soccer video clip of duration 2 minutes 25 seconds. We sampled audio at a rate of 44.1 KHz and video frames are down sampled from its original 30 frames/second to 5 frames/second. We extracted 725 video frames of this video clip. Figure 13 (c) shows the excitement clip selection based on short-time audio energy. The overall performance of the classifier at level-1 is shown in Table V. In case of poor broadcasting quality and noisy audio, performance of audio-based excitement clip extraction decreases.

2) Level-2: Replay Detection: The performance of proposed logo-based replay detection technique is shown in Table VI. Replays generally occur at the end of the excitement clips. If audio is low during the last wipe of the replay, it will not be extracted as a part of the excitement clip, and hence there will be possibility of missing replays.

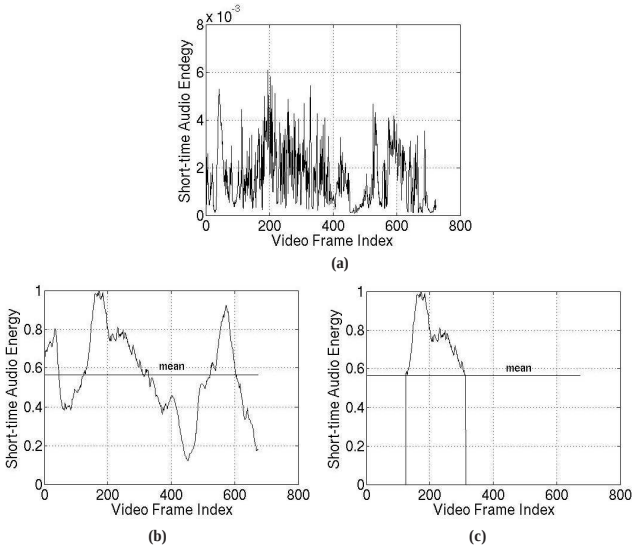

Figure 13. (a) Short-time Audio Energy vs Video Frame Number (b) Short-time Audio Energy vs Video Frame Number with window size 50 frames (c) Frames with Short-time Audio Energy greater than *mean* are selected (Frame # 125 to Frame # 312)

TABLE V.
PERFORMANCE OF LEVEL-1 OF HIERARCHICAL CLASSIFIER

Video ID	Extracted Clips Duration	N_c	N_m	N_f	Recall (%)	Precision (%)
FIFA-1	22 min	24	5	9	82.76	72.73
FIFA-2	29 min	33	6	8	84.62	80.49
FIFA-3	18 min	18	2	7	90.00	72.00
FIFA-4	16 min	10	2	4	83.33	71.43
Olympic-1	13 min	14	2	3	87.50	82.35
Olympic-2	12 min	9	1	3	90.00	75.00
Scott-1	15 min	17	1	3	94.44	85.00
Champ-1	17 min	21	3	2	87.50	91.30

TABLE VI.
PERFORMANCE OF CLASSIFIERS FROM LEVEL-2 TO LEVEL-5

Level	Class	N_c	N_m	N_f	Recall (%)	Precision (%)
2	Replay	256	46	41	84.77	86.20
	Real-time	222	34	39	86.72	85.06
3	Field-view	229	9	11	96.22	95.42
	Non-field-view	291	15	16	95.10	94.79
4a	Long-view	47	6	7	88.68	87.04
	Straight-view	36	6	6	85.71	85.71
	Corner-view	128	15	14	89.51	90.14
4b	Close-up	329	59	55	84.79	85.67
	Crowd	216	42	46	83.72	82.44
5a	Player-A	132	29	25	81.99	83.97
	Player-B	112	27	26	80.58	81.16
	Goalkeeper-A	28	8	10	77.78	73.68
	Goalkeeper-B	30	8	10	78.95	75.00
	Referee	13	3	4	81.25	76.47
5b	Players Gathering-A	77	19	18	88.21	81.05
	Players Gathering-B	67	14	15	82.72	81.70
	Spectator	74	15	15	83.15	83.15

