2023 IEEE 8th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 7-9, 2023

Cricket Video Summarization Using Deep Learning

Rahul S Bhat, Jayanth O, Pawan Prasad P, Phani Kumar Vedurumudi, Prof. Divyaprabha K N
Computer Science Department, PES University, Bangalore, India
rahulsbhat39@gmail.com, jayanth934180@gmail.com, pawanssj5@gmail.com,
pvphan.i890@gmail.com, divyaprabhamadhu@gmail.com

2023 IEEE 8th International Conference for Convergence in Technology (I2CT) | 979-8-3503-3401-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/I2CT57861.2023.10126359

***Abstract*—Video summarising is the process of choosing and presenting the most relevant or fascinating components from a longer video document in order to create a brief summary for potential consumers. Cricket is a highly regulated sport played for quite a longer duration than most other sports. This study presents a paradigm for recognising and clipping crucial occurrences in a cricket match that takes into account event-based attributes. Cues used to capture such moments include replays, audio intensity, player celebrations, and playfield scenarios. This project focuses on building a model which will summarise the original video into three types of summaries. Different Deep Learning models and Natural Language Processing techniques are used to build the model. Key frame extraction is a crucial task that is done to generate the cricket highlights. The project's purpose is to show how to extract a high-quality summary out of an original video using a high-performance approach. This research presents a robust video summarization model for the cricket videos that is able to generate high-quality cricket highlight videos along with textual and audio summaries based on the user's preferences. The focus is on the automation of video summarization process and provide a better quality of summary to the user. From the comparison of the generated highlights to the actual existing highlights manually, it was concluded that the methodology adopted gave a satisfactory result covering almost 80-90% of the events and captions with Bilingual evaluation understudy score-4 of 0.748953.**

***Index Terms*—Deep Learning, LSTM, RNN, VGG16, ResNet-50, Short time energy, BLEU score**

I. Introduction

Modernization and advancement in technology have generated data drastically as never before, especially videos and images, thus resulting in an abundant number of long videos not only on social media platforms like YouTube [3], but also on sports platforms, especially in long-duration sports like cricket. Sports videos are the most interesting and useful content, with significant economic potential [4]. Because sports recordings are typically extensive, it is not always practical for an offline audience to see the entire recorded video of a given event. Cricket is one of the world's favourite sports [5]. Cricket matches are generally played for a longer duration, with each match ranging from 3 hours to five days [5]. Due to the large number of matches taking place during the year, it is difficult for a cricket fan to watch all the matches and follow all the news. As a result, highlights are a great way for fans to catch up on all of the most significant and recent events [2].

In the current world, most people are attracted towards short-content videos due to their busy schedules. Because of this, interest in the long-duration video has gradually subsided. The increase in the popularity of social media platforms such as Instagram, Tik-Tok, and others is largely due to short content videos such as reels. According to analytical data from recent studies, the viewership of Test match videos is gradually decreasing due to the popularity of T20 matches. However, manually creating highlights takes a long time and even expert editing abilities, which makes it difficult to summarise media in a timely manner. Researchers are currently working on automating video summaries with scene categorization, OCR and frequency levels, sharpness score, boundary scores and ensemble learning along with other scoring schemes and LSTM [3], Reinforcement learning with scoring schemes which include more computations. This project focuses on generating a high-quality video summary from the original video and is an automation of the traditional manual process.

The input match video is processed as audio and higher energy video chunks are merged together to result in a highlight video. Deep Learning Algorithms and NLP techniques and different python libraries are utilised to build a model that generates three types of summaries, i.e., Highlight Video, Textual Summary, and Audio Summary. CNN models VGG16, ResNet-50 and InceptionV3 have been tried and tested for their best fit for the purpose of feature extraction. The better one among them is the ResNet-50 which has been chosen for the job. Standard LSTM model is used in building the model for caption generation. Other methods like key frame extraction, image processing have also been utilised. The novelty of this project is to generate high quality cricket video highlights, textual summary, and audio summary from the original video in either of these three formats.

Rest of the article is organized as follows. The section 2 studies various existing approaches for scene classification, summary generation of various sports and highlights the benefits and limitation of existing methodologies. In section 3, the Dataset, Preprocessing, Approach and Implementation have been discussed. In section 4, the results obtained using proposed approaches is provided. In section 5, the results has been interpreted. In last section, the work is concluded and future re-search direction is highlighted.

