Automatic Cricket Highlight Generation Using Event-Driven and Excitement Based Features Using Deep Learning

A Ph.D. Synopsis submitted to

Gujarat Technological University

For the Award of

Doctor of Philosophy

In

Electronics & Communication Engineering

By

Shingrakhia Hansben Jesabhai

Enrollment No.:179999915020

Under Supervision of

Dr. Hetal N. Patel

Professor & Head

Electronics & Communication Engineering Department

A.D. Patel Institute of Technology

V.V. Nagar, Anand, GJ, India

March 2022

INDEX

1 Title of Thesis and Abstract

1.1 Title of Thesis

AUTOMATIC CRICKET HIGHLIGHT GENERATION USING EVENT-DRIVEN AND EXCITEMENT BASED FEATURES USING DEEP LEARNING

1.2 Abstract

Video summarization plays a crucial role in diverse domains where the major application is sports video summarization. The sports broadcasting channels show interest in the summarization of sports videos based on the viewer's interest as they hold massive viewership all around the world. To gain transmission benefits and to reduce the amount of storage, they show interest in extracting the exciting clips from the lengthy cricket video. Sports video summarization is a tiring task as it includes huge variations in the camera movements, background noise, lighting conditions, editing effects, etc. As a solution, this thesis work presents reliable hybrid methods to effectively identify the key events from the video for highlight generation. The computational complexity of this strenuous task has been reduced by proposing key event recognition systems that detected and classified only the important events from the video. This step reduced the overall duration of the video making it suitable for summarization.

In the first research work, a hybrid deep neural network with emperor penguin optimization (HDNN-EPO) is proposed to generate cricket video highlights. In this work, the key events of the cricket video are identified and then summarization is done for the obtained key events. Initially, the exciting clips are extracted using audio features such as shouting, spectators cheering, and applause of the audience. Then the key frames are generated by determining the shot boundaries in the videos and these frames are identified using the hue histogram differences of neighborhood frames. The key events such as replay, players gathering, real view, umpire, batsman, fielder, spectators, and field view are then extracted to determine the importance of each clip

in summarization. After this process, the concept annotation process is carried out using the proposed HDNN-EPO algorithm for the obtained exciting clips. The EPO algorithm reduced the weight values of the DNN to reduce the error rate in the annotation process. A voting classifier is used in the model to find and label the classes with the highest votes as concepts. Finally, video summarization is carried out by determining the importance degree of each mega slot. All the mega slot highlights are concatenated to generate the cricket video highlights.

In the second research work, a hybrid machine learning framework has been proposed for the summarization of cricket videos. Initially, using the audio stream in the videos, the exciting clips are extracted. A speech-to-text framework is introduced to detect the excitement clips based on the stacked gated recurrent neural network with an attention module (SGRNN-AM). The hue histogram differences between the frames are computed to determine the shot boundaries. After this step, the keyframes from the videos are extracted using a manual threshold value that is in contrast with the hue histogram difference. Then, the shots present in each exciting clip are classified using the proposed hybrid rotation forest deep belief network (HRF-DBN). The rotation forest ensemble approach is used in the model to improve the accuracy rate of the DBN classifier. This model classified the cricket video into close-up, medium, long, and out-of-field/crowd shots. The score-card region in the video is located using the temporal running image averaging algorithm. A sequence of features is then obtained from the score-card region for summarization. The umpire gestures for the key events such as four, six, and wickets are determined and the action features are gathered. Both these features are then passed to the SGRNN-AM module for video summarization. Experimental evaluations establish the effectiveness of the proposed hybrid methods. Also, the evaluations suggest that the models can be applied to summarize any kind of lengthy sports videos to enable transmission over low-bandwidth networks.

2 A brief description on the state of the art of the research

The rapid explosion of devices and the extensive internet connectivity in the world has led to the outbreak of data like never before. This data covers textual data as well as streaming contents that are made accessible to all the users connected to the internet [1]. Among the multimedia types, streaming content gains a lot of audiences as it is more expressive and some of the examples include sports videos, CCTV footages, user videos, TV videos, etc. Based on the analysis, it is identified that sports videos are watched by most people around the world [2]. It is due to this reason that the broadcasters generate a massive collection of sports videos and transmit it through the network as it delivers commercial benefits. But the problem with this generation is that the videos are lengthy and demand more storage spaces as well as higher bandwidth requirements. Though there is a tremendous advancement in technological growth to deal with multimedia content, the exponential increase of these videos should not be overlooked [3-5].

