JSR Publications

Available Online

J. Sci. Res. **14** (2), 533-544 (2022)

JOURNAL OF SCIENTIFIC RESEARCH

www.banglajol.info/index.php/JSR

Video Summarization Using Deep Learning for Cricket Highlights Generation

D. P. Gaikwad[1*], S. Sarap[1], D. Y. Dhande[2]

[1]Department of Computer Engineering, AISSMS College of Engineering, Pune, India

[2]Department of Mechanical Engineering, AISSMS College of Engineering, Pune, India

Received 2 December 2021, accepted in final revised form 19 February 2022

Abstract

Recently, video surveillance technology has grown pervasive in many aspects of our lives. Automatic video monitoring produces massive amounts of data that need human examination at some point. The primary emphasis is on reducing storage usage by compressing or eliminating superfluous frames without sacrificing real information. The current effort seeks to close the growing gap between the amounts of real data and the volume. Searching through key events in large video collections is time-consuming and tedious. In this paper, smart surveillance for various applications by using video summarization has been presented. A method for generating highlights has presented which pre-processes extracted Video Frames. Convolutional Neural Networks are then used to evaluate these highlighted frames. The proposed technique extracts and calculates characteristics utilized to generate summary movies. For training deep neural networks, cricket datasets have been used. Experimental results show that the proposed solution attains improved results than other advanced summarization methodologies. Experimental results show that the proposed video summarization method consistently generates high-quality reviews for all types of videos. The proposed video summarization method is easy to use, and it can also help extract highlights of cricket games with high accuracy.

Keywords: Video surveillance; Convolutional neural networks; Feature extraction; Cricket highlight generation; Summarized video; Predefined highlights dataset.

doi: http://dx.doi.org/10.3329/jsr.v14i2.56856 J. Sci. Res. **14** (2), 533-544 (2022)

1. Introduction

Nowadays, viewers need help watching digital videos as the explosion of digital content on the Internet and television increases. Using the preview to understand the quality of the video is much better than watching the entire video. The video summarization system creates summaries by analyzing and shortening the video content, audio, or text from multi-modal video. Video summarization is used to generate teasers and trailers of movies and episodes of a TV serial. It is also used to present highlights of the event of sports games and music band performances [1]. These features can also be used by combining them with one another. Video summarization analysis is a process for eliminating noise in signal levels or semantic information. It is also used to create short videos or extract

[*] *Corresponding author*: dp.g@rediffmail.com

keyframes that summarize the main content of a long complete video. The purpose of video summaries is to accelerate the scrolling of large amounts of video data while allowing optimal access and interpretation of video content [2]. After reading the review, users will quickly decide on the usefulness of the video. Depending on the plan and target audience, the final evaluation may include usability studies to determine the credibility and quality of the material. There are two types of video summarization, one is static and another is dynamic [3]. The picture abstract or static storyboard refers to static video summarization. Three main approaches are used to categorize static summarization. Three approaches; Shot segmentation, Schematic segmentation, and sample-based video segmentation, are among them. This method uses the keyframe method to select a frame or part of multiple frames of the video to create a video description. The dynamic video summarization methods generate a subset of the original video [4]. It is used to preserve the time-based components by performing the shot-level analysis. Dynamic video summaries use video extraction technology to compress longer videos into shorter videos. In complex video applications, the skimming textual processing approach is used. Using video summarization, the viewer can see shorter videos from the initial footage, which are difficult to interpret. Highlights of any game can describe the entire game to comprehend the entire video. Video summarization in mobile video descriptions can save memory, extend battery life and reduce the cost of downloading videos since many people transmit information such as football, movie, and music downloads online. In this information, video summary technology can speed up browsing, especially when indexing content. Video summarisations also can be used in surveillance tracking systems. Surveillance system using video summarisation faces many challenges such as dynamic processing, visual processing, and Data management [5].