II. Literature Survey

[1] *Scene classification for Sports Video Summarization using transfer learning*

This research proposes a new approach for identifying sports video situations and uses cricket as an example and divides

scenarios into many categories like batting, balling, boundary etc. Since scene categorization is such a key part of video summarization, its accuracy is crucial.

A training dataset with five named classes (batting, bowling, crowd, boundary, and close-up) is constructed. The proposed method employs a pre-trained AlexNet convolution neural network (CNN) and three more fully connected layers for classification of scene. Dropout was applied to the first two additional layers, and then SoftMax was used to activate the last layer. Data augmentation was used to boost the performance of the model resulting in an accuracy of 99.26% across a limited dataset.

The results reveal an improvement in performance above existing state-of-the-art approaches. The proposed technique outperforms competing models in terms of performance but it took longer time for the model to acquire higher accuracy and the frames that did not fit into any category were misclassified.

[2] *A Multimodal approach for automatic cricket video summarization*

This paper presents a novel method for detecting and summarising significant occurrences in a cricket match automatically. Goal is to accept a full-length video of a cricket match as an input and return the game's most important segments as an output.

Innings are segmented into video shots which is a collection of frames captured by a single camera. In addition to the recognition of video shots, the important frames of video are also determined considering the audio. Features such as score boards and audio cues are observed. OCR is used to detect the match scores. The complete highlight of the match is then sliced together from these moments.

The comparison of official and system-generated highlights reveals that the suggested model captures practically every occurrence that should be included in a cricket match's highlights was able to detect wickets, fours and sixes with an accuracy of 89.45% but was seen to miss out on some boundaries.

[3] *Semantic Text Summarization of Long Videos*

This paper is about textual summarization of the videos using semantic knowledge of the text. Deep visual-captioning approaches provide textual and visual descriptions from key frames taken from intriguing portions. Extractive methods use captions from interesting portions to build paragraph summaries for the full video. The objective is to generate a textual summary from the videos using the different ML techniques.

The key frames are extracted from the videos using OpenCV. Interesting frame segments were identified using the cinematographic rules and measuring the different measurement scores of the segments like Boundary score, Sharpness score, Contrast score, and Attention score to pick the informative segments from the video. Each frame was passed through a ResNet CNN model and then passed to the LSTM model to generate the captions from the frame. At the end, the NLTK libraries and Python framework are used to generate a proper summary from the captions.

Though this approach performed well, there is a huge amount of performance overhead while processing large videos, as many frames are processed iteratively at each step.

[4] *Automatic Video Summarization from Cricket Videos Using Deep Learning*

This paper focuses on the automated video summarization of cricket videos. To overcome the challenges like complicated rules and long duration matches in cricket and to summarise them, the authors of this paper have developed a Deep Cricket Summarization Network (DCSN) that extracts important-shots from an input video automatically.

Deep learning and Reinforcement Learning are used to build the DCSN model. The LSTM Model and the CNN Model are also used. The key frames are extracted from the videos by using the decoder, which is an agent of Reinforcement learning. The frames are then passed to the CNN model for feature extraction. The features are then visualised and then passed to the LSTM based bi-directional Recurrent Neural network (RNN) with a fully connected neural network. Then the sigmoid function produces a probability, which is used by the Bernoulli function to sample the key frames. The various reward functions are also used to complete this robust model.

The Mean Opinion Score value of the automatically created summary videos was 4 out of 5. But, the usage of multiple bi-directional LSTMs in the model increases the space complexity of the model as each LSTM will store a significant amount of information in the network and the DCSN used had huge performance overhead.

[5] *Automatic Cricket Highlight Generation Using Event-Driven and Excitement-Based Features*

This paper proposed a methodology for recognising and clipping crucial occurrences from a video that considers both event and excitement-based aspects. The goal of the paper is to provide realistic cricket highlights by taking into account both events and enthusiasm factors. For the generation of highlights, the system relies on events as well as excitement aspects.

The four significant events in a cricket match, namely wickets, boundaries, sixes, and milestones, are extracted using event-driven features, while the remaining essential events are recognised using enthusiasm features. A full cricket match was deconstructed into a number of video clips. The absence of a scoreboard in all replays and commercials was used to detect replays. To recognise replays and playtime, a convolutional neural network (CNN) and a support vector machine (SVM) framework were utilised. Scores from the scoreboard were detected using Optical Character Recognition (OCR). The model was able to segment and recognise minor events more consistently by using video shots over video frames.