Live sports telecasted on TV and the internet are exciting but, the time duration of this telecast is high. The busy schedules of the people in modern days make it difficult to watch these lengthy videos that are only telecasted at a particular time. Therefore, most people prefer watching the highlights of these videos that are as short as possible [6, 7]. Full-length live videos comprise audio, video, and textual features that are synchronized in a proper way to deliver useful content. Video summarization is an approach to generate the highlights by evaluating the full-length videos and selecting the most useful content based on the viewer's interest [8]. This task is a tiring one as it requires a complete evaluation of the videos that are affected by several lighting and editing effects, camera orientations, background noise, etc. Among the sports events, cricket is the most popular one that is watched by several million people throughout the world. But the fact is, there are not many works concentrated on the highlight generation of cricket videos as these videos are of longer duration with more time complexities [9-12].

Cricket video summarization techniques help the broadcasters to gain useful content from the lengthy videos that match the user's interest. The live actions are monitored and the key events are identified so that the summarized version of the video can be more informative [13]. Since the manual analysis and highlight generation tasks required huge efforts and are error-prone, automated techniques are formulated to accurately generate the highlights. This thesis work presents solutions to generate informative highlights for the cricket videos through a thorough analysis of the events changes involved in the videos [14]. All the three modalities such as audio, visual, and textual features are considered to develop the framework [15]. The excitement clips in the cricket videos are highly useful to identify the interesting parts and this is differentiated by the audio energy in the video. Apart from this, the key events in the videos are determined and matched with the higher-level concepts to generate the required highlights [16, 17].

3 Problem Definition

Highlight generation for cricket videos is affected by different factors present in the cricket videos. The most common factors include the illumination conditions in the video, background noise, editing effects, game structure, etc. The key frames in the cricket video are identified by determining the exciting clips based on the variations in the frequencies of audio energy. This is done by identifying the variations of frequencies in the audio energy extracted from the cricket video. The problem in determining the audio energy is the presence of background noise which is one of the most common factor in any cricket video. Shot boundary detection step helps to identify the shot boundaries in a video that are of prime importance and can be added along with the highlights. Shot boundaries can be defined as the boundaries and transitions between consecutive shots in a cricket video. Detecting the shot boundaries include different limitations where the most important one is the orientations of the camera. The shots present on every clip may provide useful information and an ideal technique should be capable of detecting all the shots of

every excitement clips. Localization of score-card region is also important as it provides important information about the game. The transitions in the videos affect the accurate detection or localization of score-board region. The key frames in the cricket video are needed to be selected in a way that provides all the required information to generate effective highlights. For selecting the key frames, the importance of each key frame is determined and then appropriate decision is taken to form the summary. Owing to different influential factors in the cricket video like illumination effects, transition effects, gaming complexities, etc. it is difficult to accurately determine the key events of the videos. To address all the above-mentioned limitations, this thesis work proposes effective hybrid frameworks for cricket video summarization.

4 Objective and Scope of work

The main aim of this research is to explore techniques for cricket video summarization through exciting clip annotation, replay detection, key frame detection and shot boundary detection using the excitement and event-driven features. Video summarization tool is mostly used to determine the crisp representation of lengthy videos by determining the key frames. The objective of the thesis is to design and implement robust techniques that are not influenced by the editing effects, illumination conditions (i.e. difference in daylight and artificial light), background noise (i.e. audience noise), computational cost, etc.

Exciting clips in a cricket video determine the important events and it can be identified by the audio energy resulting from the spectators' cheers and commentaries. By extracting the excitement clips, the key event recognition process can be boosted by reducing the computational complexity involved in processing a full-length video. This step divides the lengthy video into smaller excitement clips and the key event recognition can be done only on the excitement clips. These exciting clips are further used in the process of concept annotation to generate the cricket highlights.

The Shot boundary is another method of processing the cricket videos by partitioning the videos into different shots that reduces the complexity involved in processing. These shots are then classified into long, medium, close-up and out of field/crowd shots. Shot classification helps to bridge the semantic gap between the high and low-level events involved in the videos that can be used for video summarization. Sample images for shot boundary is displayed in Figure 4.

