Electronic Letters on Computer Vision and Image Analysis 21(2):1-22, 2022

Cricket Video Highlight Generation Methods: A Review

Hansa Shingrakhia* and Dr. Hetal Patel+

* *PhD Scholar, Gujarat Technological University, Gujarat, India*
+ *Professor, ECE Department, A. D. Patel Institute of Technology and Engineering, Gujarat, India*

Received 19 July 2021; accepted 14 August 2022

Abstract

The key events extraction from a video for the best representation of its contents is known as video summarization. In this study, the game of cricket is specifically considered for extracting important events such as boundaries, sixes and wickets. The cricket video highlights generation frameworks require extensive key event identification. These key events can be identified by extracting the audio, visual and textual features from any cricket video.T he prediction accuracy of the cricket video summarization mainly depends on the game rules, player form, their skill, and different natural conditions. This paper provides a complete survey of latest research in cricket video summarization methods. It includes the quantitative evaluation of the outcomes of the existing frameworks. This extensive review highly recommended developing deep learning-assisted video summarization approaches for cricket video due to their more representative feature extraction and classification capability than the conventional edge, texture features, and classifiers. The scope of this analysis also includes future visions and research opportunities in cricket highlight generation.

Key Words: Cricket video summarization, shot boundary detection, key frame extraction, Excitement and event-driven highlight generation.

1 Introduction

Cricket is the second most loved game of people in the world after football. Instead, it is the most loved game in India. Particularly the Twenty20 format is very famous due to its rapidity which fascinates the audiences at the ground and the watchers at home. In the world, several individuals know and play cricket. Compared to other sports, cricket is a challenging game because of the longer match duration as well as the use of a large number of rules. Nowadays, sports broadcasters generate a massive number of recorded or live broadcasts of different sports Videos. The viewers, however, prefer to watch sports highlights than the full length videos. The highlights summarize the entire video by including all the important events [1]. The commercial importance and vast viewership make highlight generation process an important research area. Manual generation of a highlight is a tedious process and hence, automatic summarization methods, which analyse the contents of the sports video to determine the important events.

The existing sports video assessment methods can be categorized into two types: genre-specific and genre-independent. Work done on the genre-independent method is less because of the different game formats of

Correspondence to: <hansa.shingrakhia.2013@ieee.org>

Recommended for acceptance by <Angel D. Sappa>
https://doi.org/10.5565/rev/elcvia.1465
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

sports. Even though this type of general structure is well suited for highlight generation, it can't express the highlights meaningfully to the viewer. A genre-specific method is used in the sports video analysis of cricket because of its static spatial and temporal framework [2–4]. Cricket is considered as one of the most popular game in the world due to its peak rate of viewership. Compared to other sports, cricket is a challenging game because of longer match duration as well as the use of large number of rules. Hence, work done on cricket video highlight generation is less compared to other sports videos. The cricket highlights can be automatically generated by event-driven or excitement-driven methods. The game rules are considered in event-driven methods, whereas excitement-driven methods consider action from crowds, commentators and players [5, 6]. The complexities of event-driven methods are high, and more false positives weaken the performance of the excitement-driven one. The excitement-driven methods depend on the context of the video. Thus, the audio and video features are required to extract the key events of a cricket video.

Generally, the color, edge, texture, and salient-points are estimated to compute the visual features [7–9]. Some of the audio features of any audio stream are Zero crossing rate (ZCR), Mel-frequency cepstral coefficients (MFCC), short-time energy (STE), and so on [10–12]. After extracting the appropriate features, the key event identification process is reduced to sequential or temporal classification. The key events must be detected accurately without missing any important events and not including any uninteresting moments. Thus, efficient cricket video summarization methods must include entire key events in the highlights concisely. This paper thoroughly analyzes the recent highlight generation approaches for cricket video.

The paper is structured as follows: Section II provides the review objectives and selection of sources. Section III gives the generic structure of cricket video summarization along with shot boundary detection methods. Different classes of cricket video summarization methods are reviewed in Section IV. The performance measures and evaluation of existing methods for cricket video summarization are defined in Section V. Section VI provides the challenges and future recommendation. Section VII concludes the paper.

2 Review Objectives and selection of sources

The sports broadcasters summarize and distribute several sports videos over different networks allowing people to watch important events from them. The sports videos are summarized to enable transmission over low-bandwidth networks, reduce storage cost and handle time constraints. It is very difficult to select significant information from the sports video due to large number of digital video contents. An efficient video summarization method should extract the contents of the video automatically with fewer time constraints. It should capture viewer interest by exploiting only exciting segments of the original video. The automatic analysis of sports events is considered as a major demanding task over the past few years.

Cricket is the second most famous game in the world, which has approximately 2.5 Billion fans worldwide and a total of 105 countries playing this game. An officially authorized league cricketer in the world is around 1 million. The non-official cricketers are around 60 million, among that 55 million are from India. According to the statement of the Broadcast Audience Research Council (BARC), India has drawn 93% of Cricket spectators. But the spectators wish to view highlights of the matches as compared to the full-length videos. Hence, every broadcasting channel like Star Sports 1 showcases several best of live Indian and International cricket and related programs in English every day. These statistics show the look interesting of cricket highlights generation approaches to the audience. But Cricket is a challenging game compared to other sports events because of its longer match duration and the usage of the enormous number of rules. These points motivate us to study the available literary work on cricket video summarization.

This study aims to summarize the recent research on cricket highlight generation and determine the challenges associated. This paper examines the significance of various methods for key event extraction and also identifies the need for excitement and event- driven based cricket video summarization to include more meaningful content in the highlights with less complexity.

For reviewing the papers on cricket highlight generation, the online scientific databases are searched. The following keywords are used as the search strings:

1. Automatic cricket video summarization
2. Video Shot detection and classification for video summarization
3. Key event extraction methods for generating highlights

The search is directed from 2008 to 2021 using Google Scholar. From a fast look at Google Scholar for "cricket video summarization", from 2008-2021, there have been over 700 papers published. It perhaps shows a trend of the growing interest in cricket video in the past 13 years. In this review, the distributed papers in journals and conferences are selected by the accepted literature search engines and databases such as Springer Digital Library, Taylor and Francis, Google Scholar, Science Direct, ACM Digital Library, Wiley online library, Elsevier, IEEE Xplore and so on. According to the search query, the papers are selected. The articles are initially selected based on subjects. After that, the headings of the papers are verified and inappropriate headings are discarded. Then, the selected papers are reviewed and the unsuitable papers are rejected. After filtering the inappropriate and unsuitable papers from the accepted journals, the total number of papers selected for this review is around 103. Table 1 shows the number of relevant papers collected from journals/conferences of different computer societies. It reveals that 36.53 % of the papers were collected from IEEE, 21.15 % were from springer, 3.84 % were from Elsevier and 3.84% were from ACM, in total representing 65.36% of the references. The remaining were distributed across all other societies.

Sr.No.	Name of the computer society	Number of journal papers	Number of conference papers	Percentage
1	IEEE	9	29	36.53%
2	Springer	18	4	21.15%
3	Elsevier	4	-	3.84%
4	ACM	1	3	3.84%
5	Others	32	4	34.61%

Table 1: Number of published papers on different computer society from 2008-2020

3 Cricket video summarization

The general structure of cricket video highlight generation is shown in Fig.1. This structure consists of different stages, starting from shot boundary detection to key frame extraction and highlights generation [13].

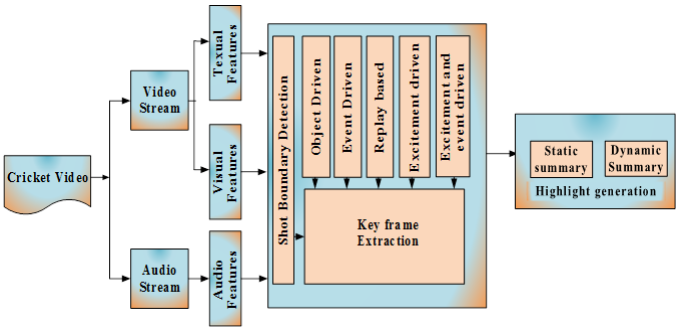

Figure 1: Generic framework of sports video summarization [13]

All the input sports video contain video as well as audio streams [14]. The video summarization methods function by extracting features like audio, video and text. The temporal video segmentation is the basic procedure for all video processing applications. It can be obtained by detecting the shot boundaries. In general, the

shot boundaries are detected for segmenting the video into shots. There is a gradual or abrupt transition in the detected boundaries. After identifying shot boundaries, the key events are extracted for generating the highlights. The significance of video shots is represented by these key events [15]. The key events can be extracted based on different contents of video like events, objects, replays and excitements. Finally, the highlights are generated by summarizing the extracted key events.

3.1 Audio descriptors

The audio stream features like speech, sound and music are represented by the audio descriptors. They play an important role in sports video summarization. The sports video contains audio contents such as the commentator's speech and the spectator's excitement. Audio keywords are created from the playing shots based on different events like applause, silence, and hitting. These audio contents detect the key events such as boundaries, fall of wicket and sixes in cricket. The highlight generation systems use the audio features due to their lower computational complexities. Also, it uniquely determines highlight-worthy segments and is complementary to visual information. The time and frequency domain features are the two major classifications of audio. One can extract the time domain features from the audio signal without transforming it into another domain. The features extracted from the audio signal after transforming it into DFT domain are called spectral or frequency domain features. The DFT is mostly used to process the audio signal, because it represents the audio signal in a meaningful way by showing its frequency distribution.

ZCR, STE and entropy are the popularly used the time domain audio features. M.H.Kolekar *et al.* [16] used the audio features like STE and short-time ZCR to extract the exciting clips. Frequency domain features include spectral flux, spectral entropy, spectral roll-off and MFCC. Baijal *et al.* [11] used MFCC for training a Gaussian Mixture Model (GMM) classifier to classify the audio into speech and non-speech signals. Vincenzo *et al.* [17] proposed a new model called SFERAnet (Selection of Football Events by Recorded Audio). A Neural Network is trained by this model for identifying key events. Here, the features are described with the spectral characteristics such as MFCC, Mel bands decomposition, centroid, spread, entropy, flux, roll-off, and ZCR.

The deep learning-based analysis of audio signals can describe the crowd-cheering and commentator tone excitement. Merler *et al.* [6] used deep Soundnet to detect audio based markers in sports video. It uses deep 1-D convolutional neural network architecture (CNN) to learn representations of environmental sounds from several unlabelled videos. In addition, the state-of-the-art deep learning (DL) approaches such as Deep Neural Network (DNN), Recurrent Neural Network (RNN), Convolutional Deep Neural Network (CNN) and the combination of these models (DNN, RNN and CNN) can be used to detect the environmental sounds [18] from the sports video. Li *et al.* [19] extracted the deep audio features with the help of a multi-stream hierarchical Deep Neural Network (DNN) to detect an acoustic event. Here, individual pre-trained Restricted Boltzman Machine (RBM) pairs build DNN.

3.2 Visual descriptors

The visual descriptors can be classified into static and dynamic descriptors. The static descriptors extract the features from the individual frames of sports video, specifically, the centre key frame of all the sports video segments. The visual features like texture, color and shape are represented by the static visual descriptors. Instead, the dynamic descriptors consider the dynamic nature of sports video and extract the motion information using all the frames of the segment. The visual descriptors can be classified into general information and specific domain information descriptors. The fundamental features like texture, color, motion, shape and regions are covered by the low-level general information descriptors. However, high-level features are depicted by specific domain information descriptors that give statistics of the objects and events. The dominance of field color in sports video is used to classify the field views of cricket. The scoreboard region and contents of cricket video such as batting and bowling statistics of each team, and batsmen and bowler's profile can be recognized using the texture and shape which further determine high-level concepts. The visual saliency features can also be

used in the video summarization model. This feature is constructed using intensity and color information [20].

Motion descriptors determine frequent camera motions in cricket videos. Specifically, they detect replay events. The key frames can be detected through the correct measurement of motion in a video. Thus, Kamoji *et al.* [21] used motion activity descriptors for extracting the key frames. They used block matching method to obtain the motion, a process considered as the key part. It was implemented using two techniques, Diamond Search and Three Step Search. The color, texture, shape and motion features are considered more valuable in the cricket highlight generation. Karmaker *et al.* [22] used motion vector to identify the cricket batting shots as Square Cut, Hook, Flick and Off Drive. This method measured the angle of the cricket shot using the motion vector. This motion vector is found using the modified approach of Lucas Kanade, that determines the optical flow for each of the frames. Optical flow measures the velocity of the image.

Roopchand *et al.* [23] proposed a bat detection method using Optical Flow and Otsu's Thresholding. The body of the batsmen and the bat moves naturally from frame to frame, while the background remains stationary. In this work, this basic concept is used to track the bat. The bat detection and tracking methods classify the batting strokes. Visual features are represented either by discretization and clustering method or by using a kernel density estimate.The video summarization approach that used conventional hand-crafted features failed to seize the data and substance from all the scenes. For tackling this issue, scene shots can be detected using motion features [24]. More recent video summarization methods use deep neural network-based features to extract higher-level concepts. Recent research efforts on CNNs have revealed that the activations of a higher layer of CNN can be powerful visual features. CNN-based image/video representations have been explored for various tasks including video summarization [25].