The model was unable to detect unexpected occurrences that can occur during a match that are important for highlights, such as harsh weather, injuries, or intense player arguments.

This method may take a little longer and may include clips that aren't included in the official highlights.

[6] *Generating Highlights of Cricket Video using Commentators and Spectators Voice*

In this paper, the proposed methodology uses audio intensity to extract key frames to generate video highlights. The objective of this method is to create a synopsis of the video while preserving the semantics of the film. There are two types of video summarising: static summary and dynamic summarization. A dynamic video summary turns a long video into a short one. Manual video summarising is a process that involves inspecting a movie to determine the most interesting and essential frames, which are then used to create a video summary.

The analysts' voices and the spectators' cheers are considered essential elements in creating the video highlights. The audio from the input video has been extracted. After that, the audio was divided into chunks. Short-term energy was computed when the audio was broken down into chunks, and a threshold value for the energy was determined to classify audio snippets as excited or not. The target video's highlights are created by combining all of the exciting segments.

This approach is useful for small size videos which are easily available from YouTube. Only audio of commentators is used for generating the highlights which is not a better approach compare to the video summarization.

[7] *Soccer Video Summarization Using Deep Learning*

This paper focuses on soccer video highlights generation from long-duration videos using deep learning techniques. Keyframe selection, ball tracking, key sub-shot selection, and skims are just a few of the prominent automated video summarising approaches. These strategies have been shown to be effective in summarising films in general, but they do not allow for the selection of the required list of action sets for a summary video. The authors of this project have used the 3D-CNN and LSTM highlight detection frameworks to develop a simple but effective approach for producing summed films, and evaluate the summary methodology using Mean Opinion Scores acquired from 48 soccer fans.

ResNet-based 3D-CNN is used for feature extraction from the video along with the ReLu activation function and annotating soccer videos manually into five soccer activity classifications for training purposes. 3D-CNN and LSTM-RNN to detect best among soccer highlights. MOS (Mean Opinion Score) is used to evaluate the summarised system, and it is usually 4 or 5 MOS for videos. They used the 3D-CNN and LSTM highlight detection frameworks to build a simple yet successful approach for producing summarised videos.

The only ball-related videos that can work well with this model is basketball. The ResNet model requires more processing power to load the model, and it takes a lot of time.

[8] *Video Summarization Using Fully Convolutional Sequence Networks*

This paper focuses on generating the highlights video using the Fully Convolutional Sequence Networks. Video summarising is treated as a keyframe selection problem in this study. The goal is to construct a summary video by selecting a subset of frames from an input video. In this article, the author uses fully convolutional networks for video summarization. In semantic segmentation, fully convolutional networks (FCN) have been widely used. Unlike previous methods that used recurrent models, this one develops a unique relationship between semantic segmentation and video summarising, then adapts popular semantic segmentation networks for video summarization. Extensive testing and analysis on two benchmark datasets have proven that the models perform well.

The GoogleNet Model is used for the feature extraction. Extensive testing and analysis on two benchmark datasets have proven that the models perform well. They have used the F-score to check the effectiveness of the model. In this study they treat video summarising as a sequence labelling task. They have used fully convolutional sequence networks (FCSN) for video summarization over LSTM to attain the parallelism in the model.

III. Proposed Methodology

A. Dataset Selection

The dataset required for this project is not readily available and has been gathered from various sources like YouTube and Hotstar. Full-length videos will be used to generate the video highlights and further processing. Alternatively, the video highlights that have been collected can also be processed to get the output in the desired format, i.e., a textual or audio summary of that video.

Dataset for Highlight generation:

Dataset for this section is a set of full-length match videos. Some videos used for testing in this case have been manually edited to remove the advertisements and the audio files to work on are generated for respective videos.

Dataset for Captioning:

Dataset for this section has been collected manually and is a sequence of frames with caption for each. Each frame is mapped to a maximum of 5 captions and is collected from an online resource and the captions have been manually modified as per the requirement so that they convey the proper sense and to ensure that the file formats are uniform.

Dataset size: 4168

B. Pre-processing

- Some videos used for testing in this case were manually edited to remove the advertisements and the audio files to work on are generated for respective videos.

- The frames from which the features are to be extracted are preprocessed into images of desired dimensions for them to be suitable to be fed to the models.
- Captions collected as a text file were processed into a dictionary with key as the file name and the values are a set of captions (max 5) and is used in this format for further process. Also, the NLP techniques of stemming and lemmatization were used to process the captions and the unique words from the captions are made into a vocabulary.