Figure 4: Sample images for shot boundary

As mentioned above, the objective is to present automatic video summarization approaches that can deliver the key events of a cricket video as highlights by reducing the duration of the video. Thus, the ultimate aim is to detect the key events occurring in a cricket video through the implementation of an effective key event detection framework. The major challenge involved in the key event detection process is the length of the video as the cricket video may last for up to 40 hours covering an approximation of 3.6 million frames. Therefore, this thesis investigates the ways to determine the keyframes of the cricket videos irrespective of the duration of the matches. To identify the keyframes, the audio, visual and textual features are determined and are used in combination to model effective techniques for cricket video summarization.

The main objectives of this thesis work are to develop, implement and evaluate the following research topics:

- ❖ Effective key event recognition and highlight generation scheme to generate highlights for cricket videos.
- ❖ An efficient shot classification and key event detection method to generate summaries for lengthy cricket videos with reduced computational complexities. The size and number of parameters in the model is reduced to minimize the overall complexity.

5 Original contribution by the thesis

The main contribution of this thesis work is as follows:

5.1 Event recognition and highlight generation

- ❖ An optimized hybrid model is proposed to automatically label the exciting concepts present in the cricket video based on the key events observed.
- ❖ Improving the efficiency of the cricket highlight generation process by combining both the event-driven and excitement-driven features.
- ❖ The hue histogram difference is computed between the neighboring frames to determine the key events from every exciting clip of the cricket video. This step reduces the overall computational time complexities.
- ❖ The proposed model is robust to any changes in camera orientations, illumination conditions, lighting and editing effects, video duration, replay speed, game structure, broadcasters, etc.

5.2 Shot classification and cricket video summarization

- ❖ A hybrid machine-learning model is proposed to automatically summarize the cricket video irrespective of the complexities and rules involved in the game.
- ❖ This approach combines event-driven, excitement-driven, and object-driven features to detect the key events for effective summarization.
- ❖ The excitement present in the audio energy is detected with an additional level of confidence by introducing a “speech to text” framework. This framework

transcript the speech or audio energy into textual cues for better performance this helped the model to gather more useful excitement clips.

- A novel hybrid shot classification model is built to classify the scenes present in every exciting clip.
- A two-stream networking structure-based classifier is proposed to generate the required summary by considering the action features extracted from the umpire gestures and score-board regions.
- The training of the proposed model is improved by providing more meaningful data that reduced the computational complexities involved in summarization and results in improved overall accuracy.

6 Methodology of research, results / comparisons

6.1 HDNN EPO: The summarization of extended cricket video is to a short-lived form is termed as a cricket highlight generation. While summarizing, it maintains the presence of essential moments in the original video. In this research, a novel method (combination of optimization and neural network) is developed to recognize the key events and also for video summarization. Initially, the excitement clips are extracted with the aid of audio features. After extracting excitement clips, the essential events such as replay, players, umpires, spectators, and players gathering are extracted. After the process of extraction, for concept labeling in the cricket video, an HDNN-EPO is proposed. Then the selection of labeled concepts is performed based on the importance degree (i.e. the degree or rank set by the user based on their priority). The efficacy of the developed method is analyzed via the simulation results and also it is compared with existing approaches to highlight the proposed work. Performance such as precision, recall, and accuracy is evaluated. The framework of the proposed method is explained in the following.

The cricket video summarization and detection of key events are done spontaneously with the aid of a highly developed optimized neural algorithm. The diagrammatic

representation of the developed event recognition and video summarization is illustrated in Figure 6.1.

