Electronic Letters on Computer Vision and Image Analysis 13(1):33-49, 2014

Automated Classification of Cricket Pitch Frames in Cricket Video

Sandesh Bananki Jayanth and Gowri Srinivasa

Department of Information Science and Engineering
PES Center for Pattern Recognition,
PESIT Bangalore South Campus, Bangalore, India.
Received 5th Dec 2013; accepted 4th Jun 2014

Abstract

The automated detection of the cricket pitch in a video recording of a cricket match is a fundamental step in content-based indexing and summarization of cricket videos. In this paper, we propose visual-content based algorithms to automate the extraction of video frames with the cricket pitch in focus. As a preprocessing step, we first select a subset of frames with a view of the cricket field, of which the cricket pitch forms a part. This filtering process reduces the search space by eliminating frames that contain a view of the audience, close-up shots of specific players, advertisements, etc. The subset of frames containing the cricket field is then subject to statistical modeling of the grayscale (brightness) histogram (SMoG). Since SMoG does not utilize color or domain-specific information such as the region in the frame where the pitch is expected to be located, we propose an alternative algorithm: component quantization based region of interest extraction (CQRE) for the extraction of pitch frames. Experimental results demonstrate that, regardless of the quality of the input, successive application of the two methods outperforms either one applied exclusively. The SMoG-CQRE combination for pitch frame classification yields an average accuracy of 98.6% in the best case (a high resolution video with good contrast) and an average accuracy of 87.9% in the worst case (a low resolution video with poor contrast). Since, the extraction of pitch frames forms the first step in analyzing the important events in a match, we also present a post-processing step, viz., an algorithm to detect players in the extracted pitch frames.

Key Words: sports video analysis, automated indexing, statistical modeling, component quantization

1 Introduction

In the recent times, a large volume of multimedia data is routinely generated and stored by various entertainment, sports and news channels. Some of this multimedia content is made available on the internet for later access by interested users. In order to retrieve those parts of a video that would be of interest to a user, intelligent indexing and summarization of content is a *sine qua non*. This indexing or summarization can be based on any one or more modes of the multimedia content such as image, audio and text data. Once multimedia content is indexed and stored, it facilitates faster retrieval. It has been reported that 2100 hours of video pictures are uploaded every 60 minutes, or 400 hours of video pictures are uploaded to YouTube every day [1]. A number

Correspondence to: <sandesh_bj@pes.edu>

Recommended for acceptance by <Angel Sappa>
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

of general strategies used to process video content for other applications (such as segmentation and tracking) can be applied to content-based video indexing [2]. It has been shown that event-semantics using a semantics-model vector is a powerful tool to detect events in a video [3]. Spatio-temporal features have also been found to be useful to index user generated videos [4].

Among the various genre of multimedia content, sports videos have a significant fan following, with different user interests ranging a wide spectrum from simply viewing the highlights to analyzing the team's (or a specific player's) performance in a game. Thus, a number of sports channels that broadcast different types of sports of live and recorded videos have started to make these recordings (or extracts from these recorded videos) available on the world wide web. It is both computationally challenging and time-inefficient to dynamically retrieve the content of interest to a specific user from these videos. This motivates the design of automated techniques for indexing and summarization, which in turn help in faster retrieval of the portions of videos of interest to a user. In particular, much work has been done towards automated analysis of sports videos, especially those of soccer, baseball and tennis. These techniques revolve around the detection of key events using cinematic features such as shot length and shot type. Such features also help in the detection of events such as a break between plays in sports videos [5]. Annotating videos, such as recording of a soccer game, based on significant events (such as a change in the score) helps in generating the highlights [6] of a game. Finite state machines are used to automatically detect semantic events in soccer videos [7]. On a similar vein, changes in the superimposed caption in a video recording of a baseball game can be used to detect (significant) events in the game. In general, optical character recognition to discern a change in the caption is used to classify an event in a sports video. In particular, events in baseball videos have been successfully annotated by combining the caption text (recognized using optical character recognition) with the detection of the *pitch view* and a *non-active view* [8]. Motion vector analysis in both spatial and temporal domains have also been used in conjunction with neural networks to track baseball events succesfully [9]. In soccer videos, excitement (discerned in the commentator's voice or the audio from the audience) has been associated with significant events on the field. Thus, video clips corresponding to audience excitement have been successfully extracted based on audio features. These features have been mined to generate semantic concepts such as goals, saves, etc. using mining techniques such as the *A priori* Algorithm [10].

The need for techniques to automate the indexing and summarization of cricket videos is patent given that cricket is the most viewed sport in India and in some of the other cricket playing nations like Pakistan and Sri Lanka. The task of automated processing of cricket videos is a challenging one because the same sport is played in different formats such as one day, test and T20. The duration of the game is several hours long and could extend to as many as five days. Further, the variation in lighting conditions is large. The data generated by different sports channels is voluminous and most viewers are interested only in certain portions of the recording. A manual annotation of such large videos to glean only specific details from a match is time consuming and tedious. Moreover, automated indexing paves way for automated analysis and user-customized extraction of information through collating multiple videos and sources of information.

To set the context for the problem addressed in this paper: the game of cricket is played in a large field with eleven players of the fielding team spread out in different directions. Further, to a large extent, the significant events revolve around the cricket pitch (or a stretch of "22 yards" typically seen in shades of brown, at the center of a green field), with a bowler, one or two fieldsmen, two batsmen and an umpire in the field of view. Thus, our attention in this work is on developing algorithms to successfully extract frames with a view of the cricket pitch and filter out those frames that provide extraneous information such as the audience, a close-up shot of a particular player, etc. Since, the extraction of frames with a view of the cricket pitch is used for annotation of key events, an example of localizing key players at the pitch is also presented here. Section 3 presents algorithms to address the primary goal of this paper, viz., is to extract pitch frames from a video recording of a cricket match (regardless of its format). And, Section 5 presents an algorithm to address the secondary goal of this work, viz., to localize key players at the pitch to facilitate further analysis of the frame.

2 Previous Work

In most sporting videos, key frames correspond to those that contain a view of the playing field. Extraction of key frames that have relevant visual information for indexing using color histogram models have been successfully implemented for soccer videos. Frames containing a view of the playing field are detected using dominant color component analysis using the HSV histogram, morphological filtering and connected component analysis. Typically, the field region in soccer videos is extracted using hue values (pixels colored green are extracted and thresholds set using the Smiths Hexagonal Cone model) [11, 12]. Supervised learning methods such as decision trees, random walks and unsupervised methods such as agglomerative clustering have also been found to be useful to discriminate field view frames from frames containing extraneous details [13, 14]. To detect individual players and the ball in these key frames, connected component analysis has been found to be promising [15]. The utility of automatic player localization, labeling and tracking is evident in applications such as player activity analysis and team tactics.