C. Assumptions

- The process of highlight video generation is dependent on Noise levels and assumes that there are no other noisy events happening in the crowd apart from the crowd cheers for the shots hit or for the wickets.
- This project assumes that original videos used do not contain any advertisement and shot replays.

D. Approach

- Key frames are extracted from the video to generate highlight video. This process uses the Energy level of the sounds in the given video.
- The generated highlight video from the above process is fed as an input to the feature extraction model.
- Features are extracted from the frames in the video using CNN model.
- Feature vectors are then given to the LSTM network which will map them to the words through the embedding layer.
- The Captions are then formatted as Textual summaries for each ball.
- The Textual Summaries are then converted to audio.

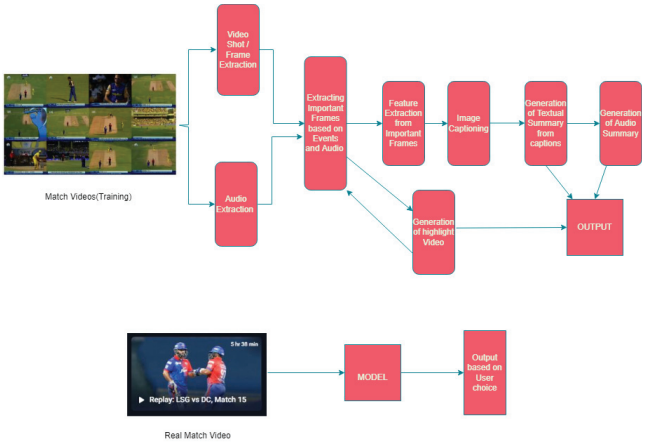

Fig.1: Architecture

E. Implementation

The Project majorly consist of three tasks:

1) Highlight generation of a Cricket Video:

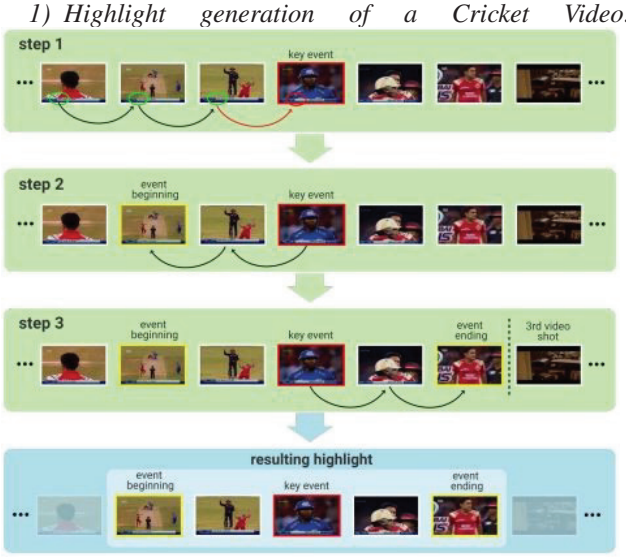

Fig.2: Highlight Generation

- In this task, the librosa module is used to extract the energy values from the audio of the cricket video

Bowler Type	Approximate Time for the ball to travel to the batsman in sec
The average professional fast bowler (137kph or 85mph)	0.52
The average professional spinner (80kph or 50mph)	0.9
The fastest ball ever recorded (161.3kph or 100.2mph)	0.45
Average club cricket fast bowler (113kph or 70mph)	0.64

Fig.3: Time approximation

- Audio is extracted from the Video using moviepy module. The audio is then split into chucks of 5sec each assuming that a shot will be played in under 5 sec.
- Short time energy for those chunks is calculated using the formula,
  Short time energy =

$$\sum a^2$$

where 'a' is an array of frequency of audio generated for every 5 sec chunks

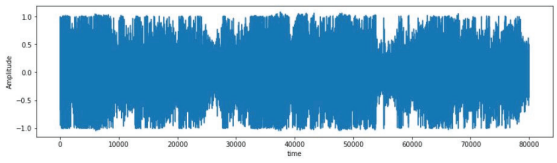

Fig.4: Short time energy graph

- Those chunks with energy value greater than the mean/threshold of energies of all the chunks are considered as excitement clip and are merged into a single video which results in a highlight video.
- The excitement clips are used as a key factor to generate the highlights of cricket video.
- The audience noise and the commentator noise are high when there is a boundary shots and wickets.
- The audience noise and the commentator noise are classified as excitement clips.
- By merging all the excitement clips a highlight of match is generated.