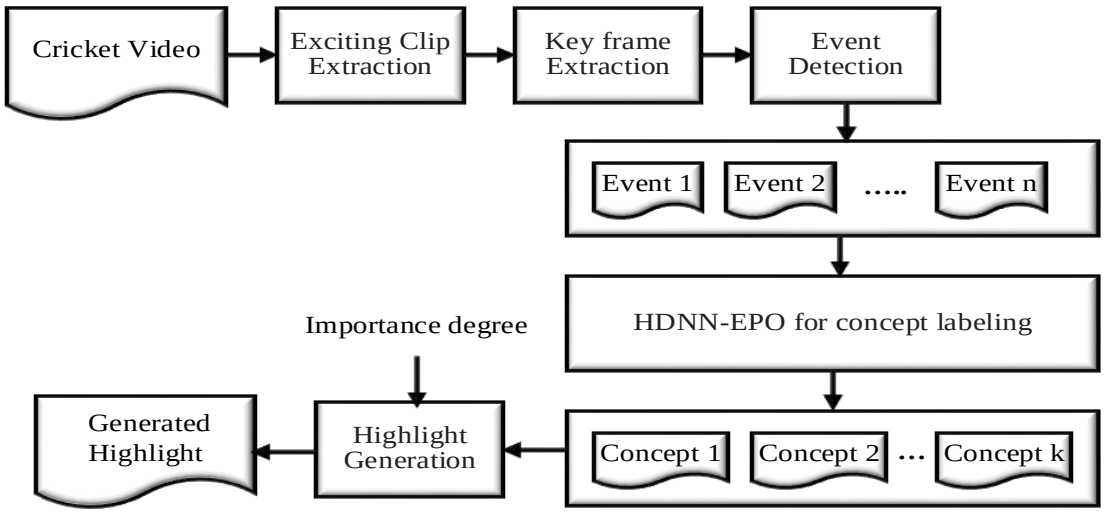

Figure 6.1: Schematic diagram of the proposed methodology

Initially, the captured cricket video is inclined as an input for the developed method. Extractions of audio features are extracted for the clip extraction from cricket video. With the aid of extracted audio features, the exciting clips from the cricket video are extracted. The emotional moments or consequences of cricket can be constituted by the utilization of spectators' cheering, shouting, and applause. An example to represent the audio energy for speech is displayed in Figure 6.2.

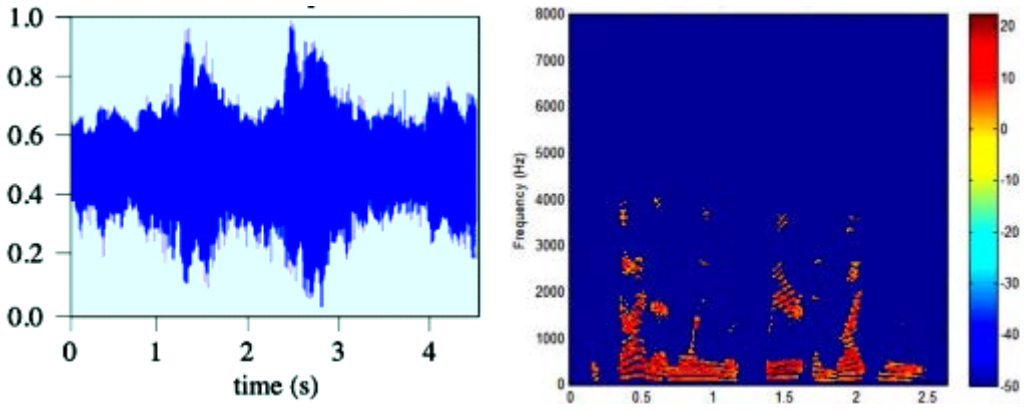

Figure 6.2: Audio energy signal for speech and its frequency range

From the exciting clips, the keyframes or short boundaries are detected by examining the deviation of the Hue histogram among the neighboring frames. By this extraction approach, the complexity of computational time gets diminished. By positioning the scorecard region, from the real frames the replay frames can be adapted. To establish the pitch view, boundary view, batsman, fielders, umpires, players gathering, and spectators the low-level features such as color and edge are extracted. The reaction of cricket players can be detected with the aid of action recognizers (i.e. it is a framework used to recognize the shots being played by the player or to determine their actions) during wickets, boundaries, injuries, and disputes/arguments between umpires and players. And for the label concept, the hybrid optimized neural network is developed. For a peculiar exciting clip, the HDNN-EPO is introduced. Finally, the highlight generation is emphasized to ordering the nominated exciting clips in the temporal order.

To provide the cricket video through the online network by machine learning-based video summarization, a hybrid machine learning approach is introduced**.** The flow of the proposed method is illustrated in figure 6.1. Initially, the cricket videos are inputted into the system then the audio streaming is extracted from the video streaming. Each of the exciting events from the video clip consists of some of the events like replay, close-up view of player; zoom in of referee, players gathering, and spectators. The inputted cricket videos are pre-processed by the operations of commercial and replay removal. In which the commercial is extracted by the cut density based on color and motion features. In commercial extraction, the audio and video cuts from the cricket video are identified after that boundaries are refined by adding the audio and video information simultaneously. The audio and video streams are extracted from the inputted video. From the audio stream, the speech to text is recognized using the ‘speech-to-text’ framework. Some pre-processing approaches such as transformation of color frames to gray-scale images and down-sampling. The commercial shots are continuous and appear in groups thus some kinds of techniques are adopted for the extraction of commercial sequences. The replay events are

examined by checking whether the scorecard is presented or not. Besides this, the real match may be telecasting previously played match highlight. As in this case, the highlight segments have high playback rate and frame transition rate as normal segments. Thus the playback rate and frame transition rate of each segment in inputted cricket video are recognized for removing the previously highlighted match from the current cricket video.