Another sport for which automated techniques have been designed is baseball. Statistical models have been incorporated for the detection of highlights in videos of baseball matches. Color, motion and texture features for each type of highlights (specific events) are calculated and these statistics are used on input videos to extract scene shots that share the same statistics [16].

In the case of cricket videos, there is a large variation in the saturation and brightness levels of specific colors (associated with the field and pitch), duration of focus on the pitch and statistics presented in frames that focus on the field or pitch. These depend on the format of the game, lighting conditions, picture quality, etc. Thus, supervised learning methods should be used with caution as the training data may not be able to capture all the variations and classifiers may not be able to adapt to novel input. While cricket is a very popular sport in countries such as India and has a large viewership, there is not much work on multimedia analytics reported in literature. Mahesh Goyani et. al. have worked on extracting semantic concepts from cricket videos. Since extraction of the frames containing the pitch is also the first step in their work, they have used the mode of the color histogram and the dominant-soil-pixel-ration for extraction [17]. Others have also worked on semantic indexing of cricket videos based on text detection, recognition and localization [18]. Another approach to frame classification has been the use of maximum likelihood estimation [19]. The focus of these foregoing efforts has been on semantic concepts for the extraction of highlights and not on content-analytics. Hence, the algorithm and experiments for the pitch detection phase are not meant to account for variations of color thresholds of the input video.

3 Proposed Method

The method presented in this paper to extract frames with a view of the cricket pitch is carried out in three phases. The first phase comprises a pre-processing step, which acts as a coarse filter. In this step, we extract frames with a view of the field ("field-frames") and discard all non-field frames using RGB color information. The field-frames include both pitch frames and extraneous information such as a view of the field near the boundary. In the second phase of the method, field-frames are further processed to separate frames with a view of the pitch ("pitch frames") from non-pitch frames using the following pitch detection algorithms: SMoG, CQRE and a combination of SMoG and CQRE methods. This second phase forms the crux of the proposed work. Finally, the third phase comprises a post-processing step, which uses the pitch frames to localizes key players in the field of view. A schematic diagram of the various stages in the proposed method is shown in Fig. 1 and each step in the method is explained in detail.

Below are the details of the proposed solution.

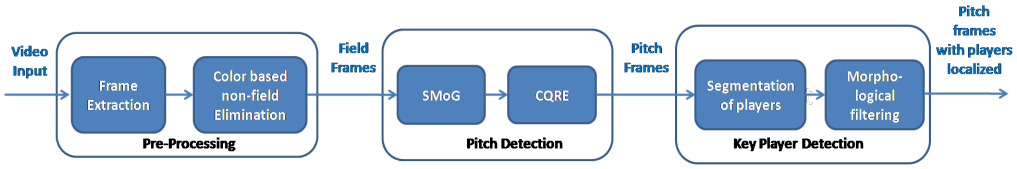


Figure 1: Schematic block diagram of the proposed system showing the three key steps: preprocessing, pitch frame detection and key player localization.

3.1 Preproccessing

In this stage, an input video clip of a cricket match is converted to frames. These frames are processed to eliminate the subset of frames containing non-field information, such as the audience or an advertisement. Thus, this stage acts as a coarse filter and reduces the search space in the subsequent phase. To achieve this coarse filtering, the preprocessing stage harnesses the observation that frames containing the cricket pitch contain mostly green pixels, except for the pitch (represented using different shades of brown), the wickets and markings on the field (typically, white) and players' jerseys (various colors, including shades of green). Whatever be the shade of green for the field, for pixels representing the cricket field, the green component is more dominant than the red and blue components for that pixel. Thus, the total number of pixels in each channel of the RGB image is computed. If green is the dominant component in a frame, that frame is extracted as one of interest. That is, an indicator function $\psi(m,n,t)$ for a pixel (m,n) is set to 1 as follows,

$$\psi(m,n,t) = \begin{cases} 1 & \text{if } 2f_g(m,n,t) - (f_r(m,n,t) + f_b(m,n,t)) \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $m = 0, 1, ..., M-1$, $n = 0, 1, ..., N-1$ and $f_g(\cdot)$, $f_r(\cdot)$ and $f_b(\cdot)$ represent the red, blue and green channels respectively of the tth input frame.

The frame $f(m,n,t)$ is classified as a "field frame" and extracted for further processing only if more than half the pixels in f satisfy the foregoing condition. That is, if

$$\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \psi(m,n,t) > \left\lfloor \frac{MN}{2} \right\rfloor$$

the frame $f(m,n,t)$ is classified as a field frame; all non-field frames are eliminated at this stage.

3.2 Cricket Pitch Detection

The frames classified as field frames by the preprocessing step are further processed to extract "pitch frames" i.e., those frames containing a dominant view of the cricket pitch.

The different algorithms used, in sequence, are as follows.

3.2.1 Statistical modeling of grayscale values (SMoG)

The grayscale version of the RGB frame $g(m,n,t)$ is first computed, followed by computation of its histogram $h(v,t)$, where $v = 0, ..., L$ represents the grayscale value of pixels in f.

It is observed that, regardless of the quality of the video and lighting conditions, histograms of cricket pitch frames have two dominant peaks, one at the grayscale value corresponding to pixels whose green component is dominant, typically pixels representing the field and another at the grayscale value corresponding to the pixels whose red component is dominant, typically brown for the pitch. Thus, histograms of cricket pitch frames are bimodal (Fig. 2(a) shows a sample pitch frame and Fig. 2(c) a non-pitch frame with their grayscale histograms

in Fig. 2(b) and Fig. 2(d) respectively). A histogram of a grayscale image represents the number of pixels at each gray level. Since the number of significant peaks in the histogram distinguishes a pitch frame from a non-pitch frame, this is evaluated. The histogram may be smoothed (to avoid detecting spurious changes as significant peaks) prior to evaluating the peaks. The number of peaks are computed by first applying a threshold τ_0 to the histogram; the threshold is computed as follows,

$$\tau_0 = \omega * M * N, \tag{1}$$

where M, N are the dimensions of the image and ω, a weight that is set at 0.04% of the total number of pixels in the frame. The rationale behind the weight computation is the following. It is seen that the mode of the histogram is usually between 4%-5% of the number of pixels in the frame. If the histogram is bimodal (as we expect in a pitch frame: the most significant peak corresponding to the green field and the second most significant peak corresponding to the pitch region), then about 10-15% of the total pixels in the most significant mode are concentrated at the second most significant peak. Thus 10% of 4% forms a reasonable threshold to assess whether the grayscale histogram is bimodal. Typically, non-pitch images do not conform to this statistical model of pitch frames. Experimental results verify that this statistical modeling of the grayscale histogram performs reliably across varying resolutions and lighting conditions. Applying threshold τ_0 to the histogram results in an indicator function φ signifying the grayscale values at which the histogram exceeds the threshold,

$$\varphi(v,t) = \begin{cases} 1 & \text{if } h(v,t) \geq \tau_0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The number of zero crossings Z for the tth frame is then computed as follows,

$$Z(t) = \sum_{v=0}^{L-1} |\text{sgn}(\varphi(v,t))|, \tag{3}$$

where L is the number of graylevels in the histogram and sgn is the signum function.