3.3 Textual descriptors

The data required for summarizing the video content based on textual cues is provided by the textual descriptors. They can be categorized based on textual overlays, closed captions, open captions, external transcripts, speech recognition, and graphical inserts. The important information about a game is provided by these textual descriptors. The cricket video may consist of different graphical inserts like wagon wheel score chart, run rate graph and player's statistics chart to represent the game statistics. The data existing in the game statistics can detect key events, which are identified by matching the game statistics with textual overlays. One can also use speech recognition to extract different events of cricket videos. The interesting moments of cricket can be detected by integrating both visual and textual features.

3.4 Shot boundary Detection

The transition between two shots is represented as a shot boundary, and the frame series captured with a single camera is represented as a shot. Every over in the cricket match mentions six acceptable deliveries. A single bowler can throw these deliveries sequentially. Every delivery holds a collection of video shots. Hence, the significant events of any cricket match will be represented using series of video shots as shown in Fig.2. Every event in the cricket match begins with a bowler's run and it is utilized for marking the start-off of the event.

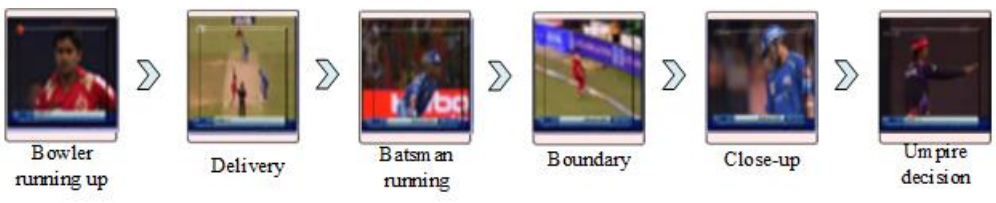

Figure 2: Collection of cricket video shots in an over

The videos can be temporally segmented by detecting the shot boundaries. All video summarization models used this shot boundary detection as a pre-processing method. The abrupt transition of the shot boundary is

also called hard cut and the gradual transition is known as fade-in and fade-out transition. It is very difficult to detect the shot boundaries in a cricket video. Most of the sports videos comprise hard cuts, and the broadcasters superimpose them at specified circumstances like replays, play breaks and so on. Most of the prevailing sports highlight generation approaches have given attention to hard cut detection.

The color histogram computation of successive frames is popularly used to determine the shot boundaries [26]. When these histograms are similar to one another, they belong to the same shot. The histogram based shot identification methods can be classified into three types; Bin to Bin, Chi Square and Histogram Intersection [27].Premaratne *et al.* [28] introduced an adaptive filter to enhance the hue histogram that improved the video shot segmentation process with great Resolution. Majumdar *et al.* [29] detected cricket video shots by comparing the SIFT features of neighboring frames. Initially, frames were obtained from the input cricket video. Then, the SIFT features were extracted from each frame. The features of the adjacent frames were matched to determine the key points. Finally, the ratio of matched key points to the total key points was compared with a threshold value to determine the shots.

The gap between low-level features and high-level events can be related with the help of suitable semantic cues provided by shot classification. Based on the views, the sports video shots can be classified into close up, medium, long and out of field view shots. Minha *et al.* [30] used AlexNet Convolutional Neural Network for classifying the field view. This network contained 5 convolutional layers and 3 fully connected layers. The training process was increased using the dropout layers on the feature maps and response normalization. Premaratne *et al.*. [31] classified the shots by detecting the ball starting and ending point video clips. The video frames between these clips were considered as one shot. The ball starting and end points were determined using camera focus area, pitch zooming, edge percentage, pitch availability, camera motion (X, Y and Z direction pattern) and camera change length (duration).

An adaptive approach was proposed by Patel *et al.* [32] to segment the shots automatically. This method changed the threshold value in accordance with the varying information. They defined a new single feature vector by considering histogram, χ^2 histogram, color histogram and DCT. This vector included image information such as frequency, color and texture domain transformation. This method detected all transitions like hard cuts, wipes, fades, and dissolves. Wei *et al.* [33] proposed a Sequence-to-Segments Network (S2N) to detect the segments of video through the contextual understanding of the whole video.

Apart from extracting only the high-level features, certain works illustrated the importance of handcrafted features extracted from the videos for video summarization. The deep features [34] can be extracted by evaluating the frames, and the layers in the deep model automatically extract the features with better clarity. On the other hand, the handcrafted features like the gray level co-occurrence matrix (GLCM) [35], histogram of oriented gradients (HOG) [4], etc., are optimistic and can be easily applied to a variety of sports videos for highlight generation. The handcrafted features are useful in accurately determining the field's texture and are more informative. The handcrafted features are, in general, extracted manually with the help of some instructed algorithms. The deep features are extracted automatically with appropriate training. This step helps to reduce the computational complexity and reduces the error rate in prediction as the manual extraction is liable to errors. Moreover, the manual extraction of features from the frame requires modifications and recalculations to make sure that the extraction made is accurate. Moreover, it is almost impossible to evaluate the extraction process completely. Meanwhile, the deep features discriminate between the features extracted from different frames. This helps the model to differentiate one frame from the other accurately. This process helps to understand the importance of each frame in the highlight, thereby choosing the required frames for highlight generation.

4 Classification of cricket video summarization

Figure 3 shows the hierarchical classification of cricket video summarization methods. The cricket video summarization methods can be classified as Static and dynamic depending on the output type, The collection of key frames (static images) is provided as output summary in static video summarization methods while the video synopsis of long length video (i.e. skimmed video) is provided by the dynamic summarization methods [36].

Some static video summarization methods use key-frame extraction algorithms like clustering [37, 38]. Ranjan *et al.* [39] selected the key frames using F-Sift, Tamura Textural, Middle level Semantic Features and k-means clustering approach. Then, the redundant key frames were eliminated by setting one threshold value. Finally, the key frames were composited to form static summarization of cricket video. The functions of the Convolution Neural Network (CNN) and the K-Means clustering algorithm were combined by Balas *et al.* [40] to summarize the cricket video efficiently. Based on the content type, cricket video summarization methods are

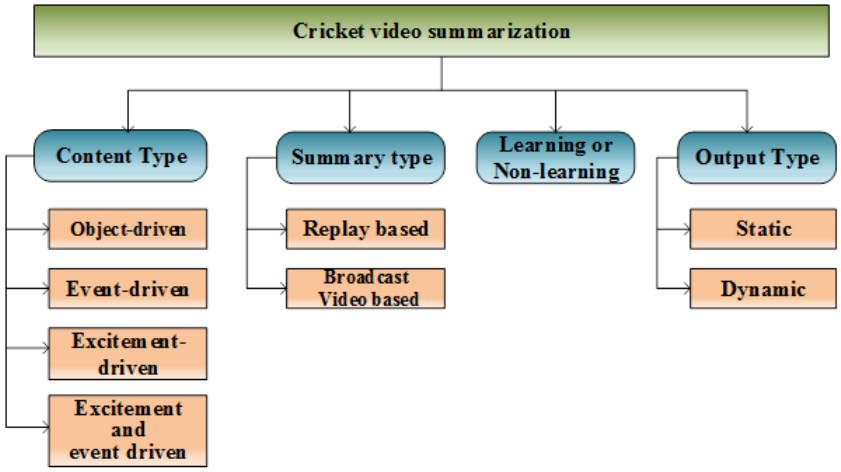

Figure 3: Classification of cricket video summarization methods.

classified as object-driven, event-driven and excitement-driven methods. The event-driven approaches use the domain knowledge of cricket for the detection of key events like boundaries, wickets and sixes [41]. Various objects, such as ball and umpires, are detected in the object-driven methods to generate highlights. Cricket highlights can also be generated by detecting the replay frames or from the live broadcasted videos. The sports broadcasters create replays for showing the exciting events when the live matches are broadcasted. Thus, an efficient way to determine significant events in cricket is replay detection process. All the existing video summarization methods depend either on learning approaches or non-learning approaches. The classification process is used in learning based approaches to determine key events. The observations made from the games such as logo placing, game transitions are used in the non-learning based approaches to generate the highlights precisely. The wrong prediction of game observations may lead to selection of unimportant key events in the non-learning-based methods. The non-learning technique may use histogram difference and contrast to identify the logo frames for replay detection.

4.1 Object-driven cricket highlight generation

The videos can be summarized on the basis of certain objects like a human, ball, etc. These are known as object driven summaries. Here, the low-level features are initially analyzed for the detection of high-level objects which are then used to create highlights. Ravi *et al.* [42] created a dataset namely, SNOW to detect the umpire pose based on SVM classifier and deep features. This paper used Inception V3 and VGG19 networks for extracting features from the umpire frames. Then, the SVM classifier was trained to detect umpire's poses during Four, Six, Wicket and Wide ball. The sample images from the SNOW dataset to illustrate the umpire pose for Four, Six, No ball, Wicket and Wide events are shown in Fig.4. The important events in cricket could be characterized based on certain gestures and postures of the umpire. The medium and close-up view shots could be used to recognize such distinctive gestures and postures [43]. Hari *et al.* [44] detected the events in cricket by capturing the umpire gestures as shown in Fig.5. Here, the scenes were initially segmented from the cricket video through the determination of bowling events. In cricket, the pitch area increases slowly from frame to frame at the time of bowling events and it consumed more region when the batsman strikes the ball.

Figure 4: Sample images of umpire pose for different events (a) Six (b) Four (c) Wicket (d) Wide.

The pitch area is not present in the rest of the frames. Thus, the authors segmented the scenes by separating the shots and by identifying the pitch. After that, the umpire frames were identified by analyzing the jersey color. As the umpires wear different colored dress it can be differentiate them from the players of both team. Then the umpire gestures were determined by projecting the horizontal and vertical intensity profiles. Finally, the random Forest classifier was trained to extract the key events. Instead, Nandyal et al [84] used HOG features and SVM classifier to detect the key frames of umpire efficiently. They also improved the umpire detection task in [45] by extracting the Area-Of-interest-based Histogram of Oriented Gradients (AOI-HOG) Features. This provided pixel-wise color information for supporting the SVM classifier to identify umpire frames.

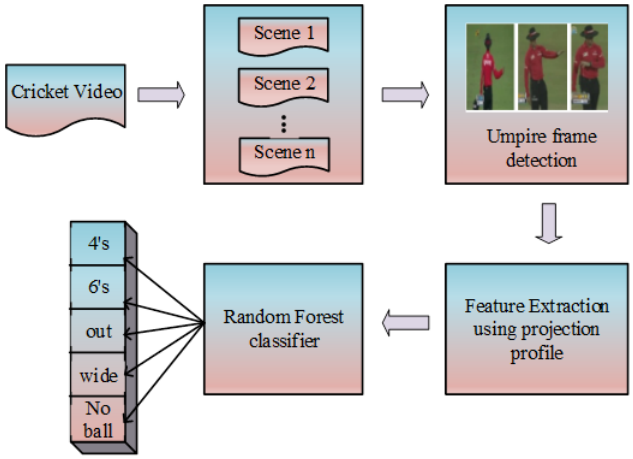

Figure 5: Cricket event summarization based on umpire frames [44]

The umpire's perception may not be correct at all events as they are human. Also, the umpires ask for confirmation of no ball events. Chowdhury *et al.* [46] applied computer vision in no ball detection approach. For that, they split the bowling crease into two sections and determined the pixel value changes in both sections by subtraction. Then, the mean of noticed pixel changes of both sections is computed and compared with a threshold value to determine the no ball event. The threshold values are determined by analyzing number of images wherein the bowler's foot heel was inside the popping crease.

Harun-Ur-Rashid *et al.* [47] used a Convolutional Neural Network to detect the waist-high no-ball event by analyzing the pitch and the ball. A well-organized Ball Detection method was proposed by B.L. Velammal *et al.* [48] for the game of Cricket. The non-ball objects were eliminated using an anti-model method and the rest of the objects were considered as ball-objects. Initially, the video was segmented using Region Growing operation. Then, the shape properties were used to classify the ball and non-ball objects. After eliminating non-ball candidates, the ball candidates were processed for the detection of the ball.

Mridul Dixit *et al.* [49] extracted the four-events from the cricket video by detecting the ball and boundary. Initially, white components were extracted from the green background. A binary search algorithm was used to determine the best fit line to locate the boundary. Then the coordinates of left-most point and right-most point were detected. For detecting the ball, the white pixels were searched in the green area below the boundary line.

The perpendicular distance between the ballpoints and the boundary line was computed after the detection of ball points and boundary lines. Finally, the four-events were extracted by defining a threshold value. When the threshold value was greater than the distance measure, the event was extracted as four. Table 2 summarizes the reviewed object detection based methods.