6.2 SGRNN-AM: In this proposed method, the key events detection process has a four-stage framework, in the first stage the key frameworks are examined by exciting frameworks through an audio stream of cricket video. Here the process is enhanced by the addition of speech to text recognition frameworks along with a stacked gated recurrent neural network with an attention module (SGRNN-AM) for the identification of excitement clips. Then the shots from each clip are classified by the proposed hybrid rotation forest deep belief network (HRF-DBN). Here the accuracy of DBN is enhanced by the rotation forest ensembles approach. The subsequent modules analyze the location of the scorecard of a particular video and additionally extract the action features of the umpire frame for observing the key events. Finally, the characters and features are extracted from the location of the scorecard, and the umpire frames are given as an input to SRGNN-AM for including the important activities from the cricket vides like four, six, and wicket in the highlight. The accuracy of SGRNN-AM is improved by adopting the gated recurrent units (GRU). A brief description of each module of the proposed method is given below Figure 6.3.

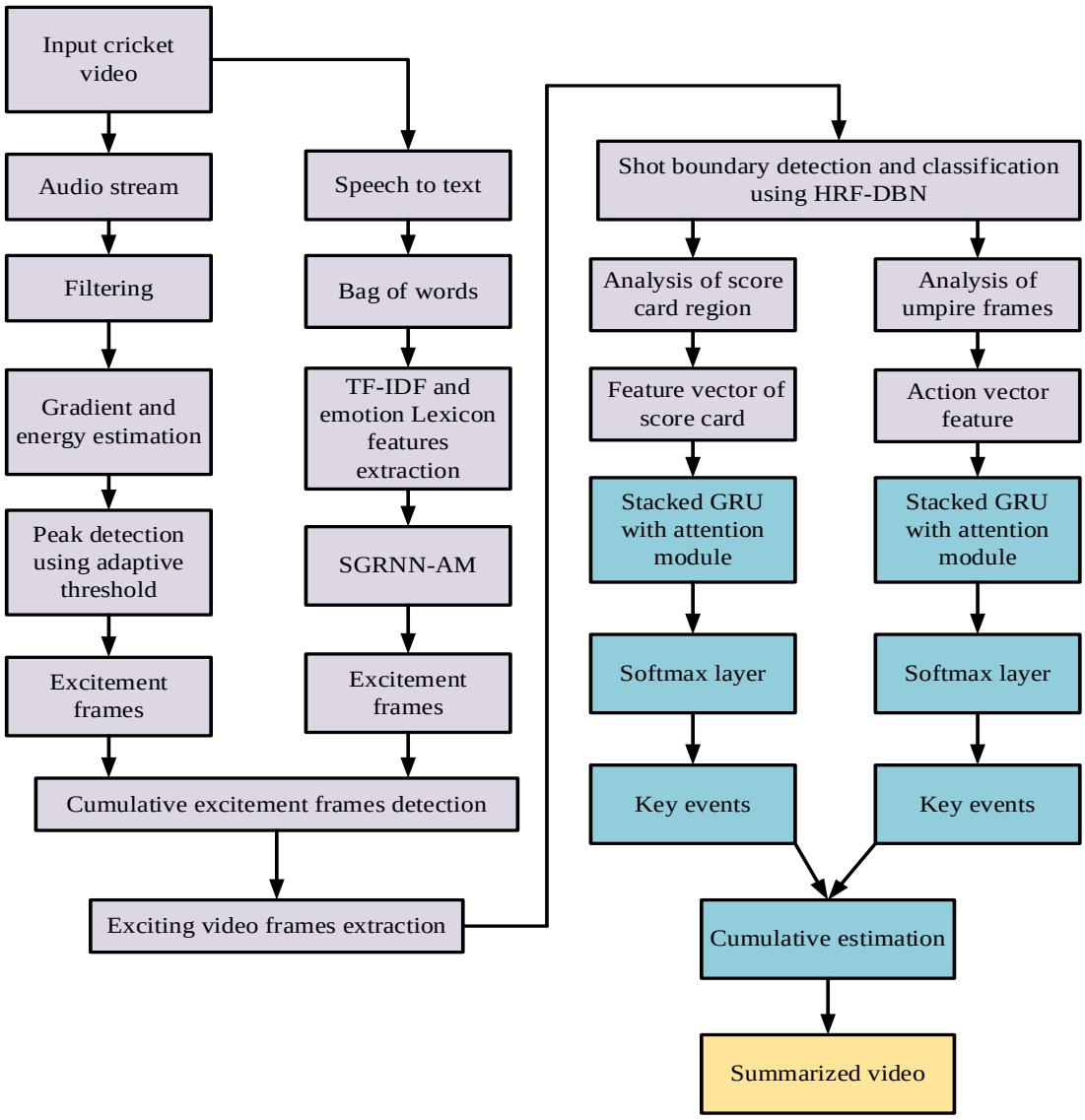

Figure 6.3: Hybrid machine learning-based video summarization

7 Achievements with respect to objectives

An optimized hybrid model for event recognition and summarization by combining both the event-driven and excitement-driven features. The exciting clips in the video are useful in the key event detection process. Thus, the audio energy present in a cricket video is determined and the exciting clips are extracted. Then, the exciting

clips are labeled effectively and the labeled concepts are chosen based on an important degree to form the cricket highlights

A novel hybrid model for cricket video summarization by considering the event-driven, excitement-driven, and object-driven features. Apart from determining only the audio energy, this approach adds additional confidence through the use of an effective "speech to text" framework to extract all the exciting clips. The characters and action features from the scoreboard and umpire gestures are determined to detect the keyframes from the exciting clips. Finally, an improved video summarization module is proposed to generate an effective summary for a cricket video.

8 Conclusion