Once the number of zero crossings is computed, SMoG classification of an input frame $f(m,n,t)$ is performed based on the following rule,

$$FrameType_t = \begin{cases} 1 & \text{if } \alpha \leq Z(t) \ \leq \beta, \\ 0 & \text{otherwise,} \end{cases}$$

where 1 indicates a pitch frame and a 0 indicates a non-pitch frame and $\alpha = 4$ and $\beta = 6$ for the upper and lower bounds on the number of zero crossings for a pitch frame. The values α and β are so chosen because if the histogram is bimodal, we can expect 4 zero crossings. There are some field-frames with a view of the pitch and players that account for three significant peaks in the thresholded histogram (one peak each corresponding to the field, the pitch and players at the pitch or other such non-stationarities respectively). Histograms that are unimodal, relatively flat or have more than three significant peaks are typically non-pitch frames. The steps of this algorithm are described in Algorithm 1.

Typical pitch and non-pitch frames with their corresponding histograms are shown in Fig. 2. Setting $\alpha < 4$ would result in classifying all frames in the video as pitch frames. Likewise, setting $\beta > 6$ would run the risk of classifying a significantly larger number of non-pitch frames as pitch frames.

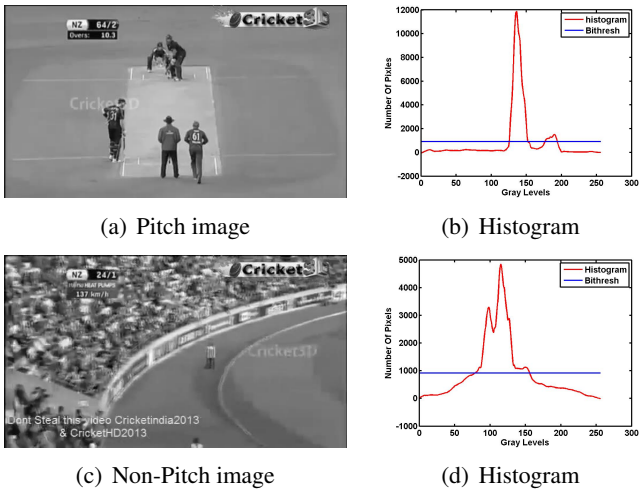
All frames with a FrameType of 1 (i.e., frames not eliminated by the SMoG algorithm) are assumed to be pitch frames. An alternative approach that is sensitive to the context (domain information) is presented in the following section.

Algorithm 1: SMoG classification of pitch frames in a cricket video

Input: $g(m,n,t)$:grayscale image corresonding to the t^{th} frame $f(m,n,t)$ of a video of T frames, L: number of gray levels of $f(m,n,t)$, M, N: height and width of the input frame, ω: a weight to compute the threshold, α: lower threshold, β: upper threshold for the number of peaks

Output: $FrameType_t$: a classification label for the tth frame input, viz., $f(m,n,t)$

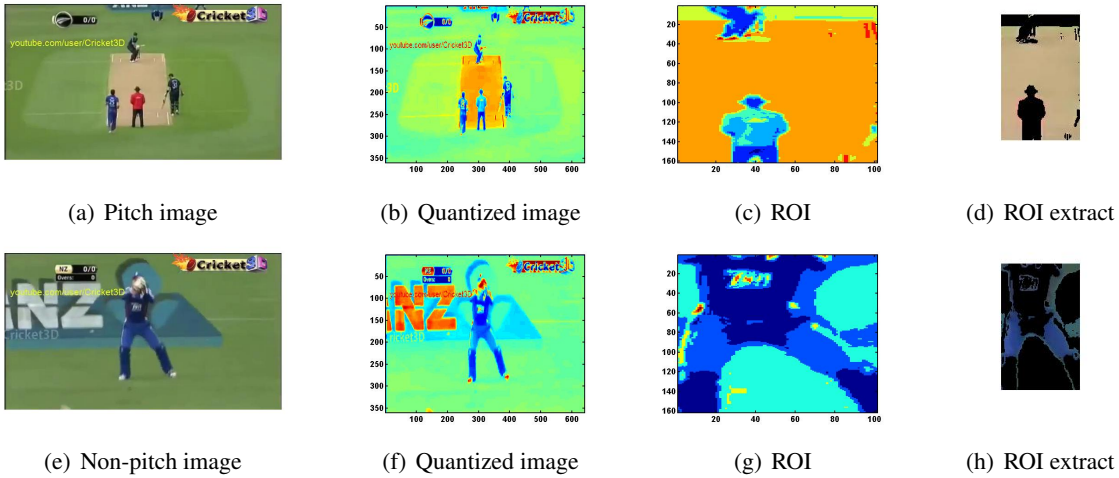
$\forall$pixels in $g(m,n,t)$, compute histogram $h(v,t)$ = $h(v,t)$+1, if $g(m,n,t) = v$;
Compute threshold $\tau_0 \leftarrow (\omega * M * N)$ as per (1);
$\forall$gray values v in $h(v,t)$, update characteristic function φ as per (2);
Compute zerocrossings $Z(t)$ as per (3);
if $\alpha \leq Z(t) \leq \beta$ **then** ;
$FrameType_t \leftarrow 1$;
else $FrameType_t \leftarrow 0$; ;
Return $FrameType_t$;



(a) Pitch image (b) Histogram

(c) Non-Pitch image (d) Histogram

Figure 2: Histogram modeling of cricket video frames (a) a typical pitch frame in a cricket video (shown in grayscale), (b) grayscale histogram of the pitch frame showing two significant peaks (above the threshold τ_0) corresponding to the field and the pitch respectively (c) a typical non-pitch frame in the same cricket video (shown in grayscale) and (d) grayscale histogram of the non-pitch frame, showing it does not conform to the statistical model of the histograms of pitch-frames.



(a) Pitch image (b) Quantized image (c) ROI (d) ROI extract

(e) Non-pitch image (f) Quantized image (g) ROI (h) ROI extract

Figure 3: (a) a pitch frame in the input cricket video (b) quantized image (8 levels) (c) region of interest (ROI) (d) largest connected component in the ROI (e) a non-pitch frame in an input video (f) quantized image (8 levels) (g) region of interest (h) largest connected component in the ROI. Whereas CQRE classifies (a) as a pitch frame (because shades of red are significant in (d), the largest connected componen in the cropped region of the input frame), CQRE elimates (e) as a non-pitch frame (since shades of red are not significant in (h), the largest connected component in the cropped region of this frame).