2) Generation of relevant Captions:

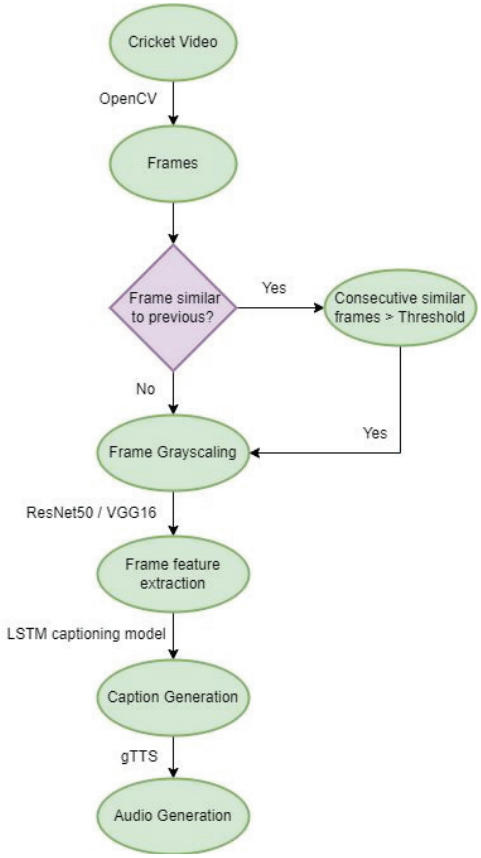

Fig.5: Flow chart of captioning system

- Streamlit library is used for UI and cricket video is taken as input from user and stored in a temporary location for captioning.
- Frames are generated using OpenCV and similar frames are considered as a single frame.
- Imagehash is used to check for similar frames generated. Similar frames are removed to ensure memory efficiency.
- Then cricket video frames are passed to the captioning system.
- Captioning system is a model trained on a train data collected. Feature extractors trained were VGG16, ResNet-50 and InceptionV3 and the captioning model trained was standard LSTM:
  - ResNet-50 trained for 12 epochs was used for the task of feature extraction as it returned the most relevant captions among the models trained for the purpose i.e. VGG16 and InceptionV3 which were trained for 12 and 50 epochs respectively. Output of ResNet-50 is a feature vector of shape 1 X 2048

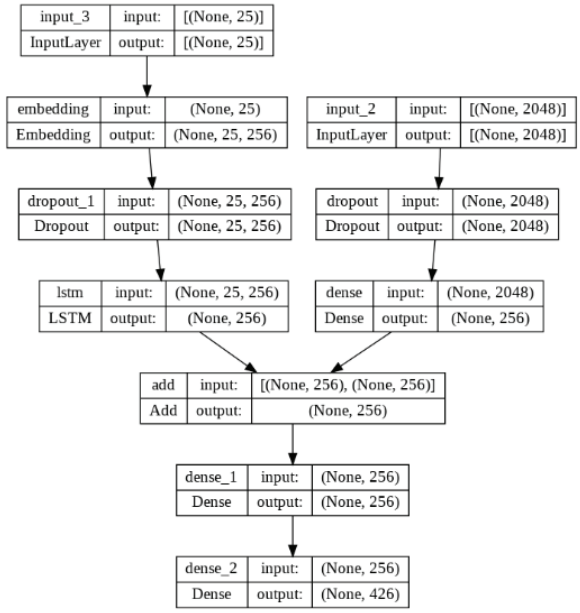

Fig.6: Captioning model architecture

  - LSTM then maps feature vectors to relevant words through the embedding layer. [Fig 6]
- Captioning of frames occurs in real time with the playing video.
- Similar frames being already removed results in unique captions which in itself is a summary for that particular delivery.
- Model can generate real time captions for an input video of size up to 200MB efficiently.

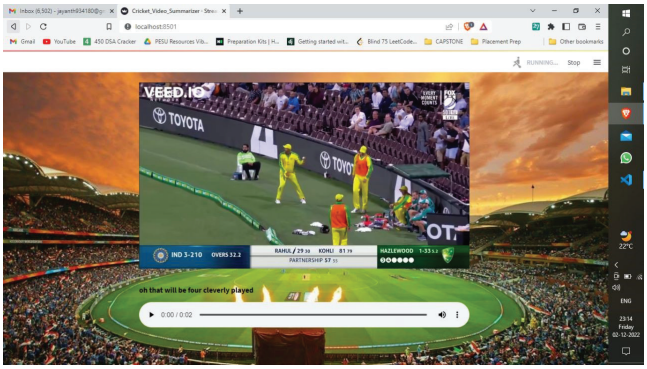

Fig.7: UI

3) Generating Audio: The captions generated are then also outputted as audio using the Google Text to Speech (gTTS) API which goes concurrently with the caption generation.

