A Novel Framework for Semantic Annotation of Soccer Sports Video Sequences

M. H. Kolekar[1], K. Palaniappan [1], S. Sengupta [2]

[1] Dept. of Computer Science, University of Missouri, Columbia, MO-65211, USA
[2] Dept. of Electronics and Electrical Communication Engg, Indian Institute of Technology, Kharagpur-721302, INDIA

Keywords: Video Content Analysis, Sports Video annotation, Replay, Close-up detection, Semantic Concept Mining

Abstract

A novel framework is presented for semantic labeling of video clips, automatically segmented from broadcast video of soccer (football) games, as highlights and excitement clips etc. The proposed framework provides a generalizable method for linking low-level video features with high-level semantic concepts defined in a commonly understood sports lexicon. Three important contributions are made to automatic annotation of sports video, as follows. First, domain knowledge combined with an event-lexicon and a four-level hierarchical classifier based on low-level video features is used to label video segments. Second, *a priori* event mining is used to establish probabilistic event-associations that are used to assign a concept-lexicon, such as goals and saves, to each highlight video segment. And, finally, the collection of highlight video clips is summarized using concept- and event-lexicons to facilitate highlight browsing, video skimming, indexing and retrieval.

1 Introduction

The well structured nature of data and large commercial interest, make sports video a prime candidate for focused research. It is observed that sports videos, especially soccer (football) videos, are one of the most popular video genres in video-on-demand services. This popularity drives the financial interest around content-based sports video applications. Sports video analysis [5], [6], [10], [9] has received increased attention recently within digital video processing. The broadcast styles of different sports genres vary dramatically. Most of the published results in this area is comprised of genre specific approaches. They have targeted the individual sports game such as soccer [19], [14], [8], tennis [16], [24], cricket [11], basketball [20], baseball [15], [17], volleyball [7], etc.

Video annotation is a basic technique which facilitates introducing and exploiting semantic abstractions in all sports video applications. Ideally, all sports video should be annotated, and the meta-data generated from each video should be stored in a database along with the original video. Such a system would allow an operator to retrieve any important event at a later date. Also, the level of annotation provided by the system should be adequate to facilitate simple text-based queries. Such a system would facilitate many sports video applications including content mining, video indexing, browsing, retrieval, highlight generation.

Manual annotation is both impractical and very expensive, due to the vast amount of data generated at a rapid rate by such videos. However, automatic annotation is a very demanding and an extremely challenging computer vision task as it involves high-level scene interpretation. In [20], authors presented a web-casting text based annotation scheme. In [1], authors proposed Finite State Machine based annotation of soccer video. Barnard et. al. proposed [3] HMM based framework to fuse audio and video features to recognize the play and break scenes in soccer video sequences. Li et. al. [12] proposed rule based algorithm using low-level audio/video features for football video summarization. Babaguchi et. al. [2] proposed event detection by recognizing the textual overlays from football video.

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 team-A", "extract all the red card events from the collection of FIFA 2006 world cup matches", etc. At higher level, user may ask the specific queries such as, "extract the replay segment from the goal clip of this soccer video?" or "extract the close-up of goalkeeper segment from the save clip of this soccer video sequences?". To answer such type of difficult queries, we propose novel framework for automatic annotation of the excitement clips and its events. This work facilitate fast retrieval, highlight generation, automatic indexing.

Our proposed system shown in figure 1 annotates the excitement clip at two-level, one at event-level and other at concept-level. First recorded soccer sports video is pre-processed using operations such as commercial removal [22]. Excitement clips are extracted from soccer sports video using audio cues [13]. Events are extracted using low-level feature based hierarchical classification tree. Each extracted event is annotated by appropriate lexicon (i.e. label). The detected events $E_1, E_2, ..., E_m$ within the clip are considered as items of a relational database D. The sequential association between the database items is computed using *a priori* algorithm [25]. Based on our proposed sequential association distance measure, appropriate lexicon is assigned to the excitement clip. The lexicon consists of words that convey meaning to the user. Each excitement clip will have one concept-lexicon