3.2.2 Component Quantization based Region Extraction (CQRE)

The output of the preprocessing stage results in a set of frames with a view of the field. This includes both pitch and non-pitch frames. The Component Quantization based Region Extraction (CQRE) method uses two domain-specific inputs. The first is that pitch frames are those captured by a camera, which contains pitch information with a view of the batsman, the bowler, an umpire and a subset of players. Such a pitch frame has the pitch (or "22 yards") at its center of focus. Thus, a frame focusing on a fieldsman near the boundary, with a peripheral view of the pitch is expected to be discarded. The second domain-specific input is that the region of the pitch component of the field appears homogeneous in the frame, but for the discontinuities due to silhouettes of the stumps, players, etc.

Thus, to separate pitch frames from those that do not contain a view of the pitch at the center of focus, each preprocessed frame is quantized as follows,

$$\dot{g}(m,n,t) = ((g(m,n,t) - g_{min})/(g_{max} - g_{min})) * (\dot{L} - 1), \tag{4}$$

where g_{min} and g_{max} are the minimum and maximum values in the grayscale frame $g(m,n,t)$ and $\dot{L}$ is the number of quantization levels. (For instance, see Fig. 3(b) and Fig. 3(f) for example images quantized to $\dot{L} = 8$ levels).

The idea is to partition the frame into contiguous regions that are fairly homogeneous. Empirically, $\dot{L} = 8$ levels is found to yield good results. The rationale behind this is that the original $L = 256$ levels contains too far too many details, whereas a number such as 2 or 4 would be too coarse a partitioning and would risk including a significant portion of non-pitch regions into the same partition as the pitch.

A subimage $q(m',n',t)$ corresponding to the region of interest (based on the camera angle, with the pitch at the center of focus) is cropped from $\dot{g}(m,n,t)$ as follows,

$$q(m',n',t) = \dot{g}\left(\frac{M}{2} + p, \frac{N}{2} + q, t\right), \tag{5}$$

where $p = -\left\lfloor \frac{M_2-M_1}{2} \right\rfloor, -\left\lfloor \frac{M_2-M_1}{2} \right\rfloor+1, ..., 0, 1, 2, ..., \left\lfloor \frac{M_2-M_1}{2} \right\rfloor, q = -\left\lfloor \frac{N_2-N_1}{2} \right\rfloor, -\left\lfloor \frac{N_2-N_1}{2} \right\rfloor+1, ..., \left\lfloor \frac{N_2-N_1}{2} \right\rfloor$ and (M_1, N_1) and (M_2, N_2) are the bounds of the region to be cropped within the quantized grayscale image $\dot{g}(\cdot)$. For cricket videos, pitch-frames usually contain the pitch in the central region of the frame. For the purpose of this application, we focus only on those frames where the pitch is in the field of focus of the camera. See for instance Fig. 3(c) and Fig. 3(g) for cropped regions of a pitch and non-pitch frame respectively.

Subsequent to the cropping, a connected component analysis is performed on each of the quantized levels. First, a characteristic function corresponding to each level $\dot{l} = 1, ..., \dot{L}$ is computed from $q(\cdot)$ as follows,

$$\varphi_{\dot{l}}(m', n', t) = \begin{cases} 1 & \text{if } q(m', n', t) = \dot{l}, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $m' = 0, 1, 2, .., (M_2 - M_1)$ and $n' = 0, 1, 2, .., (N_2 - N_1)$.

Next, the number of pixels in each connected component of the characteristic function $\varphi_{\dot{l}}$ is computed as follows,

$$N_{\dot{l},\gamma}(t) = \sum_{m'=0}^{M_2-M_1} \sum_{n'=0}^{N_2-N_1} \text{ConnectedComponent}_{\gamma}(\varphi_{\dot{l}}(m, n, t)), \tag{7}$$

where $\text{ConnectedComponent}_{\gamma}(\cdot)$ is a function that extracts the γth connected component of φ at level $\dot{l}$. The largest connected component $R_{\dot{l}}(t)$ at a quantization level $\dot{l}$ is computed as follows,

$$R_{\dot{l}}(t) = \arg\max_{\gamma}(N_{\dot{l},\gamma}(t)),$$

and the quantization level with the largest connected component at any level is computed using,

$$ROI(t) = \arg\max_{\dot{l}} R_{\dot{l}}(t).$$

It is quite possible that the centre of frame (the region in which we expect to find the cricket pitch) is a homogeneous region (indicated by the largest connected component at one of the quantization levels), but may be, for instance, a portion of the green field. To ensure the region extracted indeed corresponds to the pitch, we check whether the values in the red channel are more dominant than the corresponding blue and green channels for the cropped region in that frame. First, the relevant portion of the frame is cropped similar to (5)

$$\hat{f}(m', n', t) = f\left(\frac{M}{2} + p, \frac{N}{2} + q, t\right), \tag{8}$$

where $p = -\left\lfloor \frac{M_2-M_1}{2} \right\rfloor, -\left\lfloor \frac{M_2-M_1}{2} \right\rfloor+1, ..., 0, 1, 2, ..., \left\lfloor \frac{M_2-M_1}{2} \right\rfloor, q = -\left\lfloor \frac{N_2-N_1}{2} \right\rfloor, -\left\lfloor \frac{N_2-N_1}{2} \right\rfloor+1, ..., \left\lfloor \frac{N_2-N_1}{2} \right\rfloor$ and (M_1, N_1) and (M_2, N_2) are the bounds of the region to be cropped within the frame $f(m, n, t)$. Thereafter, for the color $c \in \{r, g, b\}$, for red, blue and green channels respectively, the total N_c of values in the largest connected component in that channel is computed as follows,

$$N_c = \sum_{m'=0}^{M_2-M_1} \sum_{n'=0}^{N_2-N_1} \varphi_{ROI(t)} \cdot * \hat{f}_c(m', n', t), \tag{9}$$

where $\hat{f}_c(m', n', t)$ is the cth color component of the subimage $\hat{f}(m', n', t)$, the operation $\cdot *$ extracts the pixels in $\hat{f}(m', n', t)$ corresponding to the largest connected component (in the $ROI(t)$th quantization level).

If a majority of the pixels in the region extracted have some shade of red (corresponding to different shades of brown for the pitch) as the significant color component, the frame is classified as a pitch frame by CQRE. If not, the frame is discarded.

As before, the $FrameType$ for the tth frame is assigned a value 1 for the pitch frame as follows,

$$FrameType_t = \begin{cases} 1 & \text{if } 2N_r - (N_b + N_g) \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Owing to the two tests (of region homogeneity and the significant color in the region expected to be the pitch), the CQRE algorithm is more aggressive in ensuring the frames selected are indeed pitch frames. The steps of the CQRE algorithm are described in Algorithm 2.