Video summarization for a cricket match in the past few years has fascinated people in automated sports material processing [6]. The explosion of sports coverage on the internet is one of the main reasons. Sports enthusiasts will find it difficult to keep up with many sports events happening year-round. As a result, highlights serve as a valuable content source and let the audience stay up to date on what is happening without wasting too much time. Several surveys on video summarization for cricket highlights generation have already appeared in the literature. In one of the first works, Asim *et al.* [7] have proposed video summarization using the keyframe extraction method. It is based on comparing color features from frame patches. In this method, only benchmark datasets have been used for video summarization to validate the performance of the proposed approach. The system has not validated using several different types of datasets. Metal [8] has proposed a deep learning-based robust scheme to find interesting events. These events have joined together into an efficient summary. In this system, DVE mechanisms have achieved accuracy up to 98.43 %. The authors have used the transfer learning method for this summarisation. The accuracy of this system can be used by proposing new architecture of deep learning. Javed *et al.* [9] have proposed an automatic method for key-events detection for video summarization for the cricket match. Initially, rule-based training is applied to detect excited audio clips in cricket videos. A decision tree framework is designed for video summarization. This system offers average accuracy of 95 %. Bhalla *et al.* [10] have proposed a novel technique for identifying and summarizing important events during a cricket match. This model takes the whole cricket match as input and produces the game's most important clips as performance. In cricket matches, visual character recognition, sound analysis, and replay detection have been utilized to extract critical occurrences such as lines, wickets, and other playfield events. From these events, the whole cricket match clip was then edited together. It also assessed the model using several qualitative and quantitative investigations. With an accuracy of the present, the suggested model identifies wickets, fours, and sixes, demonstrating the model's usefulness in real-world situations. In this method, the video clip is broken up into several video shots. The essential elements of these video snippets have been extracted as features to make match highlights. Shukla *et al.* [11] have suggested a paradigm for automatically producing sports highlights, emphasizing cricket. Cricket is a sport with complex rules and is played longer than most other sports. This work proposes a method for identifying and cutting out key events in a cricket match that takes into account both event-based and emotion-based features. Cues used to record certain activities include replays, audio intensity, player celebration, and playfield scenarios. To test the framework, s attempt to run a series of tests spanning from usage studies to a comparative study of highlights to those offered by official broadcasters. The widespread acceptance of the proposed model by consumers and the considerable overlap in both types of highlights demonstrate its utility in real-world scenarios. Jadon *et al.* [12] used a standard vision-based algorithmic methodology for correct feature extraction from video frames to overcome video summarization by unsupervised learnings. They suggested a deep learning-based feature extraction process, followed by several clustering techniques, to find an appropriate way of summarizing a video by extracting interesting mainframes. We contrasted the efficiency of these approaches on the SumMe dataset and found that deep learning-based feature extraction performed better in the case of dynamic viewpoint images. Karim [13] has presented VGRAPH, a simple yet effective video summarization tool incorporating color and texture capabilities. This approach focuses on segmenting the video into shots utilizing color features. It is used for gathering main video frames with the nearest neighbour graph built from the texture features of the shot representative frames. Furthermore, this proposal integrates and illustrates an updated appraisal strategy based on color and texture matching. The VGRAPH video summaries are compared to summaries created by others and ground reality summaries found in the literature. According to the experimental findings, VGRAPH video summaries are of better quality. Authors have used Video Fragments One frame per second pre-sampling, HSV (Hue-Saturation-Value color histogram for Temporal Video Segmentation. K-nearest neighbour graph (k-NNG) has been used to extract Key Frames. Zhang *et al.* [14] have proposed a query-conditioned three-player generative adversarial network. In this work, the joint representation of the user query is learned by the generator. The discriminator is used to discriminate the real summary from a generated and a random one. The generator and discriminator are trained using a three-player loss. This method is capable of forcing the generator to learn better summary results. It also helps to avoid the generation of random trivial summaries.

In the literature survey, it is observed that the existing system has some drawbacks such as less accuracy, more time for training, and lower detection rates. Motivated by these observations, it is aimed to fill these gaps in the literature by presenting deep-learning-based video summarization. In this paper, a new method for automatically generating cricket high-lights, photographs, emphasis on cricket have presented. The main objective of the proposed system is to abstract the video to reduce the time consumed to send and watch the video and examine the most important events from large video databases. Another objective was to save space by compressing or removing redundant frames without sacrificing actual content. Moreover, the main thought was to measure the system's performance using different datasets. The four major occurrences in cricket matches containing wickets, borders, six-pointers, and landmarks are extracted using event-driven features. The excitement features are used to distinguish the remaining significant cases. Instead of optimizing for representation, it is tried to find methods for video summaries (such as OCR, HOG, HSV, etc.), highlights, or fascinating moments in work images. Several new methods have been developed. Rather, the focus has been given to cricket videos because they provide more structure and indicators. Several research conferences have been dedicated to video summarization over the last two decades. An outline should be as short or concise as possible, according to the specification, to allow for easier video browsing.

On the other hand, it can be exact to include as many distinct and useful services as possible. To achieve both elegance and informativeness, the key to video summarization is assessing the importance of various media clips and selecting the most significant ones used for the generation of summarized videos. The rest of the paper is organized as follows. Section 2 has dedicated to describing the proposed video summarization using deep learning. In section 3, experimental results have been discussed. Finally, in section 4, the paper has concluded with future scope.