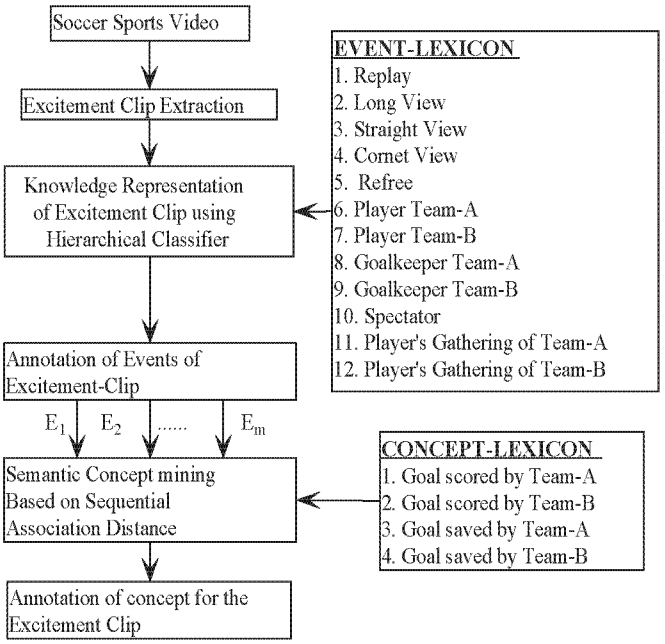

Figure 1: Proposed Semantic Concept Annotation to the Excitement Clip

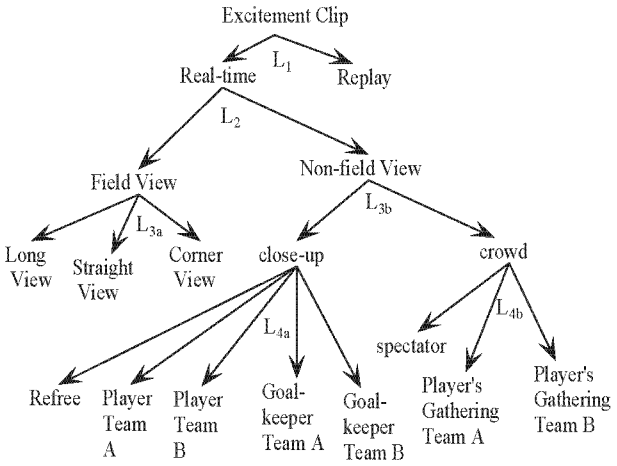

Figure 2: Hierarchical Tree

and several event-lexicons. In this paper, we use 12 lexicons for the annotation of events and 4 lexicons for annotation of concepts.

These annotated events and concepts are used for indexing and retrieval, video summarization, highlight generation purpose. In short, the goal of this work is to propose new algorithms for the extraction and management of the low-level features such as color, edge, motion in the application of storage and retrieval of soccer videos. All algorithms presented are carried out and tested on databases of soccer sports videos. We observed that by applying the low-level features to specific and constrained-application domains, we benefit from domain knowledge to derive potentially more useful semantic information.

The main contributions and the novelty of this paper are summarized as follows: (1) We propose novel hierarchical framework for extracting events of the excitement clip and each event clip is assigned appropriate event-lexicon based on domain knowledge, (2) Our proposed field-view classification based on motion-mask is new and gives very good classification accuracy, (3) We propose novel close-up detection algorithm based on edge detection, (4) We propose novel domain specific close-up classification algorithm based on skin detection and jersey color comparison, (5) Our proposed sequential association distance-based mining framework uses retrieval-friendly semantic concept lexicons.

2 Event Annotation

We have observed that during the exciting events spectator's cheer and commentator's speech becomes louder. Based on this observation, we have used two popular audio content analysis techniques- short-time audio energy and ZCR for extracting excitement clip. The details of the algorithm are published else where. In this section we will discuss, the hierarchical classification of the events of the excitement clips using various low-level feature and automated annotation of these events.

In [19] 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 classification performance is improved. (2) The approach uses top-down four-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.

2.1 Level-1: Replay Detection

As shown in figure 3 we observed that replays are generally broadcasted with flying graphics (logo) indicating the start and end of the replay. The flying graphics generally last for 10 to 20 frames. A pair of such detection gives an estimate of a replay interval. Our approach is summarized as follows:

Algorithm-1:

1: Convert the input RGB frame into HSV image format and plot 256-bins Hue-histogram.

2: Let M by N be the size of the frame. Compute Hue-Histogram Difference (HHD) between frame n and the

Figure 3: Representative frames of the replay segment. Row-1: Representative frames of flying graphics(#690 to #705), Row 2: Representative frames of replay (#706 to #894), Row-3: Representative frames of flying graphics (#895 to #910)

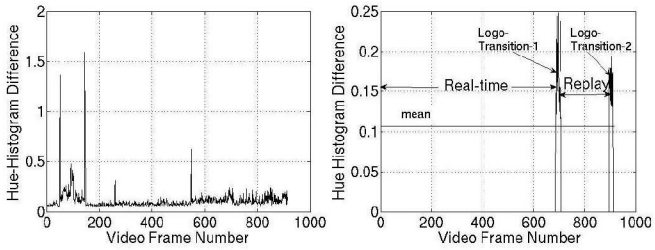

Figure 4: Graphs for the replay segment shown in figure 3. (a) Graph of HHD vs video frame number (b) Logo transition selection

previous frame $n-1$ using the following formula:

$$HHD(n) = \frac{1}{M * N} \sum_{b=1}^{B} |I_n(b) - I_{n-1}(b)| \qquad (1)$$

where B=total number of bins.

3: Plot the graph of HHD vs frame number as shown in Figure 4 (a). Select the frames which has $HHD > mean$.

4: The duration of the logo transition is mostly constant in a game video, because these segments are automatically inserted by editing software. Generally, the duration of the logo-transition is 10-20 frames. Hence, select the shots which has duration between 10 to 20 frames as shown in Figure 4 (b). These shots are the logo-transitions.

5: A 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 clip.

6: If the clip contains shot frequency higher than the average shot frequency in the clip, declare the clip as replay.