Authors	Contributions	Detected events	Type of classifier	Performance measures	Limitations
Ravi et al. [42]	Created a new SNOW image dataset to detect umpire pose	Four, Six, No ball, Wicket and Wide	SVM	True positive rate, positive prediction value	Not tested on real time cricket video.
Hari,R. et al. [44]	Detected umpire frames and used projection intensity profiles	Four, Six, No ball, Out and Wide	RF classifier	Precision, Recall	It can't detect the events of test cricket because the players and umpires wear white color dress.
Chowdhury et al. [46]	Image subtraction method is used for detecting the pixel value changes	No ball	-	Accuracy	Requires template frames for comparison.
Harun-Ur-Rashid et al. [47]	Introduced Crick-net using Artificial Intelligence	Waist High, No Balls	CNN with Inception V3	Recall, False positive rate, Specificity, Precision, F– measure, Accuracy	Not tested on real time cricket video.
B.L.Velammal et al. [48]	Used shape properties to detect the ball object	-	-	-	Performance is not guaranteed due to lack of shadow removal process.
Mridul Dixit et al. [49]	Used ball and boundary line to detect the event	Four	-	True Positive, True Negative, False Positive, False Negative rates	It gives accurate outputs only when boundary line and balls are white.
Nandyal et al. [50]	Used HoG features	-	SVM	Precision, Recall	Not tested to detect the important events of cricket video.
Nandyal et al. [45]	Introduced AOI-HOG features	-	SVM	Accuracy	Not tested to detect the important events of cricket video.

Table 2: Summary of reviewed object based methods

4.2 Replay based summarization

For presenting the description of the key events in slow motion, replays are usually incorporated after all the exciting moments of the game. Thus, most of the exciting moments of cricket are contained in replay frames. Kolekar et al [16] detected that replays are usually shown with logo transition representing the beginning and termination of the replay as illustrated in Figure 6 (a). In this figure, the illustrative frames of logo transition (# 899 to #914 and #1382 to #1397) are shown in the first and last row. The replay frames (#915 to #1381) are shown in the second and third rows. Thus, the replay frames can be determined by determining the logo frames with the help of hue histogram differences. However, the cricket matches are differed by various reference logos. As a result of this, different threshold values are required for differentiating the logo frame from the real frame of different matches. To tackle this issue scoreboard discovery approach has been developed. The replay frames usually do not contain a scoreboard region as shown in Fig6. Thus, the nonappearance of the scoreboard region has been checked to detect the replay frame.

To detect the replay frames on the basis of slow-motion and logo transition, several Learning and Non-learning approaches have been developed. Slow-motion method [51] is based on the homogeneousness of slow-motion speed of complete replay. But, the performance of this method is decreased due to the varying replay speed. Choros *et al.* [52] detected the replays by identifying the logo transition on the basis of contrast features and histogram differences. When the contrast of any frame was larger than the predefined threshold, it was included in the candidate collection of logo images. But, there might be a certain incorrect classification of logo images because of the bright color jersey of players. Thus, every classified candidate logo was checked by

Figure 6: Replay detection (a) Representative frames of the replay segment using logo transition (b) Score bar detection approach

examining their histograms. When any candidate's histogram hasn't differed adequately from the histograms of the previous and next fifteen frames, it was rejected from the candidate collection.

M. Ravinder *et al.* [53] detected the replay frames using correlation features and SVM classifiers. Here, the templates were initially created by selecting the the score card (SC) region from both non-replay frames and replay frames. The SVM classifier was trained using 500 non-replay frames and 500 replay frames. The correlation among the training frames and templates was determined to form a feature vector. This method predicted the replay frames automatically during testing. Javed *et al.* [54] summarized the cricket video by analyzing the replay frames only. The gradual transitions (fade-in and fade-out) between the video frames were analyzed to determine the replay frames by setting a threshold. Then, silhouettes and motion history image (MHI) of every key event were obtained by applying the GMM. A Confined Elliptical Local Ternary Pattern extracted the features of every MHIs. Finally, a trained Extreme Learning Machine (ELM) classifier was used for event detection. The SC region is not displayed in the replay frames which contain multiple gradual transitions (GT). Thus, Javed *et al.* [55] analyzed both SC and GTs for replay detection [13]. The GTs were determined using a dual-threshold technique and the candidate replay segments were extracted from these GT frames. After that, the SC region was detected for discriminating the replay frames from real frames. The process of replay identification based on both SC and GTs are shown in Fig.7.

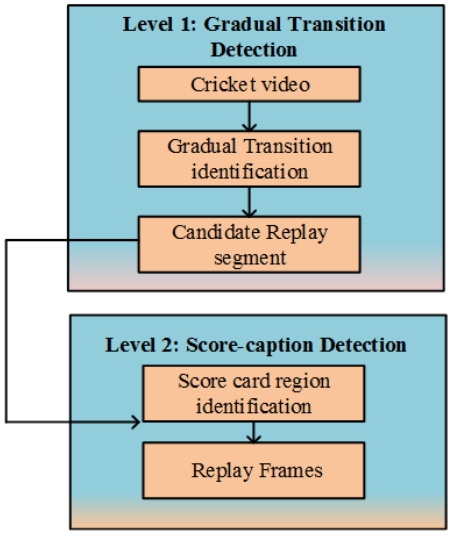

Figure 7: Block diagram of Replay detection method [13]

Narasimhan [56]summarized the cricket video with the help of a genetic algorithm (GA). Initially, the video contents were divided by identifying the shot boundaries. Then, the replay frames were identified through the analysis of the absence of score card region which is not shown during replays. In general, the summarized video should include different events, more important events and cover the entire period of the particular

match. These parameters should be optimized for the generation of user-specific highlights. Thus, the authors allowed the user for assigning the weights of such parameters. These parameters developed a multiple objective optimization problem which were solved using a genetic algorithm. The constraints of multiple objective problem was removed through the weights assigned by the user. Table 3 summarizes the reviewed Replay based methods.

Authors	Contributions	Detected events	Type of classifier	Performance measures	Limitations
Choroś *et al.* [52]	Identified the logo transition using Contrast Feature and Histogram Difference	Replay	-	Recall, precision, F-measure and Relative average time	When any candidate's histogram wasn't differed adequately from the histograms of the previous and next fifteen frames, it was rejected from the candidate collection..
M. Ravinder *et al.* [53]	Used SC region and supervised network	Replay	SVM	Precision, Recall	Required Templates for matching process.
Javed *et al.* [54]	Introduced gradual transition to detect replay frames before detecting the key events	Boundary, Six and Wickets	GMM and ELM	Precision, Recall, F-measure, Accuracy and Error	The replay segments contain number of wide and no ball events. Thus, the computation time has been increased to detect the prime concepts.
Javed *et al.* [55]	Included both SC and GT in replay detection process	Replay	-	Precision, Recall, F-measure, Accuracy and Error	The replay segments contain number of wide and no ball events.
Narasimhan *et al.* [56]	Introduced Genetic algorithm in video summarization	Six, Four, Out and Runs	-	-	Details of classification approach for each event is missing.

Table 3: Summary of reviewed Replay based methods

4.3 Excitement-driven highlight generation

The key event of any cricket video can be identified by an important cue called audio intensity. Some examples of audio cues related to the key events of cricket video are spectator's excitement, appeals and passionate commentaries. The excitement-driven approaches use loudness as an audio feature for the detection of key events. To the best of our knowledge, the literature does not contain any specific paper for excitement-driven highlight generation for cricket video. But, the methods that were designed for other sports could be generally applied for cricket video. Thus, the excitement-driven highlight generation methods developed for all sports video summarization have been reviewed in this section. Tang *et al.* [57] summarize the sports video using the audio features. Baijal *et al.* [11] developed a GMM based technique which used audio cues for summarizing the sports video. The GMM classifier was trained by MFCC features for classifying the excited and non-excited frames in two levels. Here, the speech and non-speech constituents were identified during the first level of the process. The speech constituents were examined in the second level for determining the exciting shots. The important events of cricket video had been identified by Pradeep K [58] using audio cues. Initially, the peak values of audio samples were detected and then the correlation between audio and video frames was determined to extract their respective video frames. Few frames were chosen before peak frames and few from after to construct the event as sequences of frames. Table 4 summarizes the reviewed excitement-driven methods.

4.4 Event-driven cricket highlight generation

The important events of the cricket video were used in event-driven cricket highlight generation. This event-driven summarization method used the domain knowledge of the cricket and hence it is also termed as domain-

Authors	Contributions	Type of classifier	Performance measures	Limitations
Tang *et al.* [57]	Extracted Timbre, Volume and Pitch. A hybrid local threshold and sliding window approach is used for getting the audio peak points	k-means cluster	-	Difficult to set the threshold value because it varied for different kinds of videos.
Baijal *et al.* [11]	MFCC and delta-MFCC features are extracted to identify Commentators' excited speech and spectators excitements	GMM	Precision, Recall	Includes unimportant events due to off-field distractions.
Pradeep K [58]	Based on the peak values of audio signal summarize the video	-	Fidelity	Clear explanation is missing for defining the peak values.

Table 4: Summary of reviewed excitement-driven methods

specific video summarization method [59]. The existing event-driven cricket highlight generation methods not only detected the events but also, they classified them. The OCR, replays and playfield scenarios had been used by the event-driven methods. There are two classifications available in event-driven strategies: Score caption detection based summarization and key event detection based summarization.

4.4.1 Score caption detection based summarization

In the cricket, the scoreboard represents ball-by-ball information about the match. The key events of cricket like boundaries, wickets and sixes can be predicted using these ball-by-ball statistics. Furthermore, the beginning and end of an inning can also be identified with the scoreboard. The scorecard region can be detected efficiently due the rapid changes of scorecard in the cricket game. Though, one can not localize the scoreboard region easily [60]. The name of the countries is usually used as differentiating features for locating the scoreboard region. The typical position of the score bar at the bottom of the screen. Hence, Sayyed et al [61] trimmed the text portion of the image that include the text data. Subsequently, the canny edge recognition process has been employed to identify the edges of texts. The details of the scorecard region analysis are illustrated inFig.10. In Fig.10, the score bar is positioned in the bottom most part of the display. The country name is followed by the displayed score of the presently batting country. A rectangular portion that contains runs-wicket and overs is specifically used to identify the significant event of cricket as shown in Figure 8 (b).

The text region that contained information about the score and players in cricket video was detected by Vijayakumar *et al.* [62]. In this paper, the superimposed text was detected automatically using an efficient method. This method initially determined the key frames with the help of the Color Histogram approach, which could minimize the number of video frames. The key frames were converted into gray images for the detection of text from the key frames efficiently. In general, the bottom region of the image contained superimposed text. Hence, the bottom region that contained the superimposed texts were cropped and the edges of texts were detected using the canny edge detection technique.

Nasir *et al.* [63] used running image averaging for extracting the score caption region of cricket video. Then, an Optical Character Recognition (OCR) approach was used for the recognition of score caption contents. Anjum *et al.* [64] were also used OCR for the generation of cricket highlights. The flow chart for cricket highlight generation using OCR is shown in Fig.9. Here, the events of the cricket games were identified by

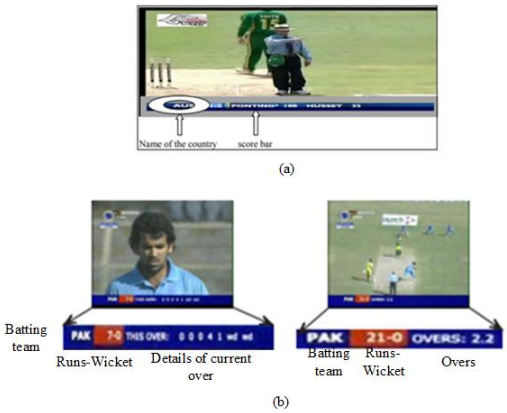

Figure 8: Score bar analysis (a) score bar localization (b) Sample score bar styles in cricket sports video

analyzing the score value and wicket number. The score and wicket values recognized in the previous frame were compared with those of current frame for predicting important events like four, six and wicket. For every key frame, the corresponding video frame collections were chosen for the generation of summarized video.

A five-layer decision tree was used in [65] to process the scorecard region and its layout design is shown in Fig.10. When the active frames contained the SC region, a score separator (SS) was used to separate the score value (SV) from the wicket value (WV). The WV of the present frame was subtracted from that of preceding frame and stored in W counter. Likewise, the SV of the present frame was subtracted from that of preceding frame and was stored in the S counter. The rules for the wicket, boundary and six eventswere given as follows:

$$R = \begin{cases} if(SC = active \wedge SS = WV \wedge W > 0) & ; wicket \\ if(SC = active \wedge SS = SV \wedge 4 \leq S < 6) & ; Boundary \\ if(SC = active \wedge SS = SV \wedge 4 \leq S \geq 6) & ; six \end{cases} \quad (1)$$

where, $W = WV^n - WV^{(n-1)}, S = SV^n - SV^{(n-1)}$
Here, the wicket events were included in the highlight when the value of W counter was greater than zero. Sunitha Abburu [66] proposed a rule-based method for the detection of key events in cricket video. She proposed a hybrid multiple layered method to solve the problems related to automatic semantic segmentation and key events such as wicket, and score detection. The low-level features of the video were analysed by the top layers for extracting the high-level semantics expressed as superimposed text on video. These texts identified the key frames. The bottom layers used the rule based methods for classifying events like wicket and score.

4.4.2 Key event detection based summarization

The important key events of the cricket are replay, boundary view, pitch view, umpire, batsman, bowler, player's gathering and spectators. Fig.**??** shown the hierarchical classification for key event detection. The real frames can be categorized as field and non-field views using the Dominant Green Color Pixel Ratio (DGPR). The crowd and close-up frames are confined into the non-field view. The percentage of edge pixels (PEP) is used to classify the crowd frames as it holds several edge pixels. Fig.12 illustrates the Frames from different Shots that define the boundary view, pitch view, close-up view of bowler, umpire frames, and crowd view of spectators and players gathering.

