Sports Highlight Generation System Based on Video Feature Extraction

Amruta D. Aphale[1], P. M. Kamde[2]

Computer Engineering Department
Sinhgad College of Engineering,
Pune, India

Abstract— **Multimedia consists of Video, Audio, Images and Text together. Of which Video type multimedia is the most rapidly being used and popular too, due to its high capability of providing information and entertainment. In order to fetch the intended information from a huge video database or a video, it is complex and lengthy task. It is because of its low level feature and high level video semantic concept. A video database describes what actually happens in a video and its perception by a human which is termed as Semantic Information. There is a need for a mechanism which could help more rapid browsing of a video and its retrieval because the availability of a video has been increased at large extent. One of the important functionality of Multimedia system is Video browsing, which permits the user an efficient way to fetch the intended information from a large amount of a video database. Whereas video retrieval facilitates user to search for a specific desired segment in a video based on the description provided. In this paper, we have presented the work for improvised video feature extraction for cricket highlight generation. These days we have chunk of national and international broadcasting sports channels which are continuously broadcasting the sport events happening across globe 24*7. Even though having these facilities one cannot stick to view complete event due to time constraints. This paper aims to outline the advance concepts for cricket highlight generation which is an effort towards summarization.**

Keywords— Browsing, event detection, multimedia, retrieval, semantic gap, video database.

I. INTRODUCTION

Video is the collection of continuous frames which is normally displayed at rate of 25 fps. The rich sports video content has much difficulty for the users to access and edit their favorite portions of sports games from huge amount of sports videos. It is clear that when accessing lengthy and voluminous sports video content, the ability to intelligently analyze that video to allow efficient browsing, indexing, enhancement and retrieval of that video content is crucial.

Since past decades extensive research has been put in for sports video analysis and application due to its high commercial potential and huge viewership across globe. [1],[2]. Various approaches and prototypes have been proposed and developed to analyze sports video content to fetch detailed events or highlights, intelligently adapt, enhance and personalize the content to meet users preferences and match network/device capabilities. Due to advancement made on object detection and tracking, applying data mining techniques in large video database has now become possible. Previous researches has focused on semantic video classification for indexing and retrieval or creation of video summary, but knowledge extraction on the activity contained in the video has been only partially addressed. Video event analysis and recognition is vital task in many applications such as incident detection in surveillance video, sports highlights, indexing human-computer interaction [4].

Cricket, is one of the most popular sports having very high viewer ship. Multiple television broadcasters like ESPN, Star Sports, Ten Sports etc have huge databases of this sport video. There are also video whose length lasts for 4-5 days, hence fetching meaningful videos having larger interests to viewer is crucial part. Each frame needs to go through a detailed analysis to suit for a specific event for which it has been requested for. Across globe countries like India, South Africa, England, Australia, Sri Lanka etc plays cricket. Even having this huge viewer ship, Cricket has not gained a position in research community [5][7]. Analyzing Cricket Video is very complex and challenging because of the complexity of game itself. If Cricket is being compared with other games like soccer, tennis, basketball etc, it has got more variable factors than any others of these. The variable factors like field area, pitches. Various formats like test series (4-5 days), one days and the most popular T20-20, day and day/night matches which causes illumination related problem and duration [7]. If time frame is compared with sport like soccer which is played for only 90 minutes, the latest version in Cricket T20-20 last at least for 180 minutes which is twice the duration. At present very few systems are implanted for cricket highlight generation. As the target rating point (TRP) of the media is hiked for the channels that are able to efficiently present the news before any competitor channel produces it. The easiest way to achieve it is to fetch one or more scalar or vector features from each frame and to define distance functions on the feature domain. Alternatively the features themselves can be used either as events for clustering the frames view.