2.2 Level-2: Field View Detection

We are using grass-pixel ratio similar to [21] to classify the real-time clips into field view and non-field view. In our experimental set-up, we consider 70 field view images in HSV format for training. We plot 256-bin histogram of the hue component of these images. We pick up the peaks of hue

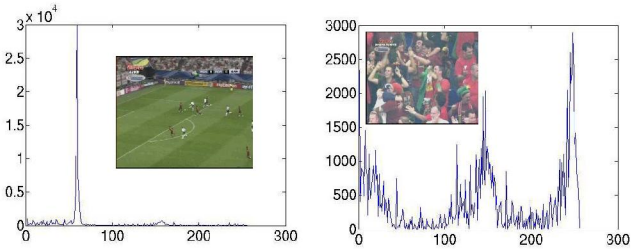

Figure 5: (a) Histogram of hue component of field view image, (b)Histogram of hue component of non-field view image (crowd).

histogram of these images. As shown in Figure 5 (a), we observed peak at bin $k = 59$ and value of the peak is 23486 for the particular image of size 288 by 352. By testing all 70 images, we observed that the green color peak occurs between bin $k = 58$ to $k = 62$. The peak of the histogram gives number of the pixels of the grass in the image. We call this number as x_g. From this, we compute the dominant grass pixel ratio ($DGPR$) as x_g/x, where x is the total number of pixels in the frame. We observed $DGPR$ values vary from 0.16 to 0.24 for the field view. For non-field view image shown in Figure 5 (b), we observed peak for bin $k = 20$ and its value is 2915. The image can be classified as field view or non-field view on the basis of $DGPR$ values. Our approach is summarized as follows:

Algorithm-2:

1:Convert input RGB image into HSV image format.

2: Plot histogram of the hue component of the image

3: Compute $DGPR$ by counting the number of grass pixels.

4:Classify the image using following condition:

if $(DGPR > 0.16)$,

then *frame belongs to class field view*

else *frame belongs to class non-field view*

2.3 Level-3a: Field View Classification

Under the constant illumination model, the optic-flow equation [4] of a spatiotemporal image volume $\mathbf{I}(\mathbf{x})$ centered at location $\mathbf{x} = [x, y, t]$ is given by Eq. 2 where, $\mathbf{v}(\mathbf{x}) = [v_x, v_y, v_t]$ is the optic-flow vector at $\mathbf{x}$,

$$\begin{aligned} \frac{d\mathbf{I}(\mathbf{x})}{dt} &= \frac{\partial \mathbf{I}(\mathbf{x})}{\partial x} v_x + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial y} v_y + \frac{\partial \mathbf{I}(\mathbf{x})}{\partial t} v_t \\ &= \nabla \mathbf{I}^T(\mathbf{x})\, \mathbf{v}(\mathbf{x}) = 0 \end{aligned} \qquad (2)$$

and $\mathbf{v}(\mathbf{x})$ is estimated by minimizing Eq. 2 over a local 3D image patch $\boldsymbol{\Omega}(\mathbf{x}, \mathbf{y})$, 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 [4]. Flux tensor is the temporal variations of the optical flow field within the local 3D spatiotemporal volume.

Computing the second derivative of Eq. 2 with respect to t, Eq. 3 is obtained where, $\mathbf{a}(\mathbf{x}) = [a_x, a_y, a_t]$ is the acceleration of

the image brightness located at $\mathbf{x}$.

$$\frac{\partial}{\partial t}\left(\frac{d\mathbf{I}(\mathbf{x})}{dt}\right) = \frac{\partial^2\mathbf{I}(\mathbf{x})}{\partial x \partial t}\, v_x + \frac{\partial^2\mathbf{I}(\mathbf{x})}{\partial y \partial t}\, v_y + \frac{\partial^2\mathbf{I}(\mathbf{x})}{\partial t^2}\, v_t + \frac{\partial\mathbf{I}(\mathbf{x})}{\partial x}\, a_x + \frac{\partial\mathbf{I}(\mathbf{x})}{\partial y}\, a_y + \frac{\partial\mathbf{I}(\mathbf{x})}{\partial t}\, a_t \quad (3)$$

which can be written in vector notation as,

$$\frac{\partial}{\partial t}(\nabla\mathbf{I}^T(x)\mathbf{v}(\mathbf{x})) = \frac{\partial\nabla\mathbf{I}^T(\mathbf{x})}{\partial t}\,\mathbf{v}(\mathbf{x}) + \nabla\mathbf{I}^T(\mathbf{x})\,\mathbf{a}(\mathbf{x}) \quad (4)$$

Using the same approach for deriving the classic 3D structure, minimizing Eq. 3 assuming a constant velocity model and subject to the normalization constraint $||\mathbf{v}(\mathbf{x})|| = 1$ leads to Eq. 5,

$$e_{ls}^F(\mathbf{x}) = \int_{\Omega(\mathbf{x},\mathbf{y})} \left(\frac{\partial(\nabla\mathbf{I}^T(\mathbf{y})}{\partial t}\,\mathbf{v}(\mathbf{x})\right)^2 W(\mathbf{x},\mathbf{y})\, d\mathbf{y} + \lambda\left(1 - \mathbf{v}(\mathbf{x})^T\mathbf{v}(\mathbf{x})\right) \quad (5)$$