K. Midhu *et al.* [67] proposed different approaches for generating highlights through the detection of key events. The replay segments were detected by identifying the absence of scorecard region. The Dominant Grass Pixel Ratio was measured to differentiate the Field and Non-field view frames. A trained naïve Bayesian classifier detected the pitch and boundary view frames. A canny edge detector determined the close-up and crowd view frames. From these frames, the bowler, batsman and umpires were classified by analyzing the

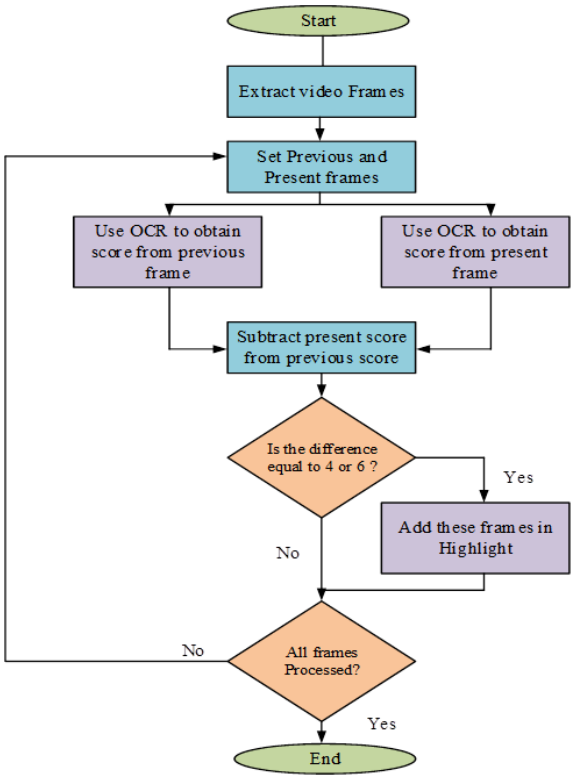

Figure 9: Flow chart for cricket highlight generation using OCR [64]

jersey color.Here, the statistics concerning the location of players face has been used to determine the block that denotes the jersey color as shown in Figure 13. Further, the statistics concerning the location of players' faces could be identified with the help of skin color. The dominance of jersey color differentiated the spectators from players gathering. Finally, an apriori algorithm was used to mine the semantic concepts of cricket. The same approaches were used by Goyani *et al.* [68] to detect all key events from cricket video except replay detection. The replay events were detected by identifying the logo frame, as they were sandwiched between two logo transition frames.

A new approach was introduced by Sandesh Bananki *et al.* [69] for classifying the pitch frames automatically. Here, the non-field frames were initially discarded by calculating the RGB color information. Then, statistical modeling of the grayscale (brightness) histogram (SMoG), component quantization based region of interest extraction (CQRE) and a combination of SMoG and CQRE methods were used to classify the field view frames as non-pitch and pitch frames.

Hetal Chudasama *et al.* [70] extracted certain low-level features for the detection of key events. The extracted features were Grass Pixel Ratio (GPR) for field detection, Pitch Pixel Ratio (PPR) for pitch detection, Total Skin Pixel Ratio (TSPR) for Close Up detection, Edge Pixel Ratio (EPR) for Crowd detection, Total Motion Pixel Ratio (TMPR) for Boundary detection, Close Up Jersey Color for Block1 (CJCB1) for player's team detection, Close Up Jersey Color for Block2 (CJCB2) for player's team detection, Fielders Jersey Color (FJC) for detecting fielders gathering , Sky Pixel Ratio for Sky detection (SkyPR), Prominent Color for Block3 (PCB3) for boundary detection, Prominent Color for Block4 (PCB4) for boundary detection and Saturation Pixel Ratio (StPR) for pitch detection.

Harikrishna *et al.* [71] classified the cricket events by identifying the visual contents in the cricket video. Initially, the boundary shots were determined by calculating the color histogram of adjacent frames. Then the visual features like GPR, PPR, EPR, Moving Edge Pixel Ratio (MEPR), Skin Color Ratio (SCR) and Dominant Hue Value (DHV) were extracted from each shot. The SVM classifier classified the shots as Run,

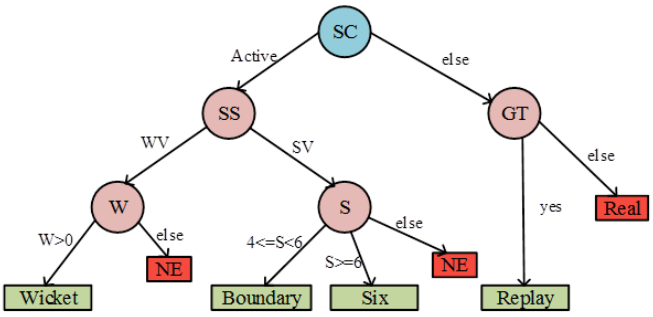

Figure 10: Decision tree-based summarization [65]

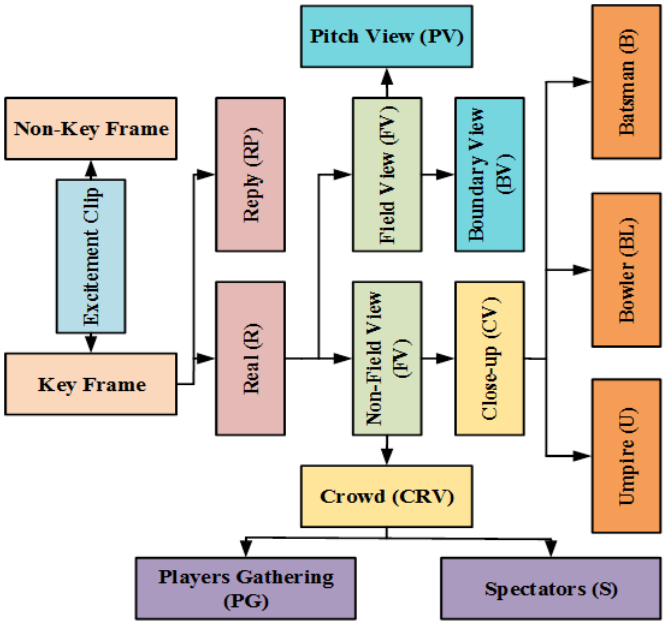

Figure 11: Classification for key event detection in cricket video [67]

Six, Four and Wicket. A sequential minimal optimization algorithm was used to train the SVM classifier. A new Representative Frame Decoration (RFD) model was proposed by Kanade *et al.* [72] for enhancing the key frames. This enhancement was required in the human perception reposed system for increasing the contrast, sharpness and for making the extracted key frames as special. They have used an unsharp mask filter for enhancing the contrast and sharpening component's edges without enhancing noise content. Table 5 summarizes the reviewed event-driven based methods.

4.5 The Excitement and Event-driven key-event detection

Shukla *et al.* [73] proposed an approach for cricket highlight generation by considering both event-based and excitement-based features. Fig.14 shows the block diagram representation of their model. It contains three levels of processes. The video was segmented into different shots at the first level. The significant cues were extracted from every video shot during the second level. These cues were utilized to generate the highlights in the final level. They utilized the audio features to detect the milestones and certain miscellaneous events such as dropped catches. After extracting the excitement clips, the scorecard regions were analyzed to detect important events like boundaries and wickets.

Khan et al [74] generated the highlights through the extraction of both audio and visual features. In this

Authors	Contributions	Detected events	Type of classifier	Performance measures	Limitations
Nasir *et al.* [63]	Running image averaging and OCR method were used	Boundary, Six and Wickets	-	Precision, Recall, Accuracy and error	Detected the score caption when it is presented at fixed location.
Anjum *et al.* [64]	Located the score region with respect to the country name.	Four, Six and Wickets	-	Accuracy	Not suitable for all videos because it considered that the score card region at the bottom of screen.
Sayyed *et al.* [61]	Introduced super imposed approach for score card detection	-	-	Precision, Recall, Accuracy	Detected the score region from all the play and non-play shots this-increased time.
SunithaAb [66]	Proposed a rule based method	Wicket	-	-	Computationally complex.
K. Midhu *et al.* [67]	Associated the key events by extracting the low-level events through deep learning methods	Wicket, Four and Six	Bayesian classifier	Precision, Recall	Required more low-level events in a clip due to the use of a priori algorithm
Goyani *et al.* [68]	Associated the key events by extracting the low-level events	Wickets	-	-	Required more low-level events in a clip due to the use of a priori algorithm
Hetal Chudasama *et al.* [70]]	Extracted low level events for shot classification	Field, boundary, pitch, close up, crowd, Field gathering, sky shots	Multi-layer perceptron multi class classifier	Accuracy	Needs to extract features from each and every frame of cricket video.
Harikrishna *et al.* [71]	Analysed visual contents of the video	Run, Six, Four and Wicket.	SVM	Precision, Recall and F-measure	Needs to extract features from each and every frame of cricket video.

Table 5: Summary of reviewed event-driven based methods

Figure 12: Sample Frames for key events

Figure 13: Close-up view of Batsman, blocks 6 and 10 are selected as skin block and blocks 7 and 11 are chosen as jersey blocks.

approach, a score bar template has been used to locate the score bar region through the learning of SIFT features. This localized region has been further processed for finding the probable text contents. Then, a deep neural network has been used to extract the information from the text contents. This information has been used to select the key events of the cricket video. An additional level of confidence has been provided using the audio-based key frame generation method. Finally, the significant frames were extracted by considering the user preferences.

Javed *et al.* [75] introduced a rule-based induction for extracting the exciting audio clip. Subsequently, a decision tree framework was developed based on scorecard to classify the key events. They improved their work [65] by using Acoustic local binary pattern features for capturing the audio stream's excitement level. Then, a trained SVM classifier was used to classify the audio frames as excited and non-excited ones. M.H.Kolekar *et al.* [16] used STE and ZCR to extract the exciting clip. The frames were considered exciting when the multiplication of short time audio energy with the ZCR surpassed the threshold value. A hierarchical classifier then detected the key events. The concept of the excitement clips was finally extracted by associating the events using the Apriori algorithm.

M.H.Kolekar *et al.* summarized the cricket video by using the caption content analysis instead of the hierarchical classifier [10]. A typical example of wicket fall concepts in cricket was illustrated by Bhawarthi *et al.* [76] using the same hierarchical classifier and Apriori algorithm. Kolekar improved his work by proposing a probabilistic Bayesian belief network (BBN) to index the excitement clips automatically [77]. The concept of the excitement clips was finally labeled with trained BBN. More recently, Shingrakhia et al [78] combined the excitement and event-driven approach using Hybrid Deep Neural Network with the Emperor Penguin Optimization (HDNN-EPO) model. But this approach needed manual thresholds to identify the events from the cricket video. Hence, a hybrid machine learning model has been developed in [78] that summarizes the cricket video without using any manual thresholds. Instead, certain action features and deep learning-based excitement detection approaches have been used. The methods such as OCR, audio and replay discovery were used by Bhalla *et al.* [79] for the extracting key in a cricket match.

The event detection method proposed by Vijayakumar [80] had two steps. Initially, the visual and audio

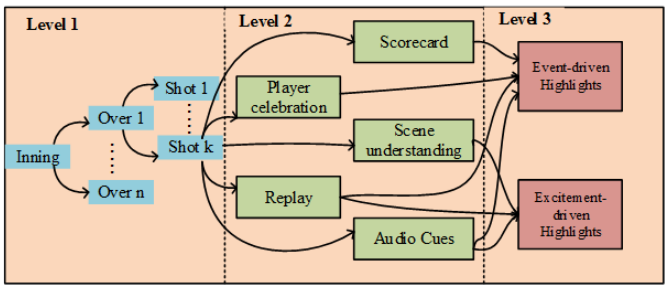

Figure 14: Model for excitement and event-driven key event recognition [73]

features were extracted. Then, a number of heuristic rules were defined to detect the semantic key events like wicket fall, and score events. This method functioned on the basis of experience used to solve, learn and detect the problems. This heuristic approach increased the processing speed providing acceptable results. Kumar *et al.* [81] used a priority curve algorithm to summarize the cricket video. Initially, the video was segmented into several blocks. Then, the visual, audio and textual features were extracted from each block of video. After that, a priority was assigned for each block on the basis of the objects and events that occurred in that blocks. Then, a graph was drawn by considering the block numbers in the x-axis and priorities in the Y-axis. Similar events were identified by determining the peak of the graphs. Then, similar blocks were merged together. Subsequently, the unwanted blocks that had lower priority levels were discarded to generate the summary. Table 6 summarizes the reviewed excitement and event-driven based methods.

4.6 Learning based approaches

The classification process is used in learning based approaches to determine key events. These methods use distinct classifiers like support vector machine (SVM), and Random Forest (RF) classifier for performing different jobs in cricket video summarization. These classifiers are trained to accurately classify the events efficiently. Learning-based techniques can offer better results at the cost of increased computational complexity. In this section, learning based methods are summarized as most of the reviewed methods use Deep Learning. Among all, the SVM classifier has been used in most of the summarization methods for different purposes due to its lower computational complexities. Ravi *et al.* [42] used the SVM classifier for umpire pose detection and M. Ravinder *et al.* [53] to classify the frames as replay and non-replay frames.