Algorithm 2: CQRE classification of pitch frames in a cricket video

Input: $f(m,n,t)$:the t^{th} frame of a video of T frames, $g(m,n,t)$, the grayscale frame corresponding to $f(m,n,t)$, $\dot{L}$: the number of levels of quantization, $(M_1,N_1),(M_2,N_2)$: coordinates of the subimage (denoting the pitch region) to be cropped

Output: $FrameType_t$: a classification label for the tth frame, $f(m,n,t)$

for $m \leftarrow 0$ ***to*** $M-1$ **do**
  for $n \leftarrow 0$ ***to*** $N-1$ **do**
    $\dot{g}(m,n,t) \leftarrow ((g(m,n,t) - g_{min})/(g_{max} - g_{min})) * (\dot{L}-1)$
$q(m',n',t) \leftarrow$subimg$(\dot{g}(m,n,t))$ as per (5);
$\forall$pixels in $q(\cdot)$, update characteristic function $\varphi(t)$as per (6)
$\forall$ levels $\dot{l}$ compute $N_{\dot{l},\gamma}$ the number of pixels in each connected component γ as per (7)
for $\dot{l} \leftarrow 0$ ***to*** $(\dot{L}-1)$ **do**
  $R_{\dot{l}}(t) \leftarrow \arg\max_{\gamma}(N_{\dot{l},\gamma})$;
$ROI(t) \leftarrow \arg\max_{\dot{l}}(R_{\dot{l}}(t))$;
$\hat{f}(m'n',t) \leftarrow$ subimg$(f(m,n,t))$ as per (8);
$N_r \leftarrow \sum_{m'=0}^{M_2-M_1}\sum_{n=0}^{N_2-N_1} \varphi_{ROI(t)} \cdot *\hat{f}_r(m,n,t)$;
$N_g \leftarrow \sum_{m=0}^{M_2-M_1}\sum_{n=0}^{N_2-N_1} \varphi_{ROI(t)} \cdot *\hat{f}_g(m,n,t)$;
$N_b \leftarrow \sum_{m=0}^{M_2-M_1}\sum_{n=0}^{N_2-N_1} \varphi_{ROI(t)} \cdot *\hat{f}_b(m,n,t)$;
$\delta \leftarrow 2N_r - (N_b + N_g)$;
if $\delta > 0$ **then** ;
$FrameType_t \leftarrow 1$;
else $FrameType_t \leftarrow 0$;;
Return $FrameType_t$;

3.2.3 Pitch frame classification using SMoG-CQRE

While SMoG and CQRE are both found to be reliable pitch frame classifiers, they use complimentary information. There are cases where each method in exclusion might be expected to misclassify a frame. For instance, if the image is a closeup shot of a player wearing a green jersey, due to the green dominance, the preprocessing stage admits the frame as a frame with a view of the field. Two dominant colors in the grayscale histogram could be green (player's jersey) and one corresponding to his skin tone (face and hands). Likewise, if it is part of a field frame with a solitary player in focus, then the statistical model of the histogram can be expected to match the model of a pitch frame (due to the dominance of green from the field and the jersey color of the player) as shown in Fig. 4(a). Such a frame is misclassified by SMoG. However, it will not pass through the CQRE algorithm, as the tests for red-dominance and connected components in a specified region are likely to fail. In fact, Fig. 4(a) discussed above is the same as Fig. 3(e) that is correctly classified as a non-pitch frame by CQRE.

On a similar note, CQRE alone might result in misclassifying an advertisement or audience frame as a pitch frame, if the red-dominance and connected component tests meet the criteria for pitch frames. In such cases, SMoG eliminates these spurious frames with ease based on the green-dominance and statistical modelling. We observe that while non-field frames are successfully eliminated by SMoG, it has a tendency to select non-pitch frames. Given a set of field frames, CQRE is the more aggressive non-pitch frame elimination algorithm. Thus, after preprocessing the frames, applying SMoG narrows down the frames to a plausible subset. Subsequent application of CQRE ensures further false positives are eliminated. Thus, the SMoG-CQRE method comprises application of Algorithm 1 followed by Algorithm 2 in succession. As discussed in Section 3.2.2, SMoG eliminates false negatives to a large extent and, CQRE eliminates most of the false positives (see Fig. 3(e)).

4 Experimental Results

In order to experimentally validate the efficacy of SMoG and CQRE to classify pitch frames automatically for further processing, we applied these algorithms on cricket video clips available online.

4.1 Data Set

Our data comprises four cricket videos from different broadcasting channels differing in their resolution, placement of text such as the logo and scores, contrast, aspect ratios, different teams (hence different jersey colors), stadia and lighting conditions. Details of the data used in the experiments are summarized in Table 1.

Table 1: Details of the dataset

video	#frames	Resolution
video-1	6819	640X360
video-2	3925	480X360
video-3	5878	480X360
video-4	3255	640X360

4.2 Parameter Settings

The input video clips are converted to frames and the SMoG, CQRE and the combination of SMoG-CQRE are applied to the frames of the input video to automatically classify them as 'pitch' or 'non-pitch' frames. Most of the parameters used in the algorithms have been detailed in the foregoing sections. In particular, α and β are the the thresholds set on the number of zero crossings for the SMoG method. As explained in Sec. 3.2.1, they are set to ensure the bimodal behavior of the grayscale histogram, while accommodating non-stationarities that may result in a graph that is trimodal. The parameter τ_0 is a threshold on the number of pixels and derived from the size of the video frame. The only parameter that needs to be set in CQRE is the field of view, to determine whether the pitch is the centre of focus in a frame. Presently, this is derived from the dimensions of the input frame as a percentage of its height (40%) and width (25%) respectively. The region of interest is cropped about the centre with provision to account for the camera zooming into the pitch. Both SMoG and CQRE and the combination SMoG-CQRE do not require any manual tuning of parameters. The resulting recall, precision and a measure of average accuracy are tabulated in Table 2.

A discussion of the results follows in the subsequent subsections.

(a) (b) (c) (d)

Figure 4: Examples of false-positives: non-pitch frames misclassified as a pitch frames by SMoG: The two significant peaks detected in the grayscale histgoram correspond to (a) and (b) green of the playing field and the colored patch for the advertisement, (c) for two different shades of green corresponding to the playing field and to the player and (d) to the two different shades of green corresponding to the playing field and the green patch at the bottom left corner and towards the top of the frame, next to the partially visible pitch.

4.3 Visual assessment of performance

A representative result of SMoG classification of pitch and non-pitch frames is shown in Fig. 2(a) and Fig. 2(c). As can be observed from representative plots of the histogram of a pitch and a non-pitch frame, the statistical model of the grayscale histogram for each class is consistent with the assumptions. The threshold, shown in blue in Fig. 2(b) shows the number of zero crossings to be 4, as expected in a pitch frame. A non-pitch frame usually does not have two significant peaks unlike the pitch frame (see Fig. 2(d)).