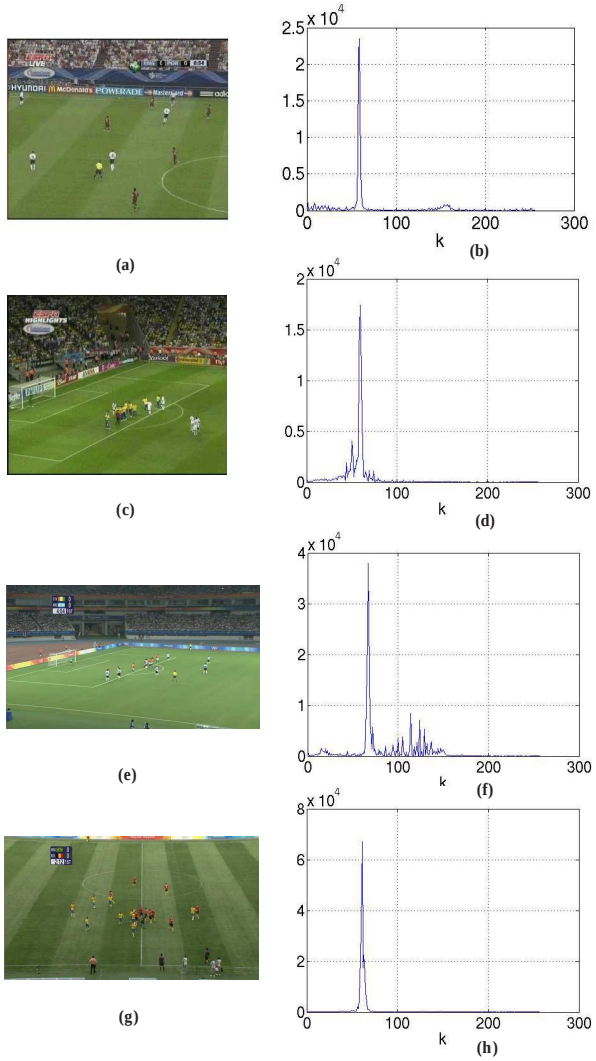

Figure 14. (a) Field view image of video FIFA-2, (b) Hue histogram of (a), (c) Field view of video FIFA-3, (d) Hue-histogram of (c), (e) Field view of video Olympic-1, (f) Hue histogram of (e), (g) Field view of video Olympic-2, (h) Hue histogram of (g)

3) Level-3: Field View Detection: Table VII shows the classification performance of field view detection of the images shown in Figure 14, 15. For non-field view, we observed the values less than 0.1. Because of this discrimination, we observed above 94 % recall and precision as shown in Table VI.

4) Level-4a: Field View Classification: The overall performance of the classifier at level-4a is shown in Table VI. Since we consider few previous frame for generating motion-mask, we observed some miss-classification near the shot boundaries. The detection of corner view is very important from semantic extraction point of view, since it is frequently observed in goal and save concepts of the soccer video.

5) Level-4b: Close-up Detection: Table VIII shows the performance of close-up detection for the images of Figure 16. Since more number of edges are observed in crowd images, feature based on edge pixels offers very good discrimination capability.