Harikrishna *et al.* [71] used it to classify the video shots as four, six and wickets. It has also been used to classify the audio frames as excited and non-excited frames [65]. Instead, Baijal *et al.* [11] trained GMM to determine excited and non-excited audio frames. Shukla *et al.* [73] trained Convolutional Neural Network (CNN) + Support Vector Machine (SVM) to classify the events presented in cricket video. Here, the features have been extracted from the fc7 layer of the pre-trained AlexNet model. This model has been trained on a multi-class linear SVM to predict the events.This multi-class linear SVM has also been used in [83] to classify the Ground view, Pitch view, Fielder view and others. Abbas et al [84] used a multi-class SVM classifier to detect Bowled Out, Caught Behind, Catch Out, and LBW Out events from the cricket video. Furthermore, Javed *et al.* [54] used an ELM classifier to classify the key events. Even though, the computational complexities of the SVM classifier were less, it does not give more representative features. Thus, the accuracy of the classifier might be reduced.

Hari *et al.* [44] used Random Forest (RF) classifier to detect the key events after extracting the umpire frames. But, this classifier requires more time for training because it creates a lot of trees to make decisions. A Naïve Bayesian Classifier has been trained by K. Midhu *et al.* [67] to classify the on-field view of cricket video. Hetal Chudasama *et al.* [70] used a supervised classifier named Multi-layer perceptron multi-class classifier to classify the video shots in cricket by learning the low-level events. In cricket video summarization methods, the semantic meaning of all excited cricket clips should be interpreted (i.e. labeled) based on the low-level events.

Authors	Contributions	Detected events	Type of classifier	Performance measures	Limitations
Shukla *et al.* [73]	Analysed audio cues, score card region and scene understanding to detect events	Boundaries, Wicket, Six and Milestone	CNN+SVM	Insertion of union metric and official highlight comparison	Is a longer process including the unimportant clips not presented in official highlights.
Khan *et al.* [74]	Introduced automatic text extraction process and speech to text model.	Wickets, Four and Six	Deep CNN	Precision, Recall, and F1	The hyperparameter of the deep learning model wasn't selected properly
Javed *et al.* [75]	Introduced rule-based induction method, SC analysis to detect events	Wicket, Boundary, Six and Replay	Decision tree classifier	Precision, Recall, accuracy, error	Caused instability due to decision tree structure
Javed *et al.* [65]	Acoustic local binary pattern for excitement detection	Wicket, Boundary, Six and Replay	SVM,Decision tree classifier	Precision, Recall, accuracy, error	Caused instability due to decision tree structure
M.H.Kolekar *et al.* [16]	Extracted STE and ZCR to detect excitement clip Associated the key events by extracting the low-level events	Wicket, Six and Four	Hierarchical classifier	Percentage selection of the premium concepts (PSPC)	Required more low-level events in a clip due to a priori algorithm.
M.H.Kolekar *et al.* [10]	Analysed caption contents to generate highlights	Wicket, Six, Four and runs	-	Precision, Recall	Affected the event detection performance due to delays in score updating.
Kolekar *et al.* [77]	Semantic concepts were labeled by mapping the low level events to high level concepts through learning process	Wicket and Hits	Bayesian belief network	Precision, Recall	Needed more relevant low level features to improve performance
Bhalla *et al.* [79]	Considered OCR, audio and replay discovery process in summarization process	Six, Wickets, Four	CNN+SVM	Accuracy and official highlight comparison	Accuracy of this method is reduced for the four and six events than wicket event
Vijayakumar *et al.* [80]	Applied number of heuristic rules to detect the events	Wicket fall, scoring event	-	Precision, Recall, F-measure, F-alarm rate	Clear explanation for heuristic rules are missing
Kumar *et al.* [81]	Introduced Priority curve algorithm	Four, Six,Out	-	-	They did not guarantee the method based on performance measures
Hansa *et al.* [82]	key events are extracted from exciting a clips these events are linked to the high level semantic concepts	Four, Six,Wicket	Hybrid deep neural network Emperor Penguin optimization (HDNN-EPO)	Precision ,Recall F1-score,accuracy error rate	Some of the events precision can still improve
Hansa *et al.* [78]	Introduced action recognition, speech to text model using deep learning.	Four, Six,Wicket	SGRNN-AM, HRF-DBN	Precision, Recall, F-1 score, Accuracy, Error rate AUC and PSPC	Not suitable to detect unusual events due to the unavailability of massive training data.

Table 6: Summary of reviewed excitement and event-driven based methods

Thus, the low-level features have been linked to the high-level semantic concepts using some machine learning algorithms like the Bayesian belief network (BBN) [5, 77].

Emon et al [85] introduced a Deep Cricket Summarization Network (DCSN) for extracting the significant shots of cricket video automatically. The excellence of the produced highlight extremely depends on the subjective perceptions of users. Yan et al [86] recommended a YOLO v3 and OpenPose networks to get highlights of sports video with greater accuracy. Among them, the YOLO-based method outperforms the OpenPose-based approach. Hence, Guntuboina et al [87] used YOLO to detect the score bar from the cricket video. Javed et al [88] proposed an efficient decision tree structure to classify the shots of cricket video. This approach could be applied for the precise detection of key events in video summarization. Here, a rule-based instruction has been employed for the creation of several rules required to categorize the cricket shots as long, medium, close-up, and out-of-field shots.

Rafiq et al [89] introduced a transfer learning approach to classify the scene of cricket video in the summarization process. Here, a pre-trained AlexNet Convolutional Neural Network (CNN) is utilized that employed fully connected layers in the encoding structure. CNN usually extracts deep semantic features to generates the summarization for any scenes [90]. Muhammad et al [91] introduced deep video event (DVE) for classifying the key frames with the help of Alexnet. Subsequently, a state machine has been used for the detection of interesting events from the keyframes. The finite state machines (FSM) were detected as an appropriate model to detect and classify the scenes and this approach has been used for the detection of the wicket, six, and four events in [92]. Samaraweera et al [93]] examined the available deep learning models for analyzing the possibility of combined network models in the classifier design of five actions including Six, No Ball, Out, Wide, None. This approach used VGG16, ResNet50V2, and MobileNetV2 CNN structures for extracting the features. The classification task has been performed based on umpire postures using SVM and Naïve Bayes. Tabish et al [94] recognized the activities of sports using the CNN model. VGG16, VGG19, ResNet50, and Inception V3 Models have been trained to extract events of clustered cricket videos.

In addition to these highlight generation methods, the learning methods have also been used to detect specific events of cricket. Specifically, Khan *et al.* [95] used deep CNN with both 2D and 3D convolutions to detect the batting shots.In 2-D CNN, the feature map has been extracted by performing convolution using 2D kernels. Instead, 2-D CNN extracts features by performing convolution in spatial and temporal dimensions of the video. Semwal *et al.* [96] used a deep CNN to represent the features of frames while detecting the cricket shots. In general, the training process of CNN layers requires different sets of images to learn rich feature representations. In this work, a fine-tuned Alex-Net model has been used that contains 5 convolution, 3 pooling layers and 2 fully Connected Layers or FC Layers. The fc7 layer of the network extracted the representations. After finding the representation for all the individual frames, they were concatenated to form a single feature vector which was input to the SVM classifier to classify the shots.

Islam *et al.* [97] introduced a CNN model for the identification of cricket bowlers by considering their bowling action through transfer learning. In transfer learning, the output layer of a pre-trained model is replaced with the nodes representing the required number of classifications. A hierarchical shot recognition method is proposed by Khan *et al.* [98] to identify the cricket shots exactly. The main aim of this work is to give an automatic assessment model to players and coaches. This model discovered the appropriateness of machine learning based sensor data analysis methods. It required a simple set up for aiding the growth of cricket batters through automatic recognition of batting behaviour and visualization of the results in standard diagrams that summarized the batting sessions to improve the player's weakness.

McGrath *et al.* [99] detected the fast bowling event with the help of an inertial measurement unit and machine learning. In this work, 17 elite fast bowlers were selected from the training set for training five machine learning models such as RF, linear SVM, polynomial SVM, neural network (NN), and gradient boosting (XGB) through bowling and executing fielding drills. The performance of these machine learning models was validated by training all the bowling stages such as pre-delivery, delivery and post-delivery. Sen et al [80] proposed a hybrid deep-neural network to classify ten cricket batting shots. Here, a CNN model and a gated recurrent unit (GRU) have been used to extract the features and identify the lengthier temporal dependencies respectively.

Yan et al [100] proposed an automated self-supervised approach to detect keyframes in the video. This approach contained a two-stream Convolutional Network and a new automated annotating structure. This model learned the deep and motion features for the detection of exceptional frames. Most of the deep learning methods broadly depend on the training data presented using professional or expert knowledge. However, it is high-priced to obtain, and hence widely accessible Web images can be used for the weak supervision of highlight generation. Kim et al [101] proposed a triplet deep ranking method for highlight generation with the help of Web images.

4.7 Effect of attention mechanism in highlight generation

Recently, attention-based neural networks are becoming quite popular for video highlight generation. This is because of the benefits of adding additional attention along with the conventional network models. In most of the works, the attention mechanism is added with the encoder-decoder framework to obtain the optimal or higher weight valued features. The attention mechanism is effective in enhancing the overall performance of the CNN and other deep neural architectures while learning the fine-grained actions from the videos. To gather more discriminative information from the videos, the temporal dependency has been exhaustively investigated in recent times. For this purpose, the temporal attention mechanisms are utilized within the encoder-decoder architectures. This is more advantageous as it selectively provides weight values to the encoded features based on different times in the videos. The video's attention mechanisms are utilized for action recognition to capture the long-range dependencies. Chang et al. [102] developed a methodology for temporal action proposal generation (TAPG) using an augmented transformer with an adaptive graph network (ATAG). The ATAG model captured both the videos' long-range and local temporal contexts. In this model, the semantic representation of the features is enhanced with the vanilla transformer that uses a self-attention layer to perform multi-head attention. For effective action recognition from the videos, Li et al. [103] introduced an attention neural cell named AttCell that captured the temporal dependencies. A unified Spatio-temporal attention network (STAN) framework is utilized to deal with multiple modalities. The STAN model extracts the feature map from every single modality and pools it with spatial attention to represent each segment. Finally, those representations are concatenated, and the combined representations are fused to the video representation for action recognition. An audio-visual network was designed by Badamdorj et al. [104] to capture the important information from both the audio and visual cues. One of the major components in the model is the presence of a bimodal attention mechanism, which is used to fuse the features and capture the interaction between the visual and audio components in the video. The model finally utilizes the fused form of the features to generate the highlights for the videos. Another similar methodology for action localization is developed by Pramono et al. [105]Based on a hierarchical self-attention network (HISAN). A two-stream CNN architecture is combined with the hierarchical bi-directional self-attention mechanism to capture the long term temporal dependency and the spatial context information for better localization. The inconsistent detection scores occurring due to clutter and occlusions are removed using the sequence rescoring (SR) algorithm. Finally, a fusion mechanism is followed to combine the appearance and motion information and motion saliency to restrict the camera motion effect. With the addition of audio features with the motion and image features, Hori et al. [106]introduced a video recognition framework to recognize the actions from three different modalities. This paper strongly addresses the problems faced by the fusion technique based on naïve concatenation. The ability of the model to discriminate the dynamic relevance of each of the features is limited by such fusion methodologies. Thus, the extracted modalities are fused in this technique using the multimodal attention model. This model selectively utilized the features from all three modalities for every word in the output description. Another model for video summarization has been put forth by Sanabria et al. [107] that grasps the important events from the video. The methodology is formulated with the hierarchical multimodal attention layer to enrich the performance. The framework initially utilized multiple instance learning methods to consider the sequential dependency among the events in the video. Later, the attention model is utilized to capture the importance of every event in action. On the whole, the attention mechanisms are used by most researchers in the world for video summarization and to identify the

most important actions from the given video.

4.8 Research Progression over time

Table 7 summarizes the classification of the existing methods for cricket video summarization. It also shows the progression of research over time. Here, the event detection methods are categorized as learning (L) and non-learning (NL) -based methods. Learning methods used different classifiers for detecting the key events. From 2008-2014, most of the methods used non-learning-based approaches. But learning-based methods present good consequences as compared to the non-learning-based method. Hence, the new techniques have been adopted using learning-based methods since 2015. Furthermore, most of the works considered low-level features including audio, visual, and text features.

But the low-level features should be linked to the high-level semantic concepts such as boundary, wicket fall, and sixes either by using association-based Apriori algorithm or any machine learning algorithms like SVM, CNN, and BBN [101, 108]. The association-based Apriori algorithm required more events in a clip and more training for the determination of the semantic meaning of the particular clip. The computational complexities of the BBN are high and it requires prior knowledge. The single-layer CNN network will not extract more representative features from the events of any clip and therefore it will give a poor performance for semantic concept annotation. These limitations are tackled by introducing deep learning methods. The recent advancement in the video summarization approach is the use of deep/transfer learning methods. From this analysis, one can understand that the deep features and transfer learning approaches play an important role in the cricket video summarization process over the past 5 years. This is due to the fact of good feature representation capability. Also, the deep learning approaches presented competent outcomes in different image processing applications.

Table 7 summarizes the classification of the existing methods for cricket video summarization.