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. 5 can be written as

$$\mathbf{J_F}(\mathbf{x},\mathbf{W}) = \int_{\mathbf{\Omega}} W(\mathbf{x},\mathbf{y})\frac{\partial}{\partial t}\nabla\mathbf{I}(\mathbf{x}) \cdot \frac{\partial}{\partial t}\nabla\mathbf{I^T}(\mathbf{x})d\mathbf{y} \quad (6)$$

and in expanded matrix form as Eq. 7.

$$\mathbf{J_F} = \begin{bmatrix} \int_{\mathbf{\Omega}} \left\{\frac{\partial^2\mathbf{I}}{\partial x\partial t}\right\}^2 d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2\mathbf{I}}{\partial x\partial t}\frac{\partial^2\mathbf{I}}{\partial y\partial t} d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2\mathbf{I}}{\partial x\partial t}\frac{\partial^2\mathbf{I}}{\partial t^2} d\mathbf{y} \\ \int_{\mathbf{\Omega}} \frac{\partial^2\mathbf{I}}{\partial y\partial t}\frac{\partial^2\mathbf{I}}{\partial x\partial t}\, d\mathbf{y} & \int_{\mathbf{\Omega}} \left\{\frac{\partial^2\mathbf{I}}{\partial y\partial t}\right\}^2 d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2\mathbf{I}}{\partial y\partial t}\frac{\partial^2\mathbf{I}}{\partial t^2} d\mathbf{y} \\ \int_{\mathbf{\Omega}} \frac{\partial^2\mathbf{I}}{\partial t^2}\frac{\partial^2\mathbf{I}}{\partial x\partial t} d\mathbf{y} & \int_{\mathbf{\Omega}} \frac{\partial^2\mathbf{I}}{\partial t^2}\frac{\partial^2\mathbf{I}}{\partial y\partial t} d\mathbf{y} & \int_{\mathbf{\Omega}} \left\{\frac{\partial^2\mathbf{I}}{\partial t^2}\right\}^2 d\mathbf{y} \end{bmatrix} \quad (7)$$

As seen from Eq. 7, 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,

$$\mathbf{trace}(\mathbf{J_F}) = \int_{\mathbf{\Omega}} ||\frac{\partial}{\partial t}\nabla\mathbf{I}||^2 d\mathbf{y} \quad (8)$$

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:

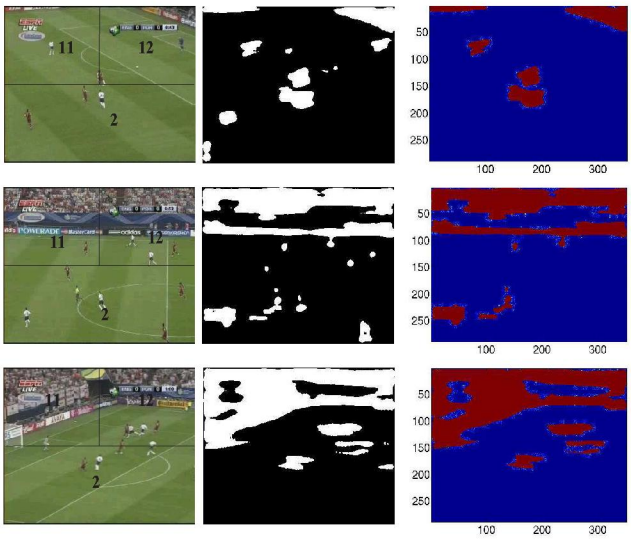

Figure 6: Row-1 shows long view: (a) Image (b) motion-mask (c) connected component image, Row-2 shows straight view: (d) Image (e) motion-mask (f) connected component image, Row-3 shows corner view: (g) Image (h) motion-mask (i) connected component image

Algorithm-3a:

1: Generate motion-mask for the input field-view frame as shown in the second column of the Figure 6.

2: Apply connected component technique to remove noisy objects from the image as shown in the third column of figure 6.

3: In the connected component image, background color is the color of object 'field'. Divide the frame into three regions 11, 12, and 2 as shown in the figure 6(a).

4: Let FP_2, FP_{11}, FP_{12} be the percentage of field pixels in the region 2, 11, 12 of the connected component image respectively. Let T_1, T_2, T_3 be the thresholds. The field-view frame is classified into long view, corner view, and straight view using following condition:

if $(FP_2 > T_1) \bigwedge ((FP_{11} + FP_{12}) > T_2)$,
then *frame belongs to class long-view*
else if $|FP_{11} - FP_{12}| > T_3$
frame belongs to class corner-view
else
frame belongs to class straight-view

2.4 Level-3b: Close-up detection

At this level, we have to classify the frames of non-field views. We observed that non-field view generally contains only close-up and crowd frames. Hence we are classifying the non-field views broadly into close-up and crowd classes. We have proposed the feature based on the percentage of edge pixels (EP) to classify the frame into crowd or close-up, since the edginess is more for crowd frame as shown in figure 7. Our approach is summarized as follows:

Algorithm-3b:

Figure 7: (a) Crowd image, (b) Edge detection results of image (a), (c) Close-up image, (d) Edge detection results of image (c)

Figure 8: Close-up frames frequently observed in broadcasted football video. (a) Player of Team-A, (b) Player of Team-B, (c) Goalkeeper of Team-A, (d) Goalkeeper of Team-B