In this paper we have presented an approach for cricket video event detection. The system facilitates a novel technique of selecting semantic concepts and the events within the concepts, according to their degree of importance. Different importance

may be assigned by different group of users. There could be viewer say general viewer who may like to have comprehensive viewing of all important actions, whereas specialist viewer may like to view actions of their choices. This approach facilitates such customized highlight generation, by assigning event importance. Event detection is hence distinct from the problem of human action recognition, where the primary task is to categorize short video sequence of an actor performing an unknown action into one of several classes. It help to assess the relevance or value of information within a shorter period of time while decision making. The use of only basic input as video is limiting the event retrieval and indexing capabilities of the user. The goal of this project is to expand the ways that people are able to interact with their computers. Visual features are extracted from individual frames and are trained to classify them into a category such as sixes, bowling, replay, crowd, etc. Namely, we wanted to enable users to interact more naturally with their computer by using simple GUI and perform various event detection for genre specific sports domain. In this project, a system is developed which enables to detect events in cricket video like SIXES and extract features like CLOSE-UPS, REPLAYS and CROWD etc. which is more directly interacted by the user and can summarize all the events which can be used for specific review. This system is simple enough to run and requires little training.

II. RELATED WORK

A multi-level hierarchical framework was used to extract semantic events from cricketing video. This approach basically utilizes audio visual details to categorize a single segment of video [4]. Alternatively, it has been shown that minimal events in a cricket video can be categorized availing camera motion parameters [8]. In addition to this a textual segmentation has been offered where cricket commentaries gets highlighted for the live match to describe a cricket video [5]. “Hidden makov” model and MPEG-7 visual signifiers are another technique which is being used to identify cricket highlights [4]. The de-tailed information of a video has two vital prospects [3][6][7].

A. Spatial Aspect:

Minute detail showcased by a video frame, like characters, location and various objects in that frame is termed as spatial aspect.

B. Temporal Aspect

Semantic details showcased by a sequence of video frames in time like various object movements, action of a character presented in the sequence. This is termed as temporal aspect.

The higher level information of a video is fetched by analyzing audio, video, and text annotation of the video to represent temporal aspects. This information considers identifying trigger events, identifying anomalous and typical patterns of an activity, predicting object centric or person centric perceptions of an activity, categorizing activities and deter-mining the interactions in entities. Temporal aspect prevents the effective browsing of these very large databases. Multiple studies are being conducted to draw association in low level visual features and high level semantic concepts for image annotation [4]

Video events contain rich semantic information which are normally defined as the interesting events which capture user attentions. For example, a soccer goal event is defined as the ball passing over the goal line without touching the goal posts and the crossbar. Kolekar et al [7] proposes a method to generate highlights based on event selection and giving that event an importance value based on user feedback (manual). In text driven temporal segmentation, annotated text data from a website and align the video to it. In this case, text annotation may not be available always especially in the case of old cricket matches. K Bhattacharya et al [5] have proposed a machine learning based approach for performing a shot segmentation in a neuro-fuzzy framework. This requires lots of annotated training data which increases manual intervention. Also the major disadvantage of real time cricket video is to generate frames as there are various video format like MPEG, AVI etc along with different frame rate, conversion rate etc. So it demands for a system which is independent of the discussed parameters and need a generic approach. So in our system we have used an approach to convert any type of video into frames by fetching screenshots at run time and crop them to increase the accuracy and save them in a folder as per the users convenient location. We have not used hierarchical approach as the retrieval time gradually increases. Hence with our approach frames are retrieved directly using the algorithms discussed in the next session

III. PROPOSED METHOD

For huge video databases, Computation time is the crucial thing. In this paper we propose a key frame detection approach for minimizing the computation time. T20-20 matches offer plenty of events in shorter duration of time as compared to ODIs. So we have concentrated our work on T20-20. The same approach can be used for ODIs. And can be extended to multiple types of videos like movies and news etc. Based on set of fixed or gradually changing parameters of camera such as Close up, audience, field etc, when these parameters are organized in a sequential frames it forms a shot. And collection of shots is termed as Scene. A series of related Scenes form a sequence and clip is the part of that sequence. A video is a composition of different story units like clips, scenes and are sequentially arranged with respect to some logical structure as defined in Screen play. In our prospect we fetch the events in form of a clip and after analyzing the same, a descriptive label is assigned to each clip. The architectural design for this system is as in figure 1:

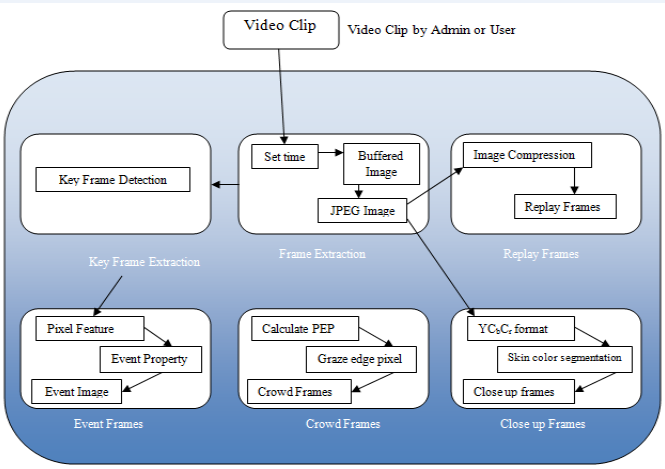

Fig. 1 System Architecture

A. *Conversion of Video to Frames*

Reading the video directly and processing it is a monotonous task and requires huge memory and high configuration systems. Instead of giving video as a direct input, we converted the video into sequence of frames first. Then every frame is an individual image and all image processing algorithms can be applied to these captured frames. Major advantage here is ease of use of image processing videos and the size of the video does not matter in this case. Also with this approach no need of specific memory requirements.

The video can be reconstructed from the frames, by simple looping operations, after performing all image processing operations. A popular method to identify frame boundaries is to compute the color histogram of consecutive frames, as in fig 2. If their color histograms are similar it means that successive frames belong to the same shot. Here we are using the RGB color histogram to compare two frames. The RGB components of a frame are quantized into 4 (red), 4 (green) and 4 (blue) bins respectively, leading to a total of 4 * 4 * 4 = 64 bins. A shot boundary gets identified when the histogram difference between two successive frames crosses a threshold. This technique works well when there transitions are rapid or hard-cut. Fig 2 shows the frames generated by this approach.

Fig. 2 Extracted frames post video conversion

Each video frame sequence consists of properties like few of the frames in a generated sequence are similar while other differs in the frame sequence. Hence to categories all such frames its computationally inefficient Consider a sequence for frames which are same to be s and d as different frames are having pattern as

s s s d s s s

For such a frame sequence in the hue histogram difference (HHD) plot there would be a spike with a rising edge and falling edge. In order to identify key and non-key frames we have to consider only rising edge.

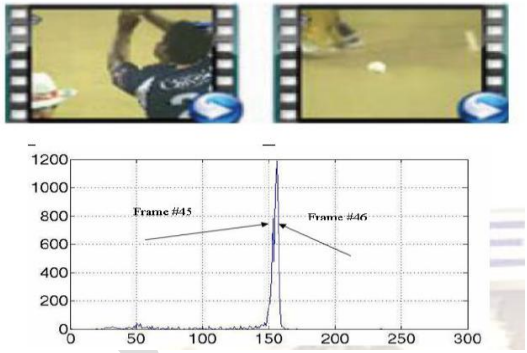

Fig. 3 Hue Histogram Difference

Algorithm Key Frame Detection:

1. Convert the image from RGB model to HSV model
2. Compute the frame wise hue difference.
3. if (diff(i) - diff(i-1)) > kfthres then
   Classify frame i as key frame
   else
   Classify frame as non key frame

B. *Visual Features*

Below are few of the visual features used for shot categorization:

i) GPR (Grass Pixel Ratio): If a hue component lies between 48 and 68 (determined experimentally) for a pixel then it's recognized as a grass pixel. For a frame first histogram is computed for a hue component which is quantized in 256 bins. GPR is the ratio of the pixel count in bins 48-68 to the total number of pixels.

ii) EPR (Edge Pixel Ratio): If a frame contains a presence of a crowd that can be detected by executing canny edge detection on a given image. Hence we calculate the ratio of edge pixels to the total number of pixels on a frame (EPR).

iii) SCR (Skin Color Ratio): The presence of fielders, umpires on a frame can be detected by looking at the percentage of skin color pixels in the frame. The Skin Color ratio is calculated by dividing frame into 16 equal blocks and calculate skin pixel ratio on each of them.