5 Performance measures for cricket highlight generation

The performances of the cricket highlight generation methods are very challenging to estimate. The performance of such approaches can be evaluated based on the shot boundary discovery, shot classification, and key-frame extraction. The unique approach or measures are not available to evaluate the video summarization methods. The following measures are often used to evaluate the performances quantitatively: precision, recall, f-score and accuracy. The classification accuracy, error rate and training time are the measures used to evaluate summarization methods employing neural networks. The proportion of number of correctly labelled events and the total number of perceived events is described as precision and the following expression is used to calculate it:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The proportion of correctly detected events and the actual events in the video is defined as the recall rate and the following expression is used to calculate it

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The trade-off of precision and recall can be measured by F-measure and it provides the harmonic mean of precision and recall. It can be expressed as:

$$\textit{F-measure} = \frac{2 * Precision * recall}{Precision + recall} \tag{4}$$

Existing methods	Types						Feature type			Output type	
	Object driven	Excitement-driven	Event driven	Excitement and event driven	Learning (L) Non-learning (NL)	Replay based (R)	Audio	Visual	Textual	Static	Dynamic
Ravi *et al.* [42]	√				L			√		√	
Hari and Wilscy [44]	√				L			√			√
Chowdhury *et al.* [46]	√				NL			√		√	
Harun-Ur-Rashid *et al.* [47]	√				L			√		√	
Mridul Dixit *et al.* [49]	√				NL			√		√	
Baijal *et al.* [11]		√			L		√				√
Tang *et al.* [57]		√					√				√
Pradeep K [58]		√			NL		√				√
Nasir *et al.* [63]			√		NL				√		√
Anjum *et al.* [66]			√		NL				√		√
Abburu Sunitha [58]			√		NL				√	√	
K. Midhu *et al.* [67]			√		L			√			√
Goyani *et al.* [68]			√		L			√		√	
Hetal Chudasama *et al.* [70]			√		L			√		√	
Harikrishna *et al.* [71]			√		L			√			√
Javed *et al.* [54]					L	√					√
Javed *et al.* [55]					H	√			√		√
Narasimhan *et al.* [56]					N	√		√	√		√
Ravinder, M *et al.* [53]					L	√				√	
Choros *et al.* [52]					NL			√			√
Shukla *et al.* [73]				√	H		√	√	√		√
Javed *et al.* [75]				√	NL		√		√		√
Javed *et al.* [65]				√	H		√	√	√		√
M.H.Kolekar *et al.* [16]				√	L		√	√			√
M.H.Kolekar *et al.* [10]				√	L		√		√		√
Bhawarthi *et al.* [76]				√	L		√	√			√
Vijayakumar [80]				√	L		√	√		√	
Maheshkumar [77]				√	L		√	√			√
Kumar *et al.* [77]				√	L		√	√	√		√
Bhalla *et al.* [79]				√	L				√		√
Khan *et al.* [95]					L			√		√	
Semwal *et al.* [96]					L			√		√	
Islam *et al.* [97]					L			√		√	
Khan *et al.* [98]					L			√			
McGrath *et al.* [99]					L			√		√	

Table 7: Summary of cricket video summarization methods on their types

The percentage of correct classification of frames is termed as accuracy and it is given as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

The ratio of the incorrectly labelled events to the total number of events is termed as error rate. This can be evaluated using the following expression:

$$ErrorRate = \frac{FP + FN}{TP + TN + FP + FN} \tag{6}$$

Here, True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are the four fundamental components in all measures. The correct labeling of positive samples is represented as TP. The correct labeling of negative samples is referred as TN. The incorrect labeling of negative samples as positive is termed as FP. The incorrect labeling of positive samples as negative is termed as FN.

5.1 Dataset description

To the best of our knowledge, there is no standard dataset to check the performance of the cricket highlight generation models. Thus, the reviewed approaches have been validated by collecting different cricket videos. Also, the existing methods have manually found the Ground truth of all the events of cricket video samples presented in the dataset for performance comparison. The dataset descriptions in the reviewed work used for the comparison are illustrated in Table 8

5.2 Evaluation of existing methods for key frame extraction

The performance of the recent methods of cricket video summarization is evaluated in terms of quantitative measures, such as precision, recall and F-measure for the prime concept identification like boundaries, sixes and wickets. Some existing methods focused on specified events, not on all prime events of cricket such as fours, sixes and wickets. For example, Mridul Dixit *et al.* [49] introduced a framework only for extracting the four-events from the cricket video. Thus, the methods that focused on specified events are not considered for comparison. Instead, the methods that focused on all the prime concepts such as fours, sixes and wickets are considered. Table 9 compares the performance of the methods that depended on the key event extraction algorithms based on object-driven, event-driven, replay based method, excitement and event- driven methods for all prime concept extraction.

Hari *et al.* [44] determined the fours, sixes, wickets and wide events by analyzing the umpire gestures. Javed *et al.* [54] used EML classifier to determine the boundary, wicket and six events from the replay frames and included the replay frames in the highlights. Nasir *et al.* [63] and Anjum *et al.* [64] analyzed the score board region without extracting the replay frames to determine the boundary, wicket and six events. K. Midhu *et al.* [67]used a learning approach to detect the boundary, wicket and six events after extracting the low level features. Likewise, the excitement and event-driven methods of Javed *et al.* [65, 75], M.H.Kolekar *et al.* [10, 16, 77]include four, six and wicket events in the highlights by analyzing audio and visual features of cricket. Khan et al [74] used the cumulative keyframe detection method to combine the individual keyframes obtained based on audio and video streams. Shingrakhia et al [78, 82] proposed different classifiers to identify the key events of cricket video by including hyperparameter selection, action recognition, and deep speech-text analyzing modules. All these works assessed the quantity of the highlights with their own dataset containing different samples of cricket videos, collected from different broadcasters as specified in Section 5.1. Also, these methods are validated by counting the number of correctly and incorrectly detected prime events in the highlights. Table 9 compares the existing methods based on quantitative, but not qualitative, measures.

This table, indicates that the quantitative performance of the existing methods for key event extraction is good enough, except those proposed by Anjum *et al.* [64] and Hari Krishna [73]. The average performance

References	Non-standard dataset details				
	Format	Frame rate	Resolution	Number of videos	Detected events
Hari *et al.* [44]	-	30fps	1280×720	3	Fours, Sixes, Wickets and Wide
Javed *et al.* [54]	AVI	25fps	640×480	20	Fours, Sixes, Wickets
Javed *et al.* [55]	AVI	25fps	640×480	22	Replay
Nasir *et al.* [63]	AVI	25fps	640×480	30	Four, Wicket and Six
Anjum *et al.* [64]	-	25fps	-	3	Four and Six
K.Midhu *et al.* [67]	-	30fps	-	3	Wicket fall, Four and Six
Javed *et al.* [75]	-	25fps	640×480	20	Four, Wicket, Six and replay
Javed *et al.* [65]	AVI	25fps	640×480	20	Four, Wicket, Six and replay
M.H.Kolekar *et al.* [16]	-	10fps	-	4	Hit, Wicket
M.H.Kolekar *et al.* [10]	-	10fps	-	2	Hit, Wicket
Kolekar *et al.* [77]	-	10fps	-	4	Hit, Wicket
Khan *et al.* [74]	-	23-30fps	-	11	Four, Wicket and Six
Shingrakhia *et al.* [82]	-	25fps	640 x 480	15	Four, Wicket and Six
Shingrakhia *et al.* [78]	-	25fps	640 x 480	100	Four, Wicket and Six

Table 8: Dataset specifications

Methods		Precision (%)	Recall (%)	F-measure (%)
Object-driven	Hari *et al.*, [44]	82.33	84.67	83.48
replay-based	Javed *et al.* [54]	96.36	95.27	95.81
	Javed *et al.* [55]	99.77	98.24	98.99
Event-driven	Nasir *et al.* [63]	93.99	87.01	90.36
	Anjum *et al.* [64]	68.46	75.42	71.77
	K. Midhu *et al.* [67]	88.01	87.91	87.95
Exitement and event-driven	Javed *et al.* [75]	91.87	89.85	90.84
	Javed *et al.* [65]	92.94	91.04	91.98
	M.H.Kolekar *et al.* [16]	85.23	88.43	86.80
	M.H.Kolekar *et al.* [10]	82.05	85.49	83.73
	Mahesh *et al.* [77]	88.16	94.66	91.29
	Khan *et al.* [74]	81.32	78.41	79.84
	Shingrakhia *et al.* [82]	93.43	92.46	91.64
	Shingrakhia *et al.* [78]	96.82	95.41	95.67

Table 9: Comparison of the existing methods based on quantitative measures

of excitement and event-driven methods was high compared to even-driven and object-driven methods. Even though the performance of the replay based methods was high for prime concept extraction, they included some unimportant activities. This can be understood by the qualitative measures.

Hari, R., and M. Wilscy [44] introduced the object-driven approach focused on umpire signals and gestures to highlight important events. This methodology is not feasible as the umpire signals are not always captured in the camera during some important events. The replay-based methods Javed et al. [54, 55] utilized the replay segments between the start and end of the gradual transitions to generate the highlights. These two methods cannot generate the highlights efficiently for the videos without gradual transitions. The event-driven method introduced by Nasir et al. [63] captured the keyframes based on the key events detected. Another similar approach by Anjum et al. [64] utilized character recognition techniques to extract the important event information. The approach focused on detecting the events such as fours, sixes and wickets to produce the match highlights. The reviewed excitement and event-driven approaches, such as those introduced by Javed et al. [65], initially captured the excitement in the audio clips. Later, a classification algorithm labels the features into excited and non-excited. The methodologies put forth by Maheshkumar H. Kolekar and Somnath Sengupta [**?**] extract the event sequence from the videos and index the excitement clips based on the extracted features.

5.3 Qualitative Analysis

The standard qualitative metrics are not available for evaluating the performance of summarisation structures. The qualitative measure is performed by considering the response of customers to a questionnaire, which includes key subjective characteristics like informativeness, representativeness, variety, curtness, exposure, semantic principle, enjoyability, understandability and so on [109]. But, most of the reviewed work did not assess the quality of the produced highlights except some. Shukla *et al.* [73] selected twenty cricket fans, in the age group of 20-50 years, to qualitatively analyze the generated highlights. These selected fans rated the generated clips in the range of 0 - 5. The high-quality clips were rated 5. If the clips were rated in the range of 2-5, they were considered as of medium quality. Also, the official highlights are selected as a baseline for quality analysis in both the works of Shukla *et al.* [73] and Bhalla *et al.* [79]

M.H.Kolekar *et al.* [16] analysed the quality of generated highlights by comparing it with that of highlights

generated by the sports channel in terms of the percentage selection of the premium concepts (PSPC).

$$PSPC_s = \frac{\text{Number of prime concepts chosen by sports channel}}{\text{Actual no. of prime concepts available in video}} \times 100 \tag{7}$$

$$PSPC_s = \frac{\text{Number of prime concepts chosen by proposed method}}{\text{Actual no. of prime concepts available in video}} \times 100 \tag{8}$$

5.3.1 Computational Analysis

The computational analysis is a very challenging measure in highlight generation applications because it needs to process a large volume of video frames. In these, several operations need to be performed on each frame to detect the key events. Thus, the computational complexity can be formulated based on the processing time required to analyze one frame. In general, the computational complexity can be written as $O(R \times C)$ Here, R represents computation cost rate of one frame, C represents total number of frames taken for analysis and O represents Big-O Notation. It shows that the computational complexity is increased with the frame rate. It is high due to massive video content in cricket video.

The systems proposed by Nasir *et al.* [63] and Anjum *et al.* [64] analyzed the score-caption using OCR to detect the prime events. These methods require 1.2 seconds to process one frame. It indicates that, the computation cost of these methods will increase with each additional frame, at the rate of 1.2 seconds per frame. The computational time is increased because they analyzed all the frames in the input video. But, Javed *et al.* [54, 55] extracted the replay frames before analysing the score-caption region, reducing the computation time significantly with the inclusion of replay detection. This could be understood by considering a video of 1000 frames with replay segments consisting of 650 frames, only these 650 frames need to be processed. Other replay based methods reduce the number of frames for analyzing the score caption. The object-driven based method, proposed by Hari, R., and M. Wilscy [44], analyzed only the umpire frames to detect the key events. Thus, the computational cost is automatically reduced in object and replay based methods compared to score-caption based methods.

A multimodal excitement and event driven method proposed by Javed *et al.* [65] takes 1.52 seconds to process a single frame for the identification of key events. Thus, a video of 1000 frames with exciting segments in only 40 key-frames needs processing of only 40 frames instead of 1000. The excitement and event-driven methods mostly include learning process. Theoretically, the computational complexity of the learning process is high compared to the non-learning process. But, the learning methods reduce the search space by measuring the audio energy. The number of frames is reduced in such methods maintaining low computational cost.

6 Challenges and future recommendations

The main aim of this review is to identify the challenges in the existing methods and provide the future research scope. The challenges can be addressed in future research. Some challenges of the existing cricket video summarization are illustrated in Figure 15.

The challenges recognized in the existing work for cricket video summarization are as follows:

1. Cricket videos are usually captured in an unrestrained atmosphere and illumination factors (daylight and artificial light). As a result of these, the visual and textual substances can't be analyzed efficiently. Especially, the illumination changes expose substantial variation in the visual features of cricket video. Hence, the effectiveness of the key-events discovery for highlight generations will be automatically reduced. The highlight generation approaches usually analyze the valued information of the score bar region. However, changes in the illumination factors affect the localization process of the score bar region. Figure 15 (a-b) presents the snapshots of sports videos recorded in daylight and artificial lights.