1:Convert input RGB image into YC_bC_r image format.
2: Apply *Canny* operator to detect the edge pixels.
3: Count the percentage of edge pixels (EP) in the image.
4: Classify the image using following condition:
if $(EP > T_4)$,
then *frame belongs to class crowd*
else *frame belongs to class close-up*

2.5 Level-4a:Close up Classification

In this level, we will classify the close-up images into player of team-A, player of team-B, goalkeeper of team-A, goalkeeper of team-B and referee. We will first find out the location of the face of the person in the close-up using skin or hair color information. In most of the close-up images, the face position will occur in the block 6,7,10,11, as shown in Figure 8. Depending on the face location, we will select the block for checking the jersey color of the player. Our approach is summarized as follows:

Algorithm-4a:
1: Convert input RGB image into YC_bC_r image format. Use the following condition for detecting skin pixels.
if $(105 < Y < 117) \bigwedge (110 < C_b < 113) \bigwedge (C_r > 128)$,
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 from the image as shown in Figure 9.
3: Divide the image into 16 blocks, and compute the percentage of skin color pixels in each block.
4: Let S_6, S_7, S_{10}, S_{11} are the skin percentages of block number $6, 7, 10, 11$ respectively. Select threshold T_5 for considering the block as skin block.
5: If no skin color found in the block $6, 7, 10, 11$, use threshold T_5 for checking hair color (black).
6: Check the condition-x of the Table 1 using skin or hair color block information. If condition-x is satisfied, result-x will give the block number for checking jersey color of the player.
7: Compute 256-bins Hue-histogram $Histo_b$ of the selected block as per condition-x.
8: Compute the average 256-bins Hue-histogram for all the known classes C.
9: Compute the Euclidean distance of the block b of frame n from the class k using following formula.

$$d_k(b) = \sqrt{\sum_{i=1}^{256} [histo_k(i) - histo_b(i)]^2} \quad where\, k = 1, 2, .., C \tag{9}$$

10: Select the value of k for which d_k has lowest value. Assign that value of k as a class-label to the particular frame n.

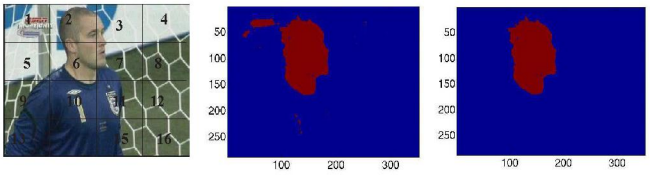

Figure 9: (a) Image of goalkeeper (b) Image showing skin detection, (c) Connected component image

Table 1: Condition for block selection for checking jersey color of the player

x	Condition-x	Result-x (block number for jersey color checking)
1	$(S_6, S_7, S_{10}, S_{11}) > T_5$	14, 15
2	$(S_6, S_{10}) > T_5$	14
3	$(S_7, S_{11}) > T_5$	15
4	$(S_6, S_7) > T_5$	10, 11
5	$(S_{10}, S_{11}) > T_5$	14, 15
6	$S_{11} > T_5$	15
7	$S_{10} > T_5$	14
8	$S_6 > T_5$	10
9	$S_7 > T_5$	11

2.6 Level-4b:Crowd Classification

At this level, we classify the crowd using jersey color information into the following three classes: Player's Gathering of Team-A, Player's Gathering of Team-B, and Spectator. Since players gather on the field after exciting event, in most of the players gathering frames will have field as a background as shown in figure 10. Hence, if we set the dominant field color (green) bins to zero, classification performance will be improved. Our approach is summarized as follows:

Algorithm-4b:
1:Convert input RGB image n into HSV format.
2: Compute 256-bin Hue-histogram for the input image n.
3: Dominant green color occur in the bins 56 to 64. So if we set these bins to zero, the error due to field color will be removed. Let $histo_n$ be the histogram of image n after setting

Figure 10: (a) Players gathering of team-A (England), (b) Players gathering of team-B (Portugal), (c) Spectator

the bins of green color to zero.
4: For all known classes, compute average 256-bins hue-histogram with setting dominant green color bins (i. e. 56 to 64) to zero.
5: Compute histogram distance of the hue-histogram of input image n from the hue-histogram of the class k using following formula.

$$d_k(n) = \sqrt{\sum_{i=1}^{256}[histo_k(i) - histo_n(i)]^2} \; where\, k = 1, 2, .., C \tag{10}$$

6: Find the value of k for which the distance $d_k(n)$ is minimum. Assign label of that class k to the frame n.

2.7 Event

We have defined the events as scenes in the video with some semantic meaning (i.e. lexicons 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 hierarchical tree. The events are replay, long view, straight view, corner view, spectator, player of team-A, player of team-B, goalkeeper of team-A, goalkeeper of team-B, players gathering of team-A, players gathering of team-B.

3 Semantic Concept Annotation

For semantic concept mining, the objective is to capture information about how events in the extracted clip are related to one another. In comparison with the classification rule [18], [23], [13], 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 hierarchical classification tree of figure 2 are arranged in their temporal order. The labels are attached to these events to form video item 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), spectator (S), player of team-A (P_A), player of team-B (P_B), goalkeeper of team-A (K_A), goalkeeper of team-B (K_B), players gathering of team-A (G_A), players gathering of team-B (G_B).