C. *Close-up (CU) Detection*

A photographic technique which tightly frames as person or object is Close-Up (CU). In films it is applied to guide audience attention and to evoke audience emotion. For CU

detection we have used Haar features wavelets which use single wavelength square waves (one high interval and one low interval). The presence of a Haar feature is identified by subtracting the average dark region pixel value from the average light-region pixel value. If the difference is above expected value (set during learning), that feature is said to be present. To identify the presence or absence of hundreds of Haar features at every image location and at several scales efficiently, integration is done. The filters at each level are trained to classify training images that passed all previous stages.

D. Crowd Detection

we see that close up or crowd frames are shown frequently whenever an exciting event occurs such as when a wicket falls, close up of batsman and bowler, then view of spectators and the players gathering of fielding team are certainly shown. The edge detection is then performed by finding the maximum gradient value of a pixel from its neighboring pixels. If the maximum value of gradient satisfies the threshold than the pixel is classified as an edge pixel [3].The percentage of edge pixels (PEP) are used to classify the frame as crowd or close-up, since we typically observe more edge pixels for crowd frames. We applied canny edge detector and use the following ratio as the close-up detection parameter:

$$PEP = \frac{\text{total number of edge pixels}}{\text{total number of pixels in the frame (EPR)}} \times 100 \quad (1)$$

The canny edge detector is used to smoothen the images to eliminate noise. Once done it then finds the image gradient to highlight regions with high spatial derivatives. The algorithm then goes after these regions and suppresses any pixel that is not at the maximum.

Algorithm crowd detection:
1) Convert the input RGB image into YCbCr model.
2) Apply canny operator to detect the edge pixels.
3) Compute Percentage Edge Pixel (PEP) for the image.
4) Classify the image using following condition:

if PEP > PPEP then
frame belongs to class crowd
else
frame belongs to class close-up

E. Replay Detection

Motion vector [5] and replay structures [7] are used to detect replays from sports video. But these methods are not robust enough to be suitable for various kinds of sports video replay detection as replays in different sports video are captured in various ways and compiled in different manners which can hardly be represented by such simple features. Hence the recent approach is to detect the accompanying logo effect of the replays in sports videos to acquire the replay segmentations. It is been commonly observed that a replay segment is always machinated between flying graphics or two logo transitions which last for 8-15 frames. The following pseudo code underlines the brief concept that is used for replay detection for two consecutive frames

Algorithm replay detection:
1) For each frame i (for Image 1)
2) For Each Frame j=i+1 (for Image 2)
3) Rgbvalue1 of each Pixel (Getting RGB Value for first image)

4) Rgbvalue2 of each Pixel (Getting RGB Value for second image)

5) if(Rgbvalue1== Rgbvalue2)

frame belongs to Replay Event
else
frame belongs to non Replay Event

F. Sixer Event Detection

In our proposed system we are working on T-20 cricket format videos. We have considered matches which are played in night. When a batsman hits a sixer, it is common observation that the ball is raised up high and will be in the air for a while. At this time cameras tracks the ball as well as the background of the shot which is of course in black color. So our system extracts the sixer frames based on the following algorithm.