Figure 15: Challenges of cricket video summarization (a) Snap of cricket video taped in artificial light (b) Snap of cricket video taped in day light (c) umpire and payers wear same color jersey (d) Replay frame representing no ball event (e) Overlapped text on score bar (f) Rare event: Disputes between umpire and players (g) Rare event: Injury (h) Rare event: Animal invasion

2. The object driven methods generated highlights by identifying the pose of the umpire or position of the ball. These events can't be detected when the umpires wear same color of jersey as players, as these methods rely only on umpire frames for key event detection. In test matches, both the umpires and players are usually wearing the white color jersey as shown in Figure 15 (c). Hence, these approaches are not suitable for all kinds of cricket matches. The method based on ball detection technique considered the ball as white. These algorithms couldn't localise the ball accurately if the ball was not white.

3. Replay based summarization methods generate the highlights by collecting all the replay segments. However, the replay segments contain not only the exiting moments but also some unimportant activities like wide ball, no ball, etc as shown in Figure 15 (d). However, most of the viewers wish to watch the most interesting events such as Four, Six, and wickets. Hence, these approaches do not consider the viewership interest in highlight generation.

4. In cricket video, the text content, like an advertisement, might overlap the scorecard region, and thus the score caption detection-based method may fail to detect the key events. This problem has been shown through the sample frame in Figure 15 (e). The actual event of this frame is belonging to 'FOUR'. But the score caption detection-based highlight generation approach missed localizing the score bar region due to the presence of certain text content. Hence, the corresponding event would not be inserted into highlights.

5. Excitement-driven methods analyse audio energy to detect the excited events. But, certain off-field distractions caused by the commentators and spectators cheering occasionally might be included in the highlights. Thus, the excitement-driven methods could embrace certain false alarms. The audio energy isn't the only feature that chooses the excitement, hence an additional level of confidence should be added.

6. More semantic contents were utilized by the event-driven methods than the excitement-driven methods. Each frame of cricket video is analyzed in event-driven based highlight generation techniques to detect the key events, increasing the processing time for long-duration videos.

7. The robustness lacked in the majority of the existing cricket highlight generation methods, and some methods were not focused on all aspects of the game.

8. The classification accuracy of the non-learning methods for highlight generation is low. When the observational information is not properly determined, the non-learning approach is not able to determine

the key events due to their static characteristics. To address these issues, the deep learning-based method needs to be adopted for cricket highlight generation. But the accuracy of the deep learning approaches is lacked due to the use of improper hyperparameters. Specifically, the weight parameter of the deep networks affects its efficiency because of its straight proportionality with the objective function.

9. All the existing approaches considered the primary concepts of cricket video such as boundaries, sixes, and wickets. But none of the approaches considered the interesting rare events such as Animal invasion, injuries, and the disputes between the umpire and players. These unexpected actions are occasional.

In the future, the issues of all the existing video summarization methods will be tackled by introducing new deep learning methods. Specifically, the issue of excitement-driven methods will be avoided by considering the game semantics on automated cricket highlight generation methods. In this scheme, an additional level of confidence will be introduced by considering an efficient speech-to-text framework in the deep learning model. Also, the proposed model will focus on both the excitement and event-driven based highlight generation methods to reduce the computational time. Furthermore, the optimization algorithms will be included in the summarization model to resolve the problems related to the learning process of existing classifiers. To tackle the issues related to both the score caption and umpire gesture methods, a new hybrid model will be developed that can identify the events effectively when one of the approaches failed to provide accurate features to identify the key events. To improve the performance of the video summarization model, a deep learning-based action recognizer will be introduced that recognizes the reactions of the umpire and players during all events. Also, a Zero-Shot Learning (ZSL) approach will be developed for including the unusual events in the highlights. It will detect unexpected events without the necessity of labeled data.

Although there are several techniques to resolve the problem of cricket highlight generation in recent times, all those methods have their own merits and demerits. One research challenge is to exploit the usage of high-performance computing and advanced machine learning models for highlight generation. Very few techniques in literature have utilized GPU computing, and therefore, it is necessary and interesting to exploit the potential of GPU for cricket highlight generation. Also, the results obtained from such experiments can be analyzed with the existing results on low-performance computing technologies. It will also be interesting to add more events from matches in the highlights apart from just highlighting the fours, sixes and wickets. These additional events include player's entry, player injuries, pre-and post-match ceremonies, etc. Also, there is a scope for building some reliable models that perform smoothening, commentary cuts and scene transitions on their own to reduce the editors' burden.

7 Conclusion

This paper provides a complete survey of the cricket video highlight generation from various perspectives. The existing cricket video summarization methods utilized various factors like objects, events, excitements and so on. They were developed on the basis of either learning- or non-learning-based methods. We have recognized certain challenges on state-of-the-art methods and provided future recommendations in this domain. The performance comparison of all existing approaches proved that, the learning based excitement and event-driven cricket video summarization methods provided better results than others in terms of classification accuracy. This work successfully identified the needs of a learning based excitement and event driven method for video summarization problem. Even though the results of every existing work on learning based excitement and event driven methods are good enough, more research in this domain is still needed to extract semantically meaningful frames with the greatest accuracy. It should characterize the video optimally without losing any exciting moments. In our future work, we plan to increase the accuracy of the highlight generation system using optimization algorithms.

References

[1] A. Tejero-de-Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, "Summarization of user-generated sports video by using deep action recognition features," *IEEE Transactions on Multimedia*, vol. 20, pp. 2000–2011, Aug 2018.

[2] M. Ramsaran, A. Pooransingh, and A. Singh, "Automated highlight generation from cricket broadcast video," *2016 8th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 251–255, 2016.

[3] D. Ringis and A. Pooransingh, "Automated highlight generation from cricket broadcasts using orb," *2015 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pp. 58–63, 2015.

[4] H. Tang, V. Kwatra, M. E. Sargin, and U. Gargi, "Detecting highlights in sports videos: Cricket as a test case," *2011 IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2011.

[5] M. H. Kolekar and S. Sengupta, "Bayesian network-based customized highlight generation for broadcast soccer videos," *IEEE Transactions on Broadcasting*, vol. 61, pp. 195–209, June 2015.

[6] M. Merler, K. C. Mac, D. Joshi, Q. Nguyen, S. Hammer, J. Kent, J. Xiong, M. N. Do, J. R. Smith, and R. S. Feris, "Automatic curation of sports highlights using multimodal excitement features," *IEEE Transactions on Multimedia*, vol. 21, pp. 1147–1160, May 2019.

[7] R. Kapela, K. McGuinness, and N. O'Connor, "Real-time field sports scene classification using colour and frequency space decompositions," *Journal of Real-Time Image Processing*, vol. 13, 06 2014.

[8] N. Homayounfar, S. Fidler, and R. Urtasun, "Sports field localization via deep structured models," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4012–4020, 2017.

[9] R. Hannane, A. Elboushaki, K. Afdel, P. Nagabhushan, and M. Javed, "An efficient method for video shot boundary detection and keyframe extraction using sift-point distribution histogram," *International Journal of Multimedia Information Retrieval*, vol. 5, pp. 89–104, 2016.

[10] M. H. Kolekar and S. Sengupta, "Caption content analysis based automated cricket highlight generation.," pp. 461–465, 2008.

[11] A. Baijal, Jaeyoun Cho, Woojung Lee, and Byeong-Seob Ko, "Sports highlights generation bas ed on acoustic events detection: A rugby case study," in *2015 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 20–23, Jan 2015.

[12] M.-H. Sigari, H. Soltanian-Zadeh, and H. Pourreza, "Fast highlight detection and scoring for broadcast soccer video summarization using on-demand feature extraction and fuzzy inference," *International Journal of Computer Graphics*, vol. 6, pp. 13–36, 01 2015.

[13] A. Javed, *EVENT DRIVEN VIDEO SUMMARIZATION FOR SPORTS*. PhD thesis, University of Engineering and Technology Taxila, Pakistan, 2016.

[14] J. Xing and X. Li, "Feature extraction algorithm of audio and video based on clustering in sports video analysis," *Journal of Visual Communication and Image Representation*, p. 102694, 11 2019.

[15] H.-c. Shih, "A survey on content-aware video analysis for sports," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, pp. 1–1, 01 2017.

[16] M. Kolekar and S. Sengupta, "Semantic concept mining in cricket videos for automated highlight generation," *Multimedia Tools Appl.*, vol. 47, pp. 545–579, 05 2010.

[17] V. Scotti, L. Sbattella, and R. Tedesco, "Sferanet: automatic generation of football highlights," 11 2019.

[18] J. B. Li, W. Dai, F. Metze, S. Qu, and S. Das, "A comparison of deep learning methods for environmental sound detection," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 126–130, 2017.

[19] Y. Li, X. Zhang, H. Jin, X. Li, Q. Wang, Q. He, and Q. Huang, "Using multi-stream hierarchical deep neural network to extract deep audio feature for acoustic event detection," *Multimedia Tools and Applications*, vol. 77, 01 2017.

[20] G. Ramya and S. Kulkarni, "Visual saliency based video summarization: A case study for preview video generation," in *Information, Photonics and Communication*, pp. 155–165, Springer, 2020.

[21] S. Kamoji, "Key frame extraction for video summarization using motion activity descriptors," *International Journal of Research in Engineering and Technology*, vol. 03, pp. 491–495, 03 2014.

[22] D. Karmaker, A. M. Chowdhury, M. S. U. Miah, M. Imran, and M. H. Rahman, "Cricket shot classification using motion vector," *2015 Second International Conference on Computing Technology and Information Management (ICCTIM)*, pp. 125–129, 2015.

[23] R. Roopchand, A. Pooransingh, and A. Singh, "Bat detection and tracking toward batsman stroke recognition," *2016 8th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 256–260, 2016.

[24] M. Z. Khan, S. Jabeen, S. ul Hassan, M. Hassan, and M. U. G. Khan, "Video summarization using cnn and bidirectional lstm by utilizing scene boundary detection," in *2019 International Conference on Applied and Engineering Mathematics (ICAEM)*, pp. 197–202, IEEE, 2019.

[25] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Video summarization using deep semantic features," *ArXiv*, vol. abs/1609.08758, 2016.

[26] S. Chakraborty and D. Thounaojam, "A novel shot boundary detection system using hybrid optimization technique," *Applied Intelligence*, vol. 49, 03 2019.

[27] J. Majumdar, M. Aniketh, B. R. Abhishek, and N. Hegde, "Video shot detection in transform domain," *2017 2nd International Conference for Convergence in Technology (I2CT)*, pp. 161–168, 2017.

[28] S. Premaratne, K. Jayaratne, and P. Sellappan, "A novel hybrid adaptive filter to improve video keyframe clustering to support event resolution in cricket videos,"

[29] J. Majumdar, M. Awale, and K. L. K. Santhosh, "Video shot detection based on sift features and video summarization using expectation-maximization," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1033–1037, Sep. 2018.

[30] R. Minhas, A. Javed, A. Irtaza, M. Mahmood, and Y. Joo, "Shot classification of field sports videos using alexnet convolutional neural network," *Applied Sciences*, vol. 9, p. 483, 01 2019.

[31] S. C. Premaratne and K. L. Jayaratne, "Structural approach for event resolution in cricket videos," in *ICVIP 2017*, 2017.

[32] V. B. T. Jay A. Patel, "An improvised algorithm for automatic shot segmentation and summarization," *International Research Journal of Engineering and Technology*, vol. 3, p. 975, 03 2016.

[33] Z. Wei, B. Wang, M. Hoai, J. Zhang, X. Shen, Z. Lin, R. Mech, and D. Samaras, "Sequence-to-segments networks for detecting segments in videos," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[34] L. Hamrouni, M. L. Kherfi, O. Aiadi, and A. Benbelghit, "Plant leaves recognition based on a hierarchical one-class learning scheme with convolutional auto-encoder and siamese neural network," *Symmetry*, vol. 13, no. 9, p. 1705, 2021.

[35] B. Khaldi, O. Aiadi, and M. L. Kherfi, "Combining colour and grey-level co-occurrence matrix features: a comparative study," *IET Image Processing*, vol. 13, no. 9, pp. 1401–1410, 2019.

[36] S. Angadi and V. Naik, "Dynamic summarization of video using minimum edge weight matching in bipartite graphs," *International Journal of Image, Graphics and Signal Processing*, vol. 8, pp. 9–18, 03 2016.

[37] K. Kumar, D. D. Shrimankar, and N. Singh, "Equal partition based clustering approach for event summarization in videos," *2016 12th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, pp. 119–126, 2016.

[38] J. Majumdar, S. Udandakar, and M. B G, *Implementation of Cure Clustering Algorithm for Video Summarization and Healthcare Applications in Big Data*, pp. 553–564. 01 2019.

[39] R. Ranjan and A. Agrawal, "Video summary based on f-sift, tamura textural and middle level semantic feature," *Procedia Computer Science*, vol. 89, pp. 870–876, 12 2016.

[40] S. J. A. Dipali M. Balas, "Video summarization using cnn and clustering algorithm," *Journal of Applied Science and Computations*, vol. 5, pp. 1249–1253, 10 2018.