3.1 Video Item 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 item in D is one transaction. The sequential correlation among the items of D would reflect the association among video items. The terminology used in the algorithm is defined as follows:
Video item is a basic unit that denotes an events with semantic meaning attached to it. In our soccer video case, R, V_l, V_s, V_c, S, P_A, P_B, K_A, K_B, G_A, G_B are video items.
Transactional database D typically includes a list of sequential patterns.
L-ItemAssociation is a sequential association that consists of L sequential items. In our case R, V_l, V_s, V_c, S, P_A, P_B, K_A, K_B, G_A, G_B are *1-ItemAssociation*.
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 2. The associations having support less than 2 are disqualified for next level.
L-ItemSet is an aggregation of all *L-ItemAssociation* with each of its member being an *L-ItemAssociation*.
L-LItemSet is an aggregation of all *L-ItemAssociation* whose *support* is no less than a given threshold.
We have used *a priori* algorithm to extract the association between the video items. It employs an iterative approach known as level-wise search, where *L-ItemSet* is used to explore *(L+1)-ItemSet*. The *a priori* algorithm [25] is described below for the ready reference.
Algorithm-5: *A priori* Algorithm
$I_1 = \{$*1-ItemSet*$\}$; $L_1 = \{$*1-ItemSet*$\}$;
for $(k = 2; L_{k-1} \neq 0;\; k++)$
begin
I_k= New candidate generated from L_{k-1};
(See candidate generation function)
L_k= Candidate in I_k with the minimum support;
end

Candidate Generation Function
1: Join the items of association in L_{k-1}
2: Insert the join results into I_k
Select $\{p.Item_1, ..., p.Item_{k-1}, q.Item_{k-1}\}$
from $\{L_{k-1\ .p},\ L_{k-1\ .q}, p \neq q\}$
where $\{p.Item_1 = q.Item_1, .., p.Item_{k-2} = q.Item_{k-2}\}$
3: Delete any member $x \in I_k$ such that some
$\{$*(k-1)-ItemAssociation*$\}$ of x is not in L_{k-1}

3.2 Video Association Classification

Our proposed method for classifying each association into a corresponding concept category is summarized as follows. Given two associations $A(a) = \{X_1, X_2, .., X_P\}$ indicating a sequential association with P items, and $A(b) = \{X_1, X_2, .., X_Q\}$ as another sequential association with Q items, we propose a distance measure, described as the **SEQ**uential **A**ssociation **D**istance ($SEQAD$) between $A(a)$ and $A(b)$ defined as

$$SEQAD\{A(a), A(b)\} = 1 - \frac{|LCS\{A(a), A(b)\}|}{min(P, Q)} *$$

$$\frac{|NCI\{A(a), A(b)\}|}{min(P, Q)} \quad (11)$$

where, LCS and NCI are the length of the *Longest Common Subsequence* and *Number of Common Items* respectively. The $SEQAD$ is used to compute the distances between association of the concept under test and the association of the known concept-class. For example, let $A(\omega c_1)$, $A(\omega c_2)$, $A(\omega c_3)$, $A(\omega c_4)$, and $A(c_T)$ are the associations for concept *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* ($Save_B$) and *Concept under test* respectively.

Example: If $A(\omega c_1) = \{P_A G_A S\}$, $A(\omega c_2) = \{P_B G_B S\}$, $A(\omega c_3) = \{P_B R\}$, $A(\omega c_4) = \{P_A R\}$ and $A(c_T) = \{P_B G_B\}$, then

$$SEQAD\{A(\omega c_1), A(c_T)\} = 1 - \frac{|LCS\{P_A G_A S, P_B G_B\}|}{min(3, 2)} *$$

$$\frac{|NCI\{P_A G_A S, P_B G_B\}|}{min(3, 2)} = 1 \quad (12)$$

Similarly,
$SEQAD\{A(\omega c_2), A(c_T)\} = 0, SEQAD\{A(\omega c_3), A(c_T)\} = 0.75$, and $SEQAD\{A(\omega c_4), A(c_T)\} = 1$.
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.

3.3 Definition of Concept

Concepts are mined from the events through a sequential association rule-base. Semantic concepts are the collection of a temporally ordered set of events. For soccer, concepts can be $Goal_A$, $Goal_B$, $Save_A$ and $Save_B$. It can be expressed as $C_a = \cup_{j=1}^{m} E_j$, where, $E_1, E_2, ..., E_m$ correspond to the events and m is the number of events associated with the a^{th} concept. $C = \{C_a\}_{a=1,2,...,n}$ represent the set of all extracted concepts.