Algorithm Sixer Detection
1) For each frame i
2) Convert RGB value of all pixels from binary to integer
3) Initially blackpixel count is 0
4) if red, green and blue values range between 0 to 50 then increment blackpixel count
6) Calculate Black Pixel Percentage as
BCP= (Blackpixelcount/total number of pixel) *100
if BCP greator than or equal to 50 then Frame belongs to Sixer event

IV. MATHEMATICAL MODEL

For Short-time audio energy,
Let
S= sample audio for corresponding to one video frame.
X = time signals
N = total number of videos

$$E(n) = \frac{1}{s} \sum_{m=1}^{s-1} [x] \quad (2)$$

Where

$$W(m) = \begin{cases} 1 & If\ 0 \\ 0 & Otherwise \end{cases}$$

Event detection of video: We consider sliding window, which helps us to identify early detection of events.

$$E'(n) = \frac{1}{L} \sum_{t=0}^{L-1} E(\text{n+1}) \quad (3)$$

and

$$Z'(n) = \frac{1}{L} \sum_{t=0}^{L-1} Z(\text{n+1}) \quad (4)$$

Here
L = Sliding window time

$$E''(n) = \frac{E'(n)}{max_{1<=i<=N} E'(i)} \quad (5)$$

$$Z''(n) = \frac{Z'(n)}{max_{1<=i<=N} Z'(i)} \quad (6)$$

Here N represents total number of videos

P-frames are created from equation 5 and 6.
At *K=1*, *K+1=2*, *X(1)* and *X(2)* are first and second edge detected frames. The difference between these 2 frames is first *P* frame. The difference between second and third edge detected frame is second *P* frame and so on.

$$P(k) = (x(k) - x'(k+1)) \quad (7)$$

Where,
k=1, 2,…,n-1
x(n) = edge_images(n)(i,j)
x'(n) = edge_images

P-frames are binary images, in which white pixels indicate the difference between two consecutive frames. Numbers of white pixels from all n-1
P-frames are computed from equation 8.

$$m(n) = \sum_{i=1}^{i=totalrows} \sum_{j=1}^{j=totalcolumn} \text{(x-x')} \quad (8)$$

Frame detection: A video frame n is labeled as:

$$P(n) = E''(n) \times z^n(\text{n}) \quad (9)$$

Extraction of correlation coefficient follows as

$$r = \frac{\sum m \sum n(\text{Amn-A})\overline{(\text{Bmn-B})}}{\sqrt{(\sum m \sum n(\text{Amn- A})^2)(\sum m \sum n(\text{Bmn-B})^2)}} \quad (10)$$

V. EXPERIMENTAL RESULTS

We have defined the events as scenes in the video with some semantic meaning and action associated with it and some features which are to be used as the system processes all the frames. So we have labeled features as close up, replay and crowd and highlighted event as sixes. We have implemented the system using java net beans with the intention to provide a hassle free solution for event and feature detection for cricket video. The dataset consists of cricket video which captures frames at rate of 20 fps (variable to change) so as to get clear images to label a specific feature or event. The smallest video which was tested was for T20-20 worldcup-2006, played for duration 4 min 52 seconds. Our system extracted 292 frames with the rate of 1 frame per second. When we perform and extract the events like Close-up, Replay, crowd and sixer we found very impressive result by the system. We have plotted the same in below graph, which matches the system extracted frame result with the human extracted frame result.

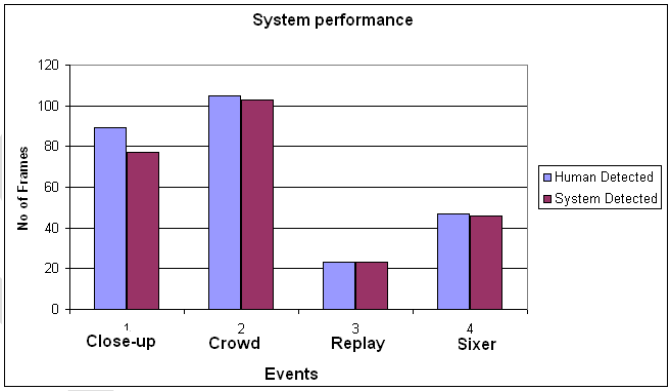

Fig. 4 System extracted frames per event

And also we perform the same kind of experiments with many videos which gives similar performance results. The aggregate result is shown in the table.