IV. Results

- From the comparison of the generated highlights to the actual existing highlights which is done, it was concluded that the methodology adopted gave a satisfactory result covering almost 80-90%.
- The errors in event detection were a result of existence of advertisements, replays and other unexpected incidents occurring on field. With a video that was edited to remove all advertisements manually was seen to give coverage of 80-90%.
- Videos of average length 20 mins were seen to result in a highlight of length 5 to 8 minutes on average.
- The length of the highlight video is subject to the number of key events occurring during that course of match. Lesser the events, shorter the highlights.

MODELS	BLEU-1	BLEU-2	BLEU-3	BLEU-4
VGG16	0.881131	0.819212	0.792201	0.728526
InceptionV3	0.699529	0.573926	0.532131	0.440226
ResNet50	0.893821	0.832808	0.807842	0.748953

Fig.8: Models Result

- Different Feature extractors trained were VGG16, ResNet-50 and InceptionV3 along with the LSTM Model which were also tested on the basis of BLEU score and relevance.
- VGG16 trained for 12 Epochs and showed BLEU-4 score of 0.728526. InceptionV3 trained for 50 Epochs beyond which they were seen to overfit and was seen to show BLEU-4 score of 0.440226 but the captions generated were meaningful in some cases.
- Out of the three, ResNet-50 trained for 12 Epochs was identified to be the best performing which gave most relevant captions with a BLEU-4 score of 0.748953.

V. Discussion

Highlight generation process under consideration assumes that there are no replays and advertisements in the input video. This assumption might works well in real time on-ground match scenarios. The time required for highlight generation and the length of the generated video are dependent on the number of important events recorded in that video. The errors in event detection were a result of existence of advertisements, replays and other unexpected incidents occurring on field. The processes involved forms a pipeline and are dependent on the previous tasks. As a result, any deviation from the expected output in the early stages will also have an impact on the final output. Working of all this tasks requires huge RAM and GPU sources to process large videos which is still a challenge.

VI. Future work and scope

The outputs of this proposed methodology under discussion are the highlights of the given cricket video, the ball by ball textual summary and the audio version of the summary generated. The textual summaries generated are of the generalized format and it does not consider any team or person names. This can be overcome by preparing a captions dataset for each individual player along with a model to identify the players. A dataset consisting of images of individual players on the ground from different angles can be prepared and can be used for the purpose of player identification. Players can be identified during the captioning stage from the same frames that will be used for captioning. This can further be extended to automated evaluation of player performance which can make use of the summaries generated and players identified during the captioning process by associating the number of boundaries and the wickets taken to each player's name as a basis to evaluate player's performance.

References

[1] Rafiq, Muhammad, et al. "Scene Classification for Sports Video Summarization Using Transfer Learning." Sensors, vol. 20, no. 6, Mar. 2020, p. 1702. Crossref, https://doi.org/10.3390/s20061702.

[2] Bhalla, Aman, et al. "A multimodal approach for automatic cricket video summarization." 2019 6th international conference on signal processing and integrated networks (SPIN). IEEE, 2019.

[3] Sah, Shagan, et al. "Semantic text summarization of long videos." 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017.

[4] Emon, Solayman Hossain, et al. "Automatic Video Summarization from Cricket Videos Using Deep Learning." 2020 23rd International Conference on Computer and Information Technology (ICCIT). IEEE, 2020.

[5] Shukla, Pushkar, et al. "Automatic cricket highlight generation using event-driven and excitement-based features." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018.

[6] Mohammed, Inayathulla, et al. "Generating Highlights of Cricket Video using Commentators and Spectators Voice." IRJTE, vol. 8 , issue. 4 , Nov. 2019.

[7] Agyeman, Rockson, Rafiq Muhammad, and Gyu Sang Choi. "Soccer video summarization using deep learning." 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2019.

[8] Rochan, Mrigank, Linwei Ye, and Yang Wang. "Video summarization using fully convolutional sequence networks." Proceedings of the European conference on computer vision (ECCV). 2018.