4 Experimental Results

We have tested our proposed approach using of 48 hours of live-recordings of soccer video. We define the threshold vector as $T = [T_1, T_2, T_3, T_4, T_5]$. We experimentally found that $T = [65, 65, 10, 8, 60]$ gives better results. We are analyzing the excitement clips based on the sequence of events in the clip. This extracted event sequence is used for mining semantic concept for the particular excitement clip. We are only interested to know whether the particular event is present or absent in the particular excitement clip. Hence, we present clip-based performance of the hierarchical classifiers. The length of the clips decreases as the level of hierarchy increases, because at each level, we divide the clips into sub-clips. For measuring the performance of classifiers at each level, we use following parameters:

$$Recall = \frac{N_c}{N_c + N_m} \quad and \quad Precision = \frac{N_c}{N_c + N_f}$$

Where, N_c, N_m, N_f represents the number of clips correctly detected, missed and false positive, respectively.

Table 2: Performance of Hierarchical Classifier

Level No.	Class	Total Clips	$N_c/N_m/N_f$	Recall (%)	Precision (%)
1	Replay	244	203/41/26	83.20	88.65
	Real-time	238	215/23/40	90.94	84.31
2	Field-view	196	192/4/11	97.96	94.58
	Non-field-view	247	236/11/4	95.55	98.33
3a	Long-view	38	33/5/4	86.84	89.19
	Straight-view	31	28/3/10	90.32	73.68
	Corner-view	122	110/12/6	90.16	94.83
3b	Close-up	362	326/36/33	90.01	90.81
	Crowd	251	221/33/36	87.01	85.99
4a	Player-A	137	112/25/18	81.75	86.15
	Player-B	123	97/26/24	78.86	80.17
	Goalkeeper-A	26	21/5/7	80.77	75.00
	Goalkeeper-B	29	22/7/9	75.86	70.97
	Referee	11	9/2/2	81.81	81.81
4b	Players Gathering-A	78	61/17/14	78.21	81.33
	Players Gathering-B	61	46/15/17	75.41	73.02
	Spectator	82	66/16/17	80.49	79.52

4.1 Performance of Event-Annotation

The performance of classifiers of hierarchical tree is shown in table 2. We observed 82 % recall and 88 % precision for replays. 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.
We observed 97.96 % recall and 94.58 % precision for field detection. For field view classification, we observed some miss-classification near the shot boundaries, since we consider few previous frame for generating motion-mask. The detection of corner view is very important from semantic concept extraction point of view, since it is frequently observed in goal and save concepts of the soccer video.
Since more number of edges are observed for crowd class, we use edge pixel density as a feature. We observed more than 90 % recall and precision for close-up detection. We have trained our classifier for classifying the close-up frames into five classes. The advantage of our close-up classification scheme is that it gives bett er results even though there is complex background as shown in Figure 11. For close-up classification, we observed the average 79.81 % and 78.82 % recall and precision respectively. For crowd classification, we used similar technique which we used for close-up classification. We observed the average 78.04 % and 77.96 % recall and precision respectively.

4.2 Performance of Concept-Annotation

First, we have computed the association of 180 excitement clips whose concepts are known to us. Then we have selected

Table 3: Association of the extracted video items

Clip No.	Sequence D	Association	D_1	D_2	D_3	D_4	Actual Concept	Observed Concept
1	$P_BV_cP_BG_BSG_BRS$	P_BG_BS	0.89	0	0.75	1	$Goal_B$	$Goal_B$
2	$P_AV_cP_ARP_AR$	P_AR	0.75	1	0.75	0	$Save_B$	$Save_B$
3	$V_cP_ARP_AV_lR$	P_AR	0.75	1	0.75	0	$Save_B$	$Save_B$
4	$V_cP_BV_cRP_BR$	V_cP_BR	1	0.89	0	1	$Save_A$	$Save_A$
5	$V_cP_ARP_AR$	P_AR	0.75	1	0.75	0	$Save_B$	$Save_B$
6	$V_lV_cP_BG_BSG_BR$	G_B	1	0	1	1	$Goal_B$	$Goal_B$
7	$P_AV_cP_AG_ARG_AS$	P_AG_A	0	1	1	0.75	$Goal_A$	$Goal_A$
8	$P_BV_cRP_BR$	P_BR	1	0.75	0	1	$Save_A$	$Save_A$
9	$P_BV_cP_BG_BSG_BSR$	P_BG_BS	0.89	0	0.75	1	$Goal_B$	$Goal_B$
10	$P_AV_cP_AG_BSRS$	P_AS	0	0.75	1	0.75	$Goal_A$	$Goal_A$

Figure 11: Examples of successfully classified close-up with complex background (FIFA world cup 2006 England vs Portugal match): Row-1: (a) Player of team-B (Portugal), (b) Player of team-A (England), (c) Player of team-B (Portugal), (d) Player of team-A (England)

Figure 12: Video Item sequence of clip no. 1 of table-3 (Goal scored by France in FIFA world cup 2006 quarter final match against Brazil): Row-1: (a) P_B, (b) V_c, (c) P_B, (d) G_B, Row-2: (e) S, (f) G_B, (g) R, (h) S

Figure 13: Video Item sequence of clip no. 2 of table-3 (Goal saved by France in FIFA world cup 2006 quarter final match against Brazil): (a) P_A, (b) V_c, (c) P_A, (d) R, (e) P_A, (f) R

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.