2. Proposed Video Summarization Using Deep Learning

In this paper, an innovative method for automatically creating cricket highlights has been proposed. Event-driven features retrieve the four main events in a cricket match. Wickets, borders, sixes, and landmarks are identified using excitement features, while other significant occurrences are identified using wickets, borders, sixes, and landmarks. Highlights are produced based on events. Many cricket matches and highlights have looked to collect a list of incidents in those broadcasts widely regarded as significant. Different analogies in video shots have been used to act boundaries, sixes, and wicket falls. When numbers of runs between two consecutive video shots are equals to four, then detected as four runs. When a border is visible, then six runs are considered. When the amount in the wicket column rises by one with two consecutive video shots, the wicket is said to have dropped. Excitement-based video summarization has been proposed based on the volume of audio, crowd appeals, loudness. These three parameters detect important game events. Milestones and random occurrences have been detected using audio features, such as boundaries, sixes, and wickets, as well as grab drops. Posed architecture of video summarization has shown in Fig. 1. The architecture consists of two main modules one is system input, and the second is system output. Following different steps of proposed architecture have been explained in short.

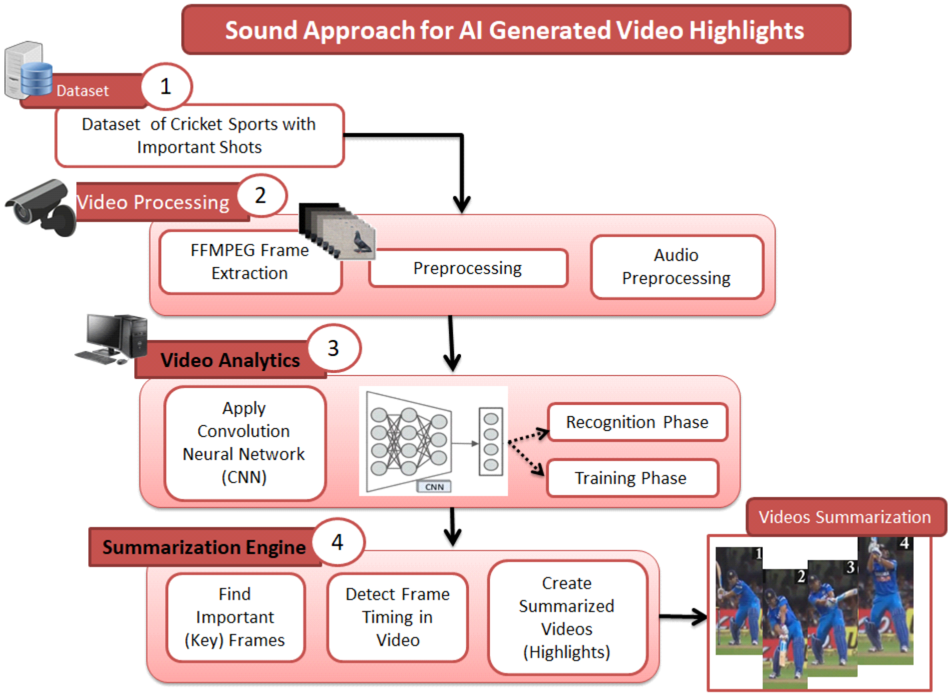
Sound Approach for AI Generated Video Highlights
Dataset
1
Dataset of Cricket Sports with Important Shots
Video Processing
2
FFMPEG Frame Extraction
Preprocessing
Audio Preprocessing
Video Analytics
3
Apply Convolution Neural Network (CNN)
CNN
Recognition Phase
Training Phase
Summarization Engine
4
Find Important (Key) Frames
Detect Frame Timing in Video
Create Summarized Videos (Highlights)
Videos Summarization
1
2
3
4

Fig. 1. The proposed system architecture.

2.1. *Video extraction and pre-processing*

Video is a combination of images in sequence. FFMPEG is used to extract frames from any video format. Video frame extraction is done at 25 frames per second. Each frame is saved with a number to create a summary later. Evolves converting the frame into the grayscale image of 128×128. Once the image is resized, it will be sent to Convolutional Neural Network for matching.

2.2. *Train dataset*

The proposed summarization framework is verified on 3 data sets, i.e., TVsum50, SumMe, ADL, for edited videos, short raw videos, and long raw videos, respectively [9].

The proposed framework is compared with various popular editing and raw video approaches based on these data sets.

2.3. *CNN training on a predefined high-lights data set*

The different layers of a CNN are the convolutional layer, the pooling layer (Max, Avg.), the ReLU layer, and the fully connected layer. CNN is used to train the video analysis engine to recognize important frames in the video.

2.4. *Feature extraction and highlight generation*

In this paper, features have been extracted using a convolution neural network and suggested by Gupta and Sowmya [15,16]. Frame number-wise important frames will be stored. Each frame number will be mapped to that video duration; an additional 5 seconds before and after duration will be added to catch the exact highlight. Generated Frame numbers are clubbed to form a 25 fps video. A convolutional neural network is used to extract features for training purposes. A Three-layer CNN network is used to train the network. Fig. 2 depicts the overall steps for feature extraction. The following sequence of the process has been followed for feature extraction.

Input → frames of timing 25 frames per second (FPS) → 45 frames for 2.1-sec → Output (Extracted Features) weightage. Fig. 2 shows different steps for feature extraction