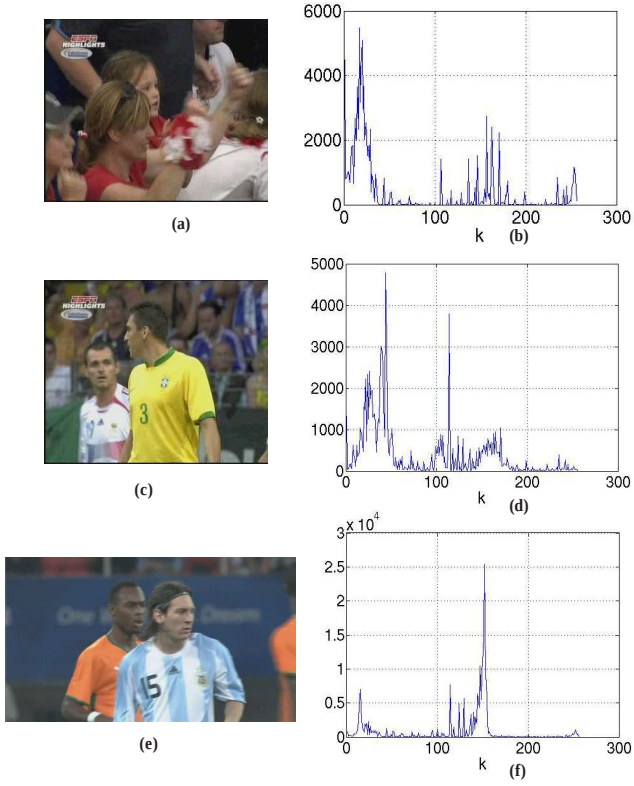

Figure 15. (a) Non-field view of video FIFA-2, (b) Hue histogram of (a), (c) Non-field view of video FIFA-3, (d) Hue-histogram of (c), (e) Non-field view of video Olympic-1, (f) Hue histogram of (e)

TABLE VII.
PERFORMANCE OF LEVEL-3 OF HIERARCHICAL CLASSIFIER

Image	G_{peak}	$DGPR$	Actual class	Observed Class
Fig. 14(a)	59	0.3058	field view	field view
Fig. 14(c)	59	0.2263	field view	field view
Fig. 14(e)	67	0.1728	field view	field view
Fig. 14(g)	61	0.3056	field view	field view
Fig. 15(a)	62	0.0013	non-field view	non-field view
Fig. 15(c)	62	0.0034	non-field view	non-field view
Fig. 15(e)	60	0.0021	non-field view	non-field view

TABLE VIII.
CLASSIFICATION OF NON-FIELD VIEWS INTO CROWD AND CLOSE-UPS AT LEVEL-4B

Fig. 16	Video ID	PEP (%)	Actual Class	Observed Class
(a)	FIFA-1	14.13	crowd	crowd
(b)		12.22	crowd	crowd
(c)		7.38	close-up	close-up
(d)		6.32	close-up	close-up
(i)	FIFA-2	9	crowd	crowd
(j)		10.43	crowd	crowd
(k)		7.62	close-up	close-up
(l)		6.44	close-up	close-up
(q)	FIFA-3	16.51	crowd	crowd
(r)		10.87	crowd	crowd
(s)		6.54	close-up	close-up
(t)		5.17	close-up	close-up

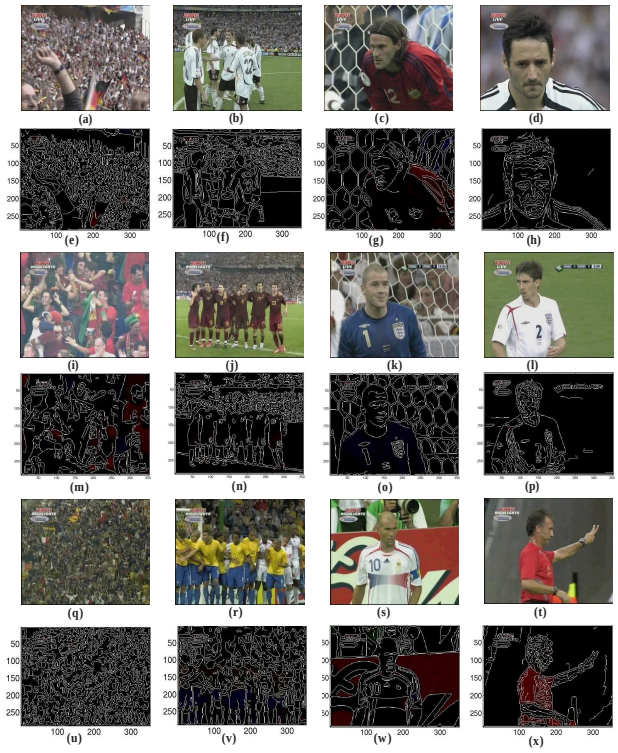

Figure 16. Row-1: Non-field view images of video FIFA-1, Row-2: Edges of the images of row-1, Row-3: Non-field view images of video FIFA-2, Row-4: Edges of the images of row-3, Row-5: Non-field view images of video FIFA-3, Row-6: Edges of the images of row-5