There are some pathological cases, which necessitate the application of CQRE, the second classifier. For instance, in Fig. 4(b), we see two significant shades of gray, despite the frame not having the cricket pitch at the center for focus. Likewise, in Fig. 4(d) is an example of a pitch frame with more non-stationarities, leading to a larger number of zero-crossings than permitted by the classifier. The latter case of valid pitch frames being misclassified as non-pitch frames are few and far between and can be corrected using the post-processing procedure. A significant number of false positives are successfully eliminated by the CQRE classifier.

Fig. 3(e) provides a representative example of the efficacy of CQRE in eliminating false positives resulting from SMoG. The method looks for a significant connected component that satisfies the homogeneity criterion and localization for the pitch in a pitch frame. A frame that may have been misclassified (due to spurious statistics), which contains significant green and blue components and misclassified by SMoG, can be successfully eliminated using component quantization as seen in Fig. 3(f) and analysis of the connected components in the region of interest as seen in Fig. 3(h). A pitch frame, on the other hand, is classified correctly (for instance, see Fig. 3(a)).

4.4 Quantitative performance measures

Table 2 summarizes the results for the various frames classified in the four video clips used in the experiments. Recall (R) and precision (P), standard measures of classifier performance are calculated as follows -

$$R = \frac{A \cap G}{G},\ P = \frac{A \cap G}{A},$$

where A is the class label "pitch frame" for an input frame f_t as output by the algorithm (SMoG, CQRE or SMoG-CQRE) and G the label, "pitch frame" as determined by the ground truth (manual annotation). R is a measure of the number of pitch frames that were correctly retrieved, whereas P is a measure of the fraction of frames retrieved that are labelled correctly. R alone might be 100% if an algorithm labels every input frame as a "pitch frame". Thus, P penalizes a method that yields too many false positives. Only an algorithm that matches exactly with the ground truth can result in 100% performance for both P and R.

Based on the design criteria explained in Section 3.2.1, we expect to see a higher recall rate (compared to the precision) for SMoG for all video clips used in the experiments. Indeed, Table 2 verifies this assumption. The rationale is that CQRE is designed to eliminate false positives resulting from SMoG, whereas it cannot recover a false-negative. In the same vein, we expect the precision values of CQRE to be higher, since CQRE is more aggressive in eliminating non-pitch frames. Consequently, the fraction of "pitch frames" output by CQRE that matches with the ground truth data is higher. The average accuracy is measured as the number of true positives and true negatives detected by the algorithms for all frames of an input video clip. This performance measure across all videos is consistently higher for SMoG-CQRE than the application of either algorithm applied in exclusion of the other. Further, the lowest accuracy of nearly 88% is noted for a video clip that is of low resolution and very poor lighting conditions and contrast. The highest accuracy of 98.6% is noted in Video-3 that is a HD video of a cricket match, with good lighting conditions and high contrast.

Table 2: Precision, recall and accuracy of classification of cricket pitch frames.

video	#frames	method	precision(%)	recall(%)	accuracy(%)
video-1	4893	SMoG	48.8	91.5	69.0
		CQRE	86.9	68.4	87.8
		SMoG-CQRE	98.2	73.4	**91.3**
video-2	1350	SMoG	21.6	95.8	62.7
		CQRE	27.6	92.3	73.6
		SMoG-CQRE	72.6	88.7	**91.1**
video-3	2463	SMoG	31.3	98.7	85.5
		CQRE	55.8	62.7	94.2
		SMoG-CQRE	90.5	91.0	**98.6**
video-4	1921	SMoG	52.3	90.0	71.5
		CQRE	70.8	78.52	87.2
		SMoG-CQRE	97.3	64.6	**87.9**

4.5 Post-processing to recover false-negatives

Despite the fact SMoG has a tendency to retain almost all the field-frames, a small number of pitch frames may be eliminated. Further, CQRE being the more aggressive algorithm may eliminate some pitch frames if the pitch is captured at a camera angle different from the one specified. Thus, a post-processing method that can be incorporated in the pitch frame detection phase after the application of SMoG-CQRE would help recover some of these false negatives. This procedure stems from the observation that a video sequence of valid pitch frames is never eliminated in its entirity. Suppose the frame rate for a video is f_s (typically, about 30 frames per second), then based on the persistence of vision, no valid change can be expected to be recorded in $f_s/3$ or fewer frames. Thus, for a frame $f(m,n,t)$ suspected to be a false negative, we study the class labels of all frames in the interval $f_t \pm f_s/3$. A weighted voting procedure (with the frame $f(m,n,t\pm1)$ having the weight 1, $f(m,n,t\pm2)$ having the weight $\frac{1}{2}$, and in general, frame $f(m,n,t\pm n)$ having the weight $\frac{1}{n}$, with a value $FrameType_t = +1$ for a positive class (pitch frame) and value of $FrameType_t = -1$ assigned to a negative class (non-pitch frame) used to decide whether the label for the frame under consideration ought to be changed. Thus a classification rule to recover false-negatives is as follows,

$$FrameType_t = \begin{cases} 1 & \text{if } \sum_{n=t-f_s/3}^{t+f_s/3} \frac{1}{n} FrameType_n \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Cricket Pitch Frame (b) Binary Image

Figure 5: Key player detection (a) a sample cricket pitch frame (b) the binarized image of the same frame showing connected components corresponding to the text and players in this frame.

This rule is run only on frames that are labeled non-pitch frames. No action is taken on those that have already been labeled as pitch frames. While the classification accuracy is higher with this post-processing method, we do not see a significant impact of these discarded frames in the further stages of our processing viz., key-player detection and towards automated indexing of significant events. Thus, although the post-processing procedure is reported for the sake of completion, its use with SMoG, CQRE and SMoG-CQRE procedures is optional.

The main purpose of automated classification of cricket pitch frames is to segment and mark key players in the pitch. Key player detection can be further used to detect the important events in a game. These event clips can be used to index a video for user-adaptive and faster retrieval. Thus, the detection of key players in the pitch frames extracted by SMoG-CQRE are discussed in the following section.

5 Key Player Detection

Given that player detection in a sporting video has ramifications in the automated indexing or retrieval of key events or player action recognition for generation of highlights, much work has been done towards player detection for sports such as soccer, baseball, tennis and basketball. In particular, facial and textual cues (number on the jersey) have been used not only to detect but even identify individual players [20]. Multiplayer detection and tracking has been used to support efficient querying. In this case, position and trajectory of the players help in semantic analysis of the game. The detection and tracking of multiple players is achieved using supervised methods such as a support vector machine (SVM) classifier in the play area of the field. Typically, features of the HSV color model of a frame are fed into an SVM [21]. Other classifiers used to recognize players and detect events (such as a goal) include probabilistic Bayesian Belief Network [22].

In tennis videos, player action recognition has been successfully done by studying the features in a small window around a detected player. This has helped in semantic and tactic analysis of the game [23]. In the case of basketball videos, deformable part models have been proposed to automatically locate and track a player [24].