Cricket is a popular sport played in the field by two teams with eleven members on each team. It has many fans and viewership around the world. Nowadays, cricket matches are viewed and shared internationally through live satellite broadcasting. However, cricket is generally a long duration and complex game with many more rules than other games like soccer, hockey, etc. So, due to its complex rules maximum range of events will be significantly based on circumstantial factors. Further, due to the emergence of a large number of sports, it has been difficult for the viewers to watch every news by consuming more time on cricket matches. Therefore, the process of cricket video highlight generation is necessarily required, and it has been considered as a significant research area due to its commercial importance and high viewership. Cricket highlight generation is defined as a process in which full-length video can be summarized into a shortened form by preserving the significant moments found in the original video. Many existing methods such as BBN, event-driven based non-learning model, distinctive algorithms, etc., have been introduced recently based on event and excitement driven. However, the event-driven model needs a long duration for generating the highlight, and the excitement-driven or other approaches take less duration but it minimizes the

efficiency. Since the existing methods use fewer annotation models and optical semantic concepts.

Therefore, by considering the above limitation, an efficient cricket highlight generation model, HDNN-EPO has been introduced using semantic meaningful concepts. In the HDNN-EPO model, the key events are extracted initially by employing low-level features namely skin tone, jersey color, edge density, and field. The exciting moments are described based on the spectator's applause, shouting, and cheering. The Hue histogram difference between the neighbors' frames is computed to identify the short boundaries or keyframes of each exciting clip. So that the complexities in the computation time can be reduced. Further, the scorecard regions can be located for differentiating the replay frames from the real frames. Then, to recognize fielders, player’s gatherings, batsmen, spectators, pitch view, umpires, and pitch view, low-level features such as edge, color has been extracted. Moreover, boundaries, disputes, wickets, injuries, etc., can be identified using an action recognizer. The action recognition process and low-level feature extraction algorithm are used to mine the events from each exciting clip. The HDNN-EPO method is employed for specific exciting clips to label the concept, and finally, arrange the selected clips in temporal order to generate the highlight. The performance is evaluated in terms of accuracy, precision, and recall. Finally, better outcomes of 93.71% accuracy, 93.43% of precision, and 92.46% recall have been achieved efficiently.