TABLE I. AGGREGATE RESULT OF VARIOUS EVENTS

Events	Performance percentage
Closeups	84.34
Crowd	97.44
Replay	100
Sixer	95.4

VI. CONCLUSION

With a minimum hardware resource an attempt is made to provide a comfortable, attractive solution for most important event detection in a more natural way.
The use of classification technique can exhibit better detection of various events and analyses of those video events which can reduce the processing time thus save lot of CPU usage. This can create a new scope to generate customized and automatic cricket video highlights for different purposes. This approach can be extended to other sports like soccer as well as other types of videos such as news, movies, etc. for video summarization applications.
The volume of databases is very huge in case of video. Thus computation time for event detection and analysis is a critical issue. We have concentrated our work on T-20 matches as they offer lots of events in short duration as compared to ODIs. The same work can be extended to ODIs as well.

Key frame event detection approach turns out to be most efficient method for minimizing the computation time with more accuracy in event detection capability. This categorization results with better classification ratio at various levels.

Following figures depicts few of the classified frames.

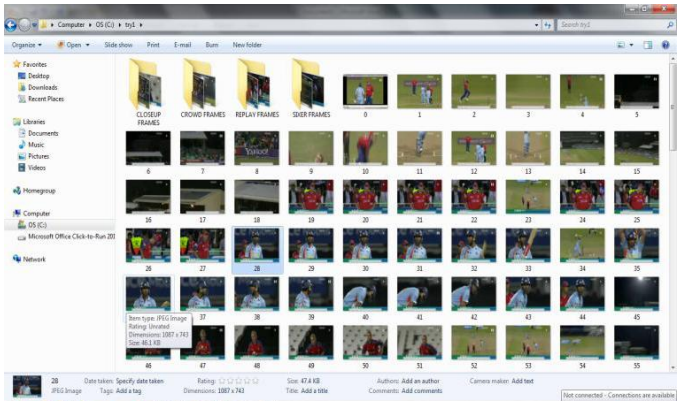

Fig. 5 Resultant output folders of event detection

Fig. 6 close-up Frames

Fig. 7 Replay Frames

Fig. 8 Crowd Frames

Fig. 9 Detection of six

REFERENCES

[1] Maheshkumar H. Kolekar, Kannappan Palaniappan Semantic Concept Mining Based on Hierarchical Event Detection for Soccer Video Indexing, journal of multimedia, vol. 4, no. 5, October 2009.

[2] Mahesh Goyani, Shreyash Dutta, Gunvatsinh Gohil,Sapan Naik, wicket fall concept mining from cricket video using a-priori algorithm, The International Journal of Multimedia and Its Applications (IJMA) Vol.3, No.1, February 2011.

[3] N.Harikrishna, Sanjeev, Satheesh, S.Dinesh, Sriram, K. S. Easwarakumar, Content Based Image Retrieval using Dominant Color Identification Based on Foreground Objects ,IEEE transactions on multimedia, vol. 10, no. 3, April 2008.

[4] Hu Min, Yang Shuangyuan, Overview of content-based image retrieval with high-level semantics 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE).

[5] Dr.N.Krishnan, M.Sheerin Banu and C.Callins Christiyana, Content Based Image Retrieval using Dominant Color Identification Based on Foreground Objects, International Conference on Computational Intelligence and Multimedia Applications , 2007 IEEE, DOI 10.1109/ICCIMA.2007.64

[6] Yu Xiaohong and Xu Jinhua, The Related Techniques of Content-based Image Retrieval, 2008 International Symposium on Computer Science and Computational Technology.

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

[8] H.B.Kekre, Dhirendra Mishra , DCT Sectorization for Feature Vector Generation in CBIR , International Journal of Computer Applications (0975 8887) Volume 9 No.1, November 2010

[9] P. B. Thawari and N. J. Janwe, CBIR Based On Color And Texture, International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 129-132