[41] M. B. P. Kumar and P. Puttaswamy, "The extraction of events and replay in cricket video," *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, pp. 54–58, 2015.

[42] A. Ravi, H. Venugopal, S. Paul, and H. R. Tizhoosh, "A dataset and preliminary results for umpire pose detection using svm classification of deep features," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1396–1402, Nov 2018.

[43] K. Choros, "Highlights extraction in sports videos based on automatic posture and gesture recognition," in *ACIIDS*, 2017.

[44] R. Hari and M. Wilscy, "Event detection in cricket videos using intensity projection profile of umpire gestures," *11th IEEE India Conference: Emerging Trends and Innovation in Technology, INDICON 2014*, 02 2015.

[45] S. Nandyal and S. L. Kattimani, "Vision based umpire detection method for event extraction in cricket video based on hog and color image segmentation," in *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pp. 1–6, IEEE, 2021.

[46] A. Z. M. E. Chowdhury, M. S. Rahim, and M. A. U. Rahman, "Application of computer vision in cricket: Foot overstep no-ball detection," *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pp. 1–5, 2016.

[47] M. Harun-Ur-Rashid, S. Khatun, M. Z. Trisha, N. Neehal, and M. Z. Hasan, "Crick-net: A convolutional neural network based classification approach for detecting waist high no balls in cricket," *ArXiv*, vol. abs/1805.05974, 2018.

[48] V. Bhagavathiappan and P. Kumar, "An efficient ball detection framework for cricket," *International Journal of Computer Science Issues*, vol. 7, 05 2010.

[49] M. Dixit and C. Bhatnagar, "A novel approach to detect fours in cricket videos," in *2014 International Conference on Computer and Communication Technology (ICCCT)*, pp. 307–311, Sep. 2014.

[50] S. Nandyal and S. L. Kattimani, "An efficient umpire key frame segmentation in cricket video using hog and svm," in *2021 6th International Conference for Convergence in Technology (I2CT)*, pp. 1–7, IEEE, 2021.

[51] C.-M. Chen and L.-H. Chen, "A novel method for slow motion replay detection in broadcast basketball video," *Multimedia Tools and Applications*, vol. 74, 06 2014.

[52] K. Choros and A. Gogol, "Improved method of detecting replay logo in sports videos based on contrast feature and histogram difference," in *ICCCI*, vol. 9875, 09 2016.

[53] T. V. G. M. Ravinder, "Replay frames classification in a cricket video using correlation features and svm," *European Journal of Applied Sciences*, pp. 92–97, 2015.

[54] A. Javed, A. Irtaza, Y. Khaliq, H. Malik, and M. Mahmood, "Replay and key-events detection for sports video summarization using confined elliptical local ternary patterns and extreme learning machine," *Applied Intelligence*, 02 2019.

[55] A. Javed, K. Bajwa, H. Malik, and A. Irtaza, "An efficient framework for automatic highlights generation from sports videos," *IEEE Signal Processing Letters*, vol. 23, pp. 1–1, 07 2016.

[56] H. Narasimhan, S. Satheesh, and D. Sriram, "Automatic summarization of cricket video events using genetic algorithm," in *GECCO '10*, pp. 2051–2054, 01 2010.

[57] S. Tang and M. Zhi, "Summary generation method based on audio feature," *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp. 619–623, 2015.

[58] P. K, "Significant event detection in sports video using audio cues," *International Journal of Innovations in Engineering and Technology*, vol. 3, pp. 144–151, 10 2013.

[59] V. Kaushal, S. Subramanian, S. Kothawade, R. Iyer, and G. Ramakrishnan, "A framework towards domain specific video summarization," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 666–675, IEEE, 2019.

[60] A. A. Khan, H. Lin, S. Tumrani, Z. Wang, and J. Shao, "Detection and localization of scorebox in long duration broadcast sports videos," in *International Symposium on Artificial Intelligence and Robotics 2020*, vol. 11574, p. 115740J, International Society for Optics and Photonics, 2020.

[61] T. Sayyed, D. Barai, and S. Kande, "Super imposed method for text extraction in a sports video," *International Journal on Future Revolution in Computer Science & Communication Engineering*, vol. 4, no. 3, pp. 332–337, 2018.

[62] V. Velusamy and N. R, "A novel method for super imposed text extraction in a sports video," *International Journal of Computer Applications*, vol. 15, 02 2011.

[63] M. Nasir, A. Javed, A. Irtaza, H. Malik, and M. Mahmood, "Event detection and summarization of cricket videos," *Journal of Image and Graphics*, vol. 6, pp. 27–32, 01 2018.

[64] M. E. Anjum, S. F. Ali, M. T. Hassan, and M. S. K. Adnan, "Video summarization: Sports highlights generation," *INMIC*, pp. 142–147, 2013.

[65] A. Javed, A. Irtaza, H. Malik, M. Mahmood, and S. Adnan, "A multimodal framework based on audio-visual features for summarization of cricket videos," *IET Image Processing*, vol. 13, 01 2019.

[66] S. Abburu, "Semantic segmentation and event detection in sports video using rule based approach," *International Journal of Computer Science and Network Security*, 10 2010.

[67] K. Midhu and N. A. Padmanabhan, "Highlight generation of cricket match using deep learning," in *Computational Vision and Bio Inspired Computing*, pp. 925–936, Springer, 2018.

[68] M. M. Goyani, S. K. Dutta, and P. Raj, "Key frame detection based semantic event detection and classification using heirarchical approach for cricket sport video indexing," in *International Conference on Computer Science and Information Technology*, pp. 388–397, Springer, 2011.

[69] S. Jayanth and G. Srinivasa, "Automated classification of cricket pitch frames in cricket video," *Electronic Letters on Computer Vision and Image Analysis*, vol. 13, pp. 33–49, 07 2014.

[70] H. Chudasama and N. Patel, "A unified framework for cricket video shot classification using low level features," *Indian Journal of Science and Technology*, vol. 10, pp. 1–6, 12 2017.

[71] N. Harikrishna, S. Satheesh, S. Sriram, and K. Easwarakumar, "Temporal classification of events in cricket videos," *National Conference on Communications (NCC)*, 01 2011.

[72] S. S. Kanade and P. Patil, "Representative frame decoration using unsharp filter in video summarization," in *2011 International Conference on Communications and Signal Processing*, pp. 570–573, IEEE, 2011.

[73] P. Shukla, H. Sadana, A. Bansal, D. Verma, C. Elmadjian, B. Raman, and M. Turk, "Automatic cricket highlight generation using event-driven and excitement-based features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1800–1808, 06 2018.

[74] A. A. Khan, J. Shao, W. Ali, and S. Tumrani, "Content-aware summarization of broadcast sports videos: An audio–visual feature extraction approach," *Neural Processing Letters*, vol. 52, no. 3, pp. 1945–1968, 2020.

[75] A. Javed, K. B. Bajwa, H. Malik, A. Irtaza, and M. T. Mahmood, "A hybrid approach for summarization of cricket videos," in *2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pp. 1–4, IEEE, 2016.

[76] D. Bhawarthi and P. S. Gadage, "Enriching feature extraction using a-priory algorithm for cricket video," 2012.

[77] M. Kolekar, "Bayesian belief network based broadcast sports video indexing," *Multimedia Tools Appl.*, vol. 54, pp. 27–54, 08 2011.

[78] H. Shingrakhia and H. Patel, "Sgrnn-am and hrf-dbn: a hybrid machine learning model for cricket video summarization," *The Visual Computer*, pp. 1–17, 2021.

[79] A. S. Bhalla, A. Ahuja, P. Pant, and A. Mittal, "A multimodal approach for automatic cricket video summarization," *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 146–150, 2019.

[80] V. Vijayakumar, "Event detection in cricket video based on visual and acoustic features," *Journal of Global Research in Computer Science*, vol. 3, no. 8, pp. 26–29, 2012.

[81] K. S. Kumar, S. Prasad, S. Banwral, and V. B. Semwal, "Sports video summarization using priority curve algorithm," *International Journal on Computer Science & Engineering*, vol. 2, no. 9, pp. 2996–3002, 2010.

[82] H. Shingrakhia and H. Patel, "Emperor penguin optimized event recognition and summarization for cricket highlight generation," *Multimedia Systems*, vol. 26, no. 6, pp. 745–759, 2020.

[83] A. Kumar, J. Garg, and A. Mukerjee, "Cricket activity detection," *International Image Processing, Applications and Systems Conference, IPAS 2014*, 02 2015.

[84] Q. Abbas and Y. Li, "Cricket video events recognition using hog, lbp and multi-class svm," in *Journal of Physics: Conference Series*, vol. 1732, p. 012036, IOP Publishing, 2021.

[85] S. H. Emon, A. Annur, A. H. Xian, K. M. Sultana, and S. M. Shahriar, "Automatic video summarization from cricket videos using deep learning," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–6, IEEE, 2020.

[86] C. Yan, X. Li, and G. Li, "A new action recognition framework for video highlights summarization in sporting events," *arXiv preprint arXiv:2012.00253*, 2020.

[87] C. Guntuboina, A. Porwal, P. Jain, and H. Shingrakhia, "Deep learning based automated sports video summarization using yolo," *Electronic Letters on Computer Vision and Image Analysis*, vol. 20, no. 1, pp. 99–116, 2021.

[88] A. Javed, K. M. Malik, A. Irtaza, and H. Malik, "A decision tree framework for shot classification of field sports videos," *The Journal of Supercomputing*, pp. 1–26, 2020.

[89] M. Rafiq, G. Rafiq, R. Agyeman, G. S. Choi, and S.-I. Jin, "Scene classification for sports video summarization using transfer learning," *Sensors*, vol. 20, no. 6, p. 1702, 2020.

[90] V. Tiwari and C. Bhatnagar, "A survey of recent work on video summarization: approaches and techniques," *Multimedia Tools and Applications*, pp. 1–35, 2021.

[91] R. Muhammad, R. Agyeman, H.-K. Shin, R. Ali, K.-M. Kim, and G.-S. Choi, "Deep video events (dve): A deep learning approach for sports video summarization,"

[92] V. Ellappan and R. Rajkumar, "Classification of cricket videos using finite state machines," *International Journal of Information Technology and Management*, vol. 20, no. 1-2, pp. 83–94, 2021.

[93] W. Samaraweera, S. Premaratne, and A. Dharmaratne, "Deep learning for classification of cricket umpire postures," in *International Conference on Neural Information Processing*, pp. 563–570, Springer, 2020.

[94] M. Tabish, M. Shaheen, *et al.*, "Activity recognition framework in sports videos," *Multimedia Tools and Applications*, pp. 1–23, 2021.

[95] M. Z. Khan, M. A. Hassan, A. Farooq, and M. U. G. Khan, "Deep cnn based data-driven recognition of cricket batting shots," in *2018 International Conference on Applied and Engineering Mathematics (ICAEM)*, pp. 67–71, IEEE, 2018.

[96] A. Semwal, D. Mishra, V. Raj, J. Sharma, and A. Mittal, "Cricket shot detection from videos," in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6, IEEE, 2018.

[97] M. N. Al Islam, T. B. Hassan, and S. K. Khan, "A cnn-based approach to classify cricket bowlers based on their bowling actions," in *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pp. 130–134, IEEE, 2019.

[98] A. Khan, J. Nicholson, and T. Ploetz, "Activity recognition for quality assessment of batting shots in cricket using a hierarchical representation," *PACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)*, vol. 1, pp. 62:1–62:31, 09 2017.

[99] J. McGrath, J. Neville, T. Stewart, and J. Cronin, "Cricket fast bowling detection in a training setting using an inertial measurement unit and machine learning," *Journal of Sports Sciences*, pp. 1–7, 12 2018.

[100] X. Yan, S. Z. Gilani, M. Feng, L. Zhang, H. Qin, and A. Mian, "Self-supervised learning to detect key frames in videos," *Sensors*, vol. 20, no. 23, p. 6941, 2020.

[101] H. Kim, T. Mei, H. Byun, and T. Yao, "Exploiting web images for video highlight detection with triplet deep ranking," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2415–2426, 2018.

[102] S. Chang, P. Wang, F. Wang, H. Li, and J. Feng, "Augmented transformer with adaptive graph for temporal action proposal generation," *arXiv preprint arXiv:2103.16024*, 2021.

[103] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 416–428, 2018.

[104] T. Badamdorj, M. Rochan, Y. Wang, and L. Cheng, "Joint visual and audio learning for video highlight detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8127–8137, 2021.

[105] R. R. A. Pramono, Y.-T. Chen, and W.-H. Fang, "Hierarchical self-attention network for action localization in videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 61–70, 2019.

[106] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proceedings of the IEEE international conference on computer vision*, pp. 4193–4202, 2017.

[107] M. Sanabria, F. Precioso, and T. Menguy, "Hierarchical multimodal attention for deep video summarization," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 7977–7984, IEEE, 2021.

[108] M. M. Elgamml, F. S. Abas, and H. A. Goh, "Semantic analysis in soccer videos using support vector machine," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 34, no. 09, p. 2055018, 2020.

[109] M. Sreeja and B. C. Kovoor, "Towards genre-specific frameworks for video summarisation: A survey," *J. Visual Communication and Image Representation*, vol. 62, pp. 340–358, 2019.