However, the HDNN-EPO model requires more optimal features to train and classify the input properly for attaining better accuracy. Therefore, to enhance the accuracy and minimize the computation complexity (by reducing the time taken for computations), there is a need for training through meaningful data. So, a hybrid machine learning method is introduced for automatically detecting key events and summarizing cricket videos. The hybrid model examines event, object, and excitement-based features from the cricket video to find the key event. First, the audio contents are examined by a speech-to-text framework, adaptive threshold (i.e.

binarization), and SGRNN-AM to extract the exciting clips. Then, HRF-DBN is introduced to classify the scenes of every exciting clip. Subsequently, the action and character features from the scorecard region of every umpire frame and key-frames of exciting clips have been extracted. Eventually, the key events such as wickets, sixes, and four are detected by employing the SGRNN-AM model. Moreover, accuracy is enhanced by using the attention module in the hidden output of GRU to choose the most important features. As a result, the proposed hybrid model has achieved 96.32% accuracy, 96.82% precision, 95.41% recall, 0.18% error rate, and 95.97% F1-score.

9 Copies of papers published and a list of all publications

Shingrakhia, H., Patel, H. SGRNN-AM and HRF-DBN: a hybrid machine learning model for cricket video summarization. *Computer Vision* (2021). https://doi.org/10.1007/s00371-021-02111-8

Shingrakhia, H., Patel, H. Emperor Penguin optimized event recognition and summarization for cricket highlight generation. *Multimedia Systems* **26,** 745–759 (2020). https://doi.org/10.1007/s00530-020-00684-3

10 Patent

Patent number: 2021102956

The Commissioner of Patents (Australian Govt.) has granted the above patent on 15 September 2021 and certifies that the below particulars have been registered in the Register of Patents.

Title of the invention:

A METHOD FOR EMPEROR PENGUIN OPTIMIZED EVENT RECOGNITION AND SUMMARIZATION FOR CRICKET HIGHLIGHT GENERATION

Name of inventor(s): SHINGRAKHIA, HANSA and NIKUNJ PATEL, HETAL

11 References

[1] Nasir, M., et al. "Event detection and summarization of cricket videos." *Journal of Image and Graphics* 6.1 (2018).

[2] Rafiq, Muhammad, et al. "Scene classification for sports video summarization using transfer learning." *Sensors* 20.6 (2020): 1702.

[3] Stevens, Tim, and Stephen Appleby. "Video delivery and challenges: Tv, broadcast and over the top." *MediaSync*. Springer, Cham, 2018. 547-564.

[4] 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.

[5] Ma, Mingyang, et al. "Video summarization via block sparse dictionary selection." *Neurocomputing* 378 (2020): 197-209.

[6] 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.

[7] Sanabria, Melissa, Frédéric Precioso, and Thomas Menguy. "A deep architecture for multimodal summarization of soccer games." *Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*. 2019.

[8] Khan, Abdullah Aman, et al. "Content-Aware summarization of broadcast sports Videos: An Audio-Visual feature extraction approach." *Neural Processing Letters* (2020): 1-24.

[9] Guntuboina, Chakradhar, et al. "Deep Learning-Based Automated Sports Video Summarization using YOLO." *Electronic Letters on Computer Vision and Image Analysis*20.1 (2021): 99-116.

[10] Rajpoot, Vinay, and Sheetal Girase. "A study on application scenario of video summarization." *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2018.

[11] Kaushal, Vishal, et al. "A framework towards domain-specific video summarization." *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019.

[12] Rajpoot, Vinay, and Sheetal Girase. "A study on application scenario of video summarization." *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2018.

[13] Javed, Ali, et al. "A hybrid approach for summarization of cricket videos." *2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. IEEE, 2016.

[14] Shingrakhia, Hansa, and Hetal Patel. "Emperor Penguin optimized event recognition and summarization for cricket highlight generation." *Multimedia Systems* 26.6 (2020): 745-759.

[15] Premaratne, S. C., K. L. Jayaratne, and P. Sellapan. "Improving Event Resolution in Cricket Videos." *Proceedings of the 2nd International Conference on Graphics and Signal Processing*. 2018.

[16] M. M. Salehin and M. Paul, "Fusion of Foreground Object, Spatial and Frequency Domain Motion Information for Video Summarization," Image and Video Technology – PSIVT 2015 Workshops, Springer, Cham, pp. 319-331, 2016.

[17] J. Zhu, S. Liao and S. Z. Li, "Multicamera Joint Video Synopsis," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 6, pp. 1058-1069, June 2016.