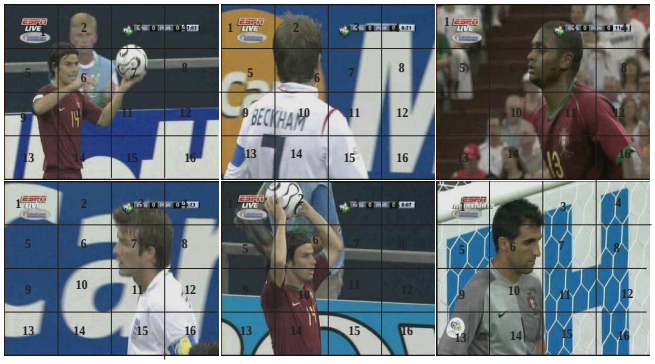

Figure 17. Examples of successfully classified close-up with complex background of video FIFA-2, Row-1: (a) Player Team-B (Portugal), (b) Player Team-A (England), (c) Player Team-B (Portugal), Row-2: (d) Player Team-A (England), (e) Player Team-B (Portugal), (f) Goalkeeper Team-B (Portugal)

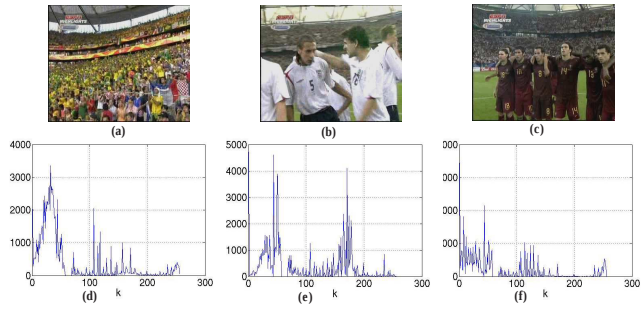

Figure 18. Examples of correctly classified crowd clips of video FIFA-2, (a) Spectator, (b) Players Gathering Team-A (England), (c) Players Gathering Team-B (Portugal), (d) Hue-histogram of (a), (e) Hue-histogram of (b), (f) Hue-histogram of (c)

TABLE IX.
CLOSE-UP CLASSIFICATION OF FIGURE 17

Fig. 17	JB	$d_1(JB), d_2(JB), d_3(JB), d_4(JB), d_5(JB)$	Actual Class	Observed Class
(a)	10	2134, 1793, 1341, 1055, 2116	P_B	P_B
(b)	10	2637, 2326, 1166, 2165, 2889	P_A	P_A
(c)	11	2556, 2174, 1789, 983, 2449	P_B	P_B
(d)	15	2273, 2031, 1170, 1969, 2587	P_A	P_A
(e)	10	2201, 1835, 1374, 1128, 2145	P_B	P_B
(f)	10	2642, 1887, 2035, 2129, 2533	K_B	K_B

TABLE X.
CROWD CLASSIFICATION OF FIGURE 18

Fig. 18	d_1 S	d_2 G_A	d_3 G_B	Actual Class	Observed Class
(a)	8055	10469	9692	S	S
(b)	12589	9473	10601	G_A	G_A
(c)	14347	13206	3431	G_B	G_B

6) Level-5a: Close-up Classification: Figure 17 shows the examples of successfully classified close-ups from video $FIFA-2$. The advantage of our close-up classification scheme is that it gives better results even though there is complex background. The close-up images of Figure 17 have complex background such as banner with large characters, spectator, net in the background. Our classifier has classified these close-up frames successfully as shown in Table IX. As shown in Table VI, we observed the average recall and precision as 80.11% and 78.06% respectively for close-up classification.

7) Level-5b: Crowd Classification: Table X shows the classification of Figure 18 into *Spectator*, *Players Gathering Team-A* and *Players Gathering Team-B*. We observed the average 84.69% and 81.96% recall and precision respectively as shown in Table VI.