In the present work, our focus is on detecting (and not recognizing) players in the pitch frame. The grayscale histogram of the pitch frame used in the SMoG or SMoG-CQRE method is used for this purpose. As discussed in Section 3.2.1, we expect the histogram of a pitch frame to approximate a bimodal distribution. Assuming the distribution is a mixture of two gaussians, curve-fitting is used first to identify the two significant peaks, which are estimated as the two means, μ_1 and μ_2 of the Gaussian mixture model. Using these peak values the respective variances σ_1^2 and σ_2^2 are estimated and estimated as the upper and lower threshold are calculated as follows,

$$\tau_1 = \mu_1 - \sigma_1^2 \text{ and } \tau_2 = \mu_2 + \sigma_2^2,$$

where $\mu_1 < \mu_2$ (that is, the first peak lies to the left of the second).

The value τ_1 and τ_2 are used to threshold the grayscale version $g(m,n,t)$ of the tth frame to extract a characteristic function $\phi(t)$ of the frame, which contains the segmented players. The characteristic function (or binary image with the segmented players) is computed akin to an extreme value distribution as follows,

$$\phi(m,n,t) = \begin{cases} 1 & \text{if } \tau_2 \leq g(m,n,t) \leq \tau_1, \\ 0 & \text{otherwise.} \end{cases}$$

The majority of pixels, i.e., those with grayscale levels between τ_1 and τ_2 account for the grayscale equivalents of the field and pitch, the two most prominent regions in any pitch frame. Thus, a compliment of these regions yields the players and any textual information in the frame. The binary image corresponding to the original frame in Fig. 5(a) after thresholding is shown Fig. 5(b). The algorithm for computing the peaks and thereafter obtaining a binary image of players in a pitch frame is detailed in Algorithm 3.

Algorithm 3: PeakFinder Method applied on BiModal Histogram to segment the key Players in the Pitch frames Of Cricket Video

Input: $g(m,n,t)$:grayscale equivalent of the t^{th} frame of a video of T frames, L: number of gray levels, M: the number of rows and N: the number of columns, τ_1: lower threshold, τ_2: upper threshold, μ_1: mean value of the first peak, μ_2: mean value of the second peak, σ_1^2: variance of the first peak, σ_2^2: variance of the second peak,

Output: $\phi(t)$: a characteristic function of pixels representing players at the pitch in the tth frame

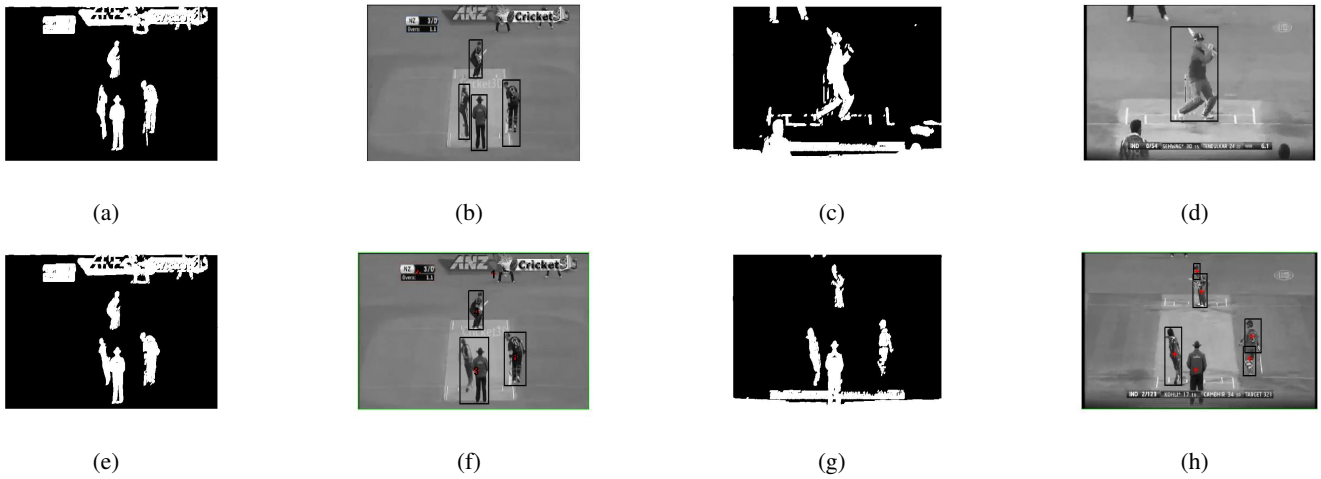
```
τ1 ← μ1 − σ1^2;
τ1 ← μ2 + σ2^2;
∀ the pixels in g(m,n,t) update characteristic function φ(m,n,t)
for m ← 0 to M − 1 do
    for n ← 0 to N − 1 do
        if τ1 ≤ g(m,n,t) ≤ τ2 then ;
        φ(m,n,t) ← 1;
        else ;
        φ(m,n,t) ← 0;
Return φ(t);
```

Once a binary image with the objects of interest is obtained, isolated pixels in the binary image $\phi(t)$ are removed using the morphological operation *clean*. Further, using vertical and diagonal structuring elements, the image is dilated to fill holes inside objects. Subsequently, small objects spuriously connecting two connected components (such as two players or the player and umpire) are removed using the area-open operation. This operation is designed to eliminate any connected component that is fewer than P pixels in area (P is about 0.1% of the size of the image). For each object detected, the centroid is computed and a bounding box is drawn around each object in the binary image φ for subsequent processing such as tracking or action analysis.

5.1 Preliminary Results

Data used for the experiments are those labelled "pitch frame" by the SMoG-CQRE algorithm. There are not many parameters to be set in this procedure and, as explained in the foregoing sections, the parameters that need to be set are done so automatically: parameters μ_1,μ_2 represents the two values about which the distribution peaks in the grayscale histogram. σ_1^2 and σ_2^2 are the variance values about the two significant peaks. These statistics are used to calculate the lower and upper thresholds, τ_1 and τ_2 to obtain characteristic functions with the field and pitch set to 0 and players (and captions) set to 1.

Sample frames for player detection in the pitch frames are provided in Fig. 6. As can be noted, Fig 6(a) is a perfect binary image with no overlaps between binary objects corresponding to the players (or umpire)



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 6: Qualitative assessment automated key player detection (a-b) binarized pitch frame and bounding box around the players correctly marked (c-d) binarized pitch frame and bounding box around the players for a different configuration of players (e-f) binarized pitch frame and bounding box of two players merged due to the proximity of players (g-h) binarized pitch frame and a split in the bounding box around a player due to shadow effects.

in the pitch. Thus, each player in the field of view is detected accurately. Fig 6(c) shows another image that yields nearly perfect detection of the players in the field. For the sake of completion, we also report the results where this segmentation could yield faulty outcomes leading to a merge, split or both. In the case of Fig. 6(e), the bowler is too close to the umpire to be separated using the area open operation. Thus, there is a merge, resulting in a common bounding box for the umpire and bowler. Finally, Fig. 6(g) demonstrates an anomaly in the topology (after morphological operations) of the batsman at the non-striker's end that gives the appearance of two separate objects. Thus, two bounding boxes delineate the same player resulting in a split.