Let $A(\omega c_1)$, $A(\omega c_2)$, $A(\omega c_3)$, $A(\omega c_4)$ be the key-association for the concepts $Goal_A$, $Goal_B$, $Save_A$ and $Save_B$ respectively. We found following key associations for our four concept-classes. $A(\omega c_1) = \{P_AG_AS\}$, $A(\omega c_2) = \{P_BG_BS\}$, $A(\omega c_3) = \{P_BR\}$, $A(\omega c_4) = \{P_AR\}$. Let $A(c_T)$ be the association for the concept under test. Let $D_1 = SEQAD\{A(\omega c_1), A(c_T)\}$, $D_2 = SEQAD\{A(\omega c_2), A(c_T)\}$, $D_3 = SEQAD\{A(\omega c_3), A(c_T)\}$, $D_4 = SEQAD\{A(\omega c_4), A(c_T)\}$. The concept-class-$i$ is assigned to the concept under test if $D_i = min(D_1, D_2, D_3, D_4)$ for $i = 1, 2, 3, 4$.

The examples of the association of clips are shown in the table 3. Figure 12 and 13 show the video items (events) for the clip-1 and clip-2 of the table 3 respectively. The overall performance of the semantic concept mining is shown in the table 4.

5 Conclusion

In this paper, we have presented a novel framework for annotation of the excitement clips of the soccer sports video. This framework creates significant links between a set of low-level features extracted directly from sports video and high-level semantic concepts defined in the lexicon. The low-level features used by hierarchical classifier are field color, motion, edge density, skin tone, player's jersey color, etc.

The sports domain semantic knowledge encoded in the hierarchical classification is used to annotate the events of the clip. Sequential association distance is used to annotate semantic concepts for the excitement clip. Each excitement clip will have one concept-lexicon and several event-lexicons. Our results demonstrate that our proposed concept-annotation technique is effective and we could achieve average recall 83.06 % and precision 82.09 %.

For soccer video, we have considered only goal and save concept-lexicons. In future, we will introduce more number of lexicons such as foul, corner kick, free kick, red-card, yellow-card, etc. The proposed framework can be readily generalized to other sports domains as well as other types of video.

References

[1] J. Assfalg, M. Bertini, C. Colombo, A. Bimbo, and W. Nunziati. Semantic annotation of soccer videos:

Table 4: Performance of semantic concept mining

Semantic Concept	Total Clips	N_c	N_m	N_f	Recall (%)	Precision (%)
$Goal_A$	39	32	7	8	82.05	80.00
$Goal_B$	25	21	4	7	84.00	75.00
$Save_A$	73	62	11	8	84.93	88.57
$Save_B$	48	39	9	7	81.25	84.78

automatic highlights identification. *in Elsevier Journal on Computer Vision and Image Understanding*, 2003.

[2] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi. Personalized Abstraction of Broadcasted American Football Video by Highlight Selection. *in IEEE Trans. on Multimedia*, 6(4), 2004.

[3] 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.

[4] 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*, 2(4), 2007.

[5] N. Dimitrova, H. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor. Applications of video-content analysis and retrieval. *in proc. of IEEE Multimedia*, 9(3), 2002.

[6] 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.

[7] L. Duan, M. Xu, Q. Tian, C. Xu, and J. Jin. A unified framework for semantic shot classification in sports video. *in IEEE Trans. on Multimedia*, 7(6), 2005.

[8] 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.

[9] A. Hanjalic. Adaptive extraction of highlights from a sport video based on excitement modeling. *in IEEE Trans. on Multimedia*, 7(6), 2005.

[10] 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*, 23(2), 2006.

[11] M. H. Kolekar and S. Sengupta. Event-importance based customized and automatic cricket highlight generation. *in IEEE Int. Conf. on Multimedia and Expo*, 2006.

[12] B. Li, H. Pan, and I. Sezan. A General Framework for Sports Video Summarization with its application to Soccer. *in IEEE ICASSP*, 2002.

[13] 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*, 3852:633–642, 2006.

[14] R. Ren and J. M. Jose. Football video segmentation based on video production strategy. *in proc. of 27th European Conference on Information Retrieval*, 2005.

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

[16] 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.

[17] C. Theobalt, I. Albrecht, J. Haber, M. Magnor, and H. P. Seidel. Pitching a baseball - tracking high-speed motion with multi-exposure images. *ACM Trans. on Graphics (SIGGRAPH'04)*, 23(3), 2004.

[18] 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.

[19] J. Wang, E. Chng, C. Xu, H. Lu, and Q. Tian. Generation of personalized music sports video using multimodal cues. *in IEEE Trans. on Multimedia*, 9(3), 2007.

[20] C. Xu, J. Wang, H. Lu, and Y. Zhang. A novel framework for semantic annotation and personalized retrieval of sports video. *in IEEE Trans. on Multimedia*, 10(3), 2008.

[21] 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.

[22] L. Zhang, Z. Zhu, and Y. Zhao. Robust commercial detection system. *in proc. of Int. Conf. on Multimedia and Expo*, 2007.

[23] W. Zhou, A. Vellaikal, and C. Kuo. Rule-based video classification system for basketball video indexing. *in Proc. ACM workshop on Multimedia*, 2000.

[24] 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 Trans. on Multimedia*, 9(6), 2007.

[25] 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 Trans. on Knowledge and Data Engg*, 17(5), 2005.