B. Semantic Concept Mining

Video event sequence is generated for all extracted excitement clip. The clips with event sequence length less than four are filtered prior to concept mining. We compute $SEQAD$ distance between clip under test and key associations $A(\omega c_1) = \{P_A G_A S\}$, $A(\omega c_2) = \{P_B G_B S\}$, $A(\omega c_3) = \{P_B R\}$, $A(\omega c_4) = \{P_A R\}$. Let $D_1 = SEQAD\{A(\omega c_1), A(c_T)\}$, $D_2 = SEQAD\{A(\omega c_2), A(c_T)\}$, $D_3 = SEQAD\{A(\omega c_3), A(c_T)\}$, $D_4 = SEQAD\{A(\omega c_4), A(c_T)\}$ are the $SEQAD$ distances between the concept under test and key associations of known classes. The concept class which has minimum distance from the concept under test is considered as the class of that concept and its label is assigned as the concept-label. The examples of the association of clips are shown in the Table XI. Figure 19 shows the video event sequences for the clips of the Table XI. The overall performance of the semantic concept mining is shown in the Table XII.

V. CONCLUSION

In this paper, we have presented a hierarchical framework for analyzing high-level events in soccer

TABLE XI.
VIDEO ASSOCIATION CLASSIFICATION

Clip No.	Video ID	Video Event Sequence D	Video Event Association	D_1	D_2	D_3	D_4	Actual Concept	Observed Concept
1	FIFA-1	$V_cG_BRP_BG_B$	G_B	1	0.75	1	1	Goal_B	Goal_B
2	FIFA-2	$V_cP_BRP_BR$	P_BR	1	0.75	0	0.75	Save_A	Save_A
3	FIFA-3	$P_BV_cP_BG_BSG_BRS$	P_BG_BS	0.89	0	0.75	1	Goal_B	Goal_B
4	FIFA-3	$P_AV_cP_ARP_AR$	P_AR	0.75	1	0.75	0	Save_B	Save_B
5	Olympic-1	$V_cP_BV_cG_BSG_BR$	V_cG_B	1	0.75	1	1	Goal_B	Goal_B
6	Olympic-2	$V_cG_ASG_ASG_AR$	G_AS	0	0.75	1	1	Goal_A	Goal_A
7	Scott-1	$V_cP_BRP_BR$	P_BR	1	0.75	0	0.75	Save_A	Save_A

Figure 19. Video Event Sequence for the clips of Table XI

TABLE XII.
PERFORMANCE OF SEMANTIC CONCEPT MINING

Semantic Concept	N_c	N_m	N_f	Recall (%)	Precision (%)
$Goal_A$	12	1	3	92.31	80.00
$Goal_B$	11	1	2	91.66	84.62
$Save_A$	46	5	6	90.20	88.46
$Save_B$	41	5	3	89.13	93.18

video by combing low level feature analysis with high level semantic knowledge. The sports domain semantic knowledge encoded in the hierarchical classification not only reduces the cost of processing data drastically, but also significantly increases the classifier accuracy. The hierarchical framework enables the use of simple features and organizes the set of features in a semantically meaningful way.

The next task addressed in this paper is semantic concept mining based on proposed sequential association distance. We observed promising results for mining of the *Goal scored by team-A*, *Goal scored by team-B*, *Goal saved by team-A*, and *Goal saved by team-B* concepts. Our results demonstrate that our proposed mining technique is effective and we could achieve average recall 90.83% and precision 86.57%.

For soccer video, we have considered only goal and save concept. We are working on to include more number of concepts such as foul, corner kick, free kick, red-card, yellow-card, etc. The proposed semantic mining framework based on hierarchical event classification can be readily generalized to other sports domains as well as other types of video. Our future work includes probabilistic modeling framework for building the semantic hierarchy, semantic concept extraction based on

the classified events for applications such as highlight generation and video summarization.

REFERENCES

[1] N. Dimitrova, H. Zhang, B. Shahraray, I. S. T. Huang, and A. Zakhor, "Applications of video-content analysis and retrieval," *in proc. of IEEE Multimedia*, vol. 9, no. 3, pp. 42–55, 2002.