In the context of this work that focuses on detection of pitch frames, it is noteworthy that there is a potential for improving the detection of key players to account for shadow effects and merges due to proximity of players. However, we defer the discussion on improvements to a future work and note that the SMoG-CQRE procedure ensures pitch frames are satisfactorily classified and does not eliminate any of the significant action sequences.

6 Conclusion

In this paper we proposed techniques for automating the classification of cricket pitch frames. The statstical modeling of the grayscale histogram (SMoG) is found to be a lenient classifier that has a high recall. Of the subset of images classified by SmOG, the second algorithm, component quantization based region extraction(CQRE), eliminates most of the false-positives, ensuring a high precision. Experiments conducted on four videos differing in various aspects of content and quality demonstrated the efficacy of the algorithms and applying them in succession. We have also demonstrated how the pitch frames extracted by the SMoG-CQRE algorithm can be used for further processing such as segmenting key players at the pitch. While the results of the player detection algorithm look promising, these results can be improved by accounting for shadow effects and topological anomalies of the players detected in the pitch frames.

Acknowledgment

The authors would like to acknowledge the sources of cricket video clips used in this paper: Cricket3D, ESPNStar, CrkTorrents and thank the users who have generously shared their home-recorded video clips and made them available for download on YouTube.

References

[1] Scott. YouTube official Blog. Last accessed on 2013-12-05 21:37. [Online]. Available: http://youtube-global.blogspot.in/2010/11/great-scott-over-35-hours-of-video.html

[2] W. Hu, N. Xie, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 41, pp. 797–819, Jul. 2011.

[3] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Trans. on Multimedia*, vol. 14, pp. 88–101, Feb. 2012.

[4] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Elli, M. Shah, and S.-F. Chang, "Combining multiple modalities, contextual concepts, and temporal matching," *Intl. Conf. on Columbia-UCF TRECVID2010 Multimedia Event Detection*, 2010.

[5] A. Ekin and A. M. Tekalp, "Generic play-break event detection for summarization and hierarchical sports video analysis," *Proc. of Intl. Conf. on Multimdeia and Expo (ICME)*, 2003.

[6] X. Qian, G. Liu, H. Wang, Z. Li, and Z. Wang, "Soccer video event detection by fusing middle level visual semantics of an event clip," *Proc. Intl. Conf. on Advances in Multimedia Information*, pp. 439–451, 2010.

[7] L. Bai, S. Lao, W. Zhang, G. J. F. .Jones, and A. F. Smeaton, "A semantic event detection approach for soccer video based on perception concepts and finite state machines," *Proc. Intl. Workshop on Image and Audio Analysis for Multimedia Interactive Services*, 2007.

[8] D. Zhang and S.-F. Chang, "Event detection in baseball video using superimposed caption recognition," *Proc. Tenth ACM Intl. Conf. on Multimedia*, pp. 315–318, 2002.

[9] W.-N. Lie, T.-C. Lin, and S.-H. Hsiai, "Motion-based event detection and semantic classification for baseball sport videos," *Proc. Intl. Conf. on Multimedia and Expo (ICME)*, 2004.

[10] M. H. Kolekar and K. Palaniappan, "Semantic concept mining based on hierarchical event detection for soccer video indexing," *Journal Of Mulitmeda*, vol. 4, pp. 298–312, Sep. 2009.

[11] O. Utsumi, K. Miura, I. Ide, S. Sakai, and H. Tanaka, "An object detection method for describing soccer games from video," *Proc. IEEE Conf. on Multimedia and Expo (ICME)*, pp. 45–48, Aug. 2002.

[12] M. Y. Eldib, B. S. A. Zaid, H. M. Zawbaa, M. El-Zahar, and M. El-Saban, "Soccer video summarization using enhanced logo detection," *Proc. Intl. Conf. on Image Processing (ICIP)*, p. 42894292, 2009.

[13] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, H. Wang, B. Yang, L. Sun, and S. Yang, "Automatic player detection, labeling and tracking in broadcast soccer video," *Patteren Recognition Letters*, vol. 30, pp. 103–113, 2009.

[14] C.-Y. Chiu, P.-C. Lin, W.-M. Chang, H.-M. Wang, and S.-N. Yang, "Detecting pitching frames in baseball game video using markov random walk," *Proc. Intl. Conf. on Image Processing (ICIP)*, 2010.

[15] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing, "Players and ball detection in soccer videos based on color segmentation and shape analysis," *Proc. Intl. Conf. on Multimedia Content Analysis and Mining*, vol. 4577, pp. 416–425, 2007.

[16] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden markov models," *Proc. Int. Conf. on Image Processing (ICIP 02)*, vol. 1, pp. 601–612, 2002.

[17] M. Goyani, S. Dutta, G.Gohil, and S. Naik, "Wicket fall concept mining from video using a-priori algorithm," *Proc. International Journal of Multimedia and Its Applications (IJMA)*, vol. 3, Feb. 2011.

[18] S. Abburu, "Multilevel semantic extraction for cricket video text processing," *Intl. Journal of Engineering Science and Technology*, vol. 2, pp. 5377–5384, 2010.

[19] M. H. Kolekar and K. Palaniappan, "Hidden markov model based structuring of cricket video sequences using motion and color features," *Proc. Intl. Conf. on Vision, Graphics and Image Processing (ICVGIP)*, Nov. 2005.

[20] M. Bertini, A. D. Bimbo, and W. Nunziati, "Automatic detection of players identity in soccer videos using faces and text cues," *Proc. of ACM Multimedia*, pp. 1663–1666, 2006.

[21] G. Zhu, C. Xu, Q. Huang, and W. Gao, "Automatic multi-player detection and tracking in broadcast sports video using support vector machine and particle filter," *Intl. Conf. on Multimedia and Expo (ICME)*, pp. 1629–1632, 2006.

[22] M. H. Kolekar and K. Palaniappan, "Event detection and semantic identification using bayesian belief networks," *Proc. of IEEE Workshop on Video-Oriented Object and Event Classification (ICCV)*, pp. 554–561, 2009.

[23] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing, "Player action recognition in broadcast tennis video with applications to semantic analysis of sports game," *Proc. ACM Multimedia*, pp. 431–440, 2006.

[24] W.-L. Lu, J.-A. Ting, K. P. Murphy, and J. J. Little, "Identifying players in broadcast sports videos using conditional random fields," *Proc. Intl. Conf. on Computer Vision and Patteren Recognition (CVPR)*, pp. 431–440, 2011.