[2] M. H. Kolekar and S. Sengupta, "Semantic Indexing of News Video Sequences: A Multimodal Hierarchical Approach Based on Hidden Markov Model," *in IEEE Int. Region 10 Conference (TENCON)*, 2005.

[3] M. Shyu, Z. Xie, M. Chen, and S. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *in IEEE Transactions on Multimedia*, vol. 10, no. 2, 2008.

[4] M. H. Kolekar and S. Sengupta, "Event-importance Based Customized and Automatic Cricket Highlight Generation," *in IEEE Int. Conf. on Multimedia and Expo*, pp. 1617–1620, 2006.

[5] C. Xu, J. Wang, H. Lu, and Y. Zhang, "A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video," *in IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 421–436, 2008.

[6] A. Kokaram, N. Rea, R. Dahyot, M. Tekalp, P. Bouthemy, P. Gros, and I. Sezan, "Browsing sports video: trends in sports-related indexing and retrieval work," *in IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 47–58, 2006.

[7] Y. Li, J. Smith, T. Zhang, and S. Chang, "Multimedia database management systems," *in Elsevier Journal of Visual Communication and Image Representation*, pp. 431–440, 2004.

[8] L. Duan, M. Xu, T. Chua, Q. Tian, and C. Xu, "A mid-level representation framework for semantic sports video analysis," *in proc. of ACM Int. Conf on Multimedia*, 2003.

[9] J. Wang, E. Chng, C. Xu, H. Lu, and Q. Tian, "Generation of personalized music sports video using multimodal cues," *in IEEE Transaction on Multimedia*, vol. 9, no. 3, pp. 576–588, 2007.

[10] A. Ekin and A. M. Tekalp, "Automatic soccer video analysis and summarization," *in Sym. Electronic Imaging: Science and Technology: Storage and Retrieval for Image and Video database IV*, 2003.

[11] X. Tong, Q. Liu, and H. Lu, "Shot classification in broadcast soccer video," *in Electronics Letters on Computer Vision and image Analysis*, vol. 7, no. 1, pp. 16–25, 2008.

[12] G. Sudhir, J. Lee, and A. K. Jain, "Automatic classification of tennis video for high-levelcontent-based retrieval," *in proc. of IEEE Int. Workshop on Content-Based Access of Image and Video Database*, 1998.

[13] G. Zhu, Q. Huang, C. Xu, L. Xing, W. Gao, and H. Yao, "Human Behavior Analysis for Highlight Ranking in Broadcast Racket Sports Video," *in IEEE Transactions on Multimedia*, vol. 9, no. 6, pp. 1167–1182, 2007.

[14] C. Cheng and C. Hsu, "Fusion of audio and motion information on HMM based highlight extraction for baseball games," *in IEEE Transaction on Multimedia*, vol. 8, no. 3, 2006.

[15] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," *in eighth ACM Int. Conf. Multimedia*, 2000.

[16] L. Duan, M. Xu, Q. Tian, C. Xu, and J. Jin, "A Unified Framework for Semantic Shot Classification in Sports Video," *in IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1066–1083, 2005.

[17] A. Hanjalic, "Adaptive extraction of highlights from a sport video based on excitement modeling," *in IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1114 – 1122, 2005.

[18] D. Sadlier and N. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," *in IEEE Transaction on Circuits and Systems for Video Technology*, vol. 15, no. 10, 2005.

[19] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang, "Audio events detection based highlight extraction from baseball, golf and soccer games in a unified framework," *in proc. ICASSP*, vol. 3, 2003.

[20] B. Li, H. Pan, and I. Sezan, "A General Framework for Sports Video Summarization with its application to Soccer," *in IEEE Int. Conf. on Audio, Speech, and Signal Processing*, vol. 3, pp. 169–172, 2003.

[21] S. Lefevre, B. Maillard, and N. Vincent, "3 classes segmentation for analysis of football audio sequences," *in Int. Conf. on Digital Signal Processing*, 2002.

[22] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized Abstraction of Broadcasted American Football Video by Highlight Selection," *in IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 575–586, 2004.

[23] K. Wan and C. Xu, "Recent soccer highlight generation with a novel dominant speech feature extractor," *in IEEE Int. Conf. on Multimedia and Expo*, vol. 1, 2004.

[24] Y. Ding and G. Fan, "Segmental Hidden Markov Model for View-based Sports Video Analysis," *in IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[25] M. Barnard and J. Odobez, "Multi-modal audio-visual event recognition for football analysis," *in IEEE Int. workshop on Neural Networks for Signal Processing*, 2003.

[26] R. Ren and J. Jose, "Football video segmentation based on video production strategy," *in Lecture Notes in Computer Science*, vol. 3408, pp. 433–446, 2005.

[27] N. Ancona, G. Cicirelli, A. Branca, and A. Distante, "Goal detection in football by using support vector machines for classification," *in proc. of Int. Joint Conf. on Neural Networks*, vol. 1, 2001.

[28] J. Assfalg, M. Bertini, C. Colombo, A. Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," *in Elsevier Journal on Computer Vision and Image Understanding*, 2003.

[29] L. Wang, M. Lew, and G. Xu, "Offense based temporal segmentation for event detection in soccer video," *in Poc. Of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, 2004.

[30] T. Yu and Y. Zhang, "Retrieval of video clips using global motion information," *in Electronics Letters*, vol. 37, no. 14, 2001.

[31] M. Xu, J. Orwell, and G. Jones, "Tracking football players with multiple cameras," *in IEEE Int. Conf. on Image Processing*, vol. 5, 2004.

[32] Y. Li, A. Dore, and J. Orwell, "Evaluating the performance of systems for tracking football players and ball," *in IEEE Int. Conf. Advanced video and signal based surveillance*, 2005.

[33] P. Nillius, J. Sullivan, and S. Carlsson, "Multi-target tracking- linking identities using bayesian network inference," *in IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2006.

[34] X. Zhu, X. Wu, A. K. Elmagarmid, Z. Feng, and L. Wu, "Video Data Mining: Semantic Indexing and Event Detection from the Association Perspertive," *in IEEE Transcation on Knowledge and Data Engineering*, vol. 17, no. 5, 2005.

[35] R. Agarwal and R. Shrikant, "Fast Algorithm for mining association rules," *in proc. of Int Conf on Very Large Data Bases*, 1994.

[36] V. Chasanis, A. Likas, and N. Galatsanos, "Scene Detection in Videos Using Shot Clustering and Symbolic Sequence Segmentation," *in IEEE Workshop on Multimedia Signal Processing*, 2007.

[37] C. Ngo, T. Pong, and H. Zhang, "On clustering and retrieval of video shots through temporal slices analysis," *in IEEE Transactions on Multimedia*, vol. 4, no. 4, 2002.
[38] L. Wang, X. Liu, S. Lin, G. Xu, and H. Shum, "Generic slow-motion replay detection in sports video," *in IEEE Int. Conf. on Image Processing*, vol. 3, pp. 1585–1588, 2004.
[39] H. Pan, B. Li, and M. Sezan, "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transition," *in IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2002.
[40] P. Xu, L. Xie, S. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," *in IEEE Int. Conf. on Multimedia and Expo*, 2001.
[41] F. Bunyak, K. Palaniappan, S. Nath, and G. Seetharaman, "Flux Tensor Constrained Geodesic Active Contours with Sensor Fusion for Persistent Object Tracking," *in Journal of Multimedia*, vol. 2, no. 4, 2007.
[42] V. Tovinkere and R. J. Qian, "Detecting semantic events in soccer games: towards a complete solution," *in proc. of IEEE Int. Conf. on Multimedia and Expo*, 2001.
[43] W. Zhou, A. Vellaikal, and C. Kuo, "Rule-based video classification system for basketball video indexing," *in Proc. ACM workshop on Multimedia*, 2000.
[44] M.H.Kolekar and S. Sengupta, "A Hierarchical Framework for Generic Sports Video Classification," *in Lecture Notes on Computer Science (LNCS), Springer-Verlag Berlin Heidelberg*, vol. 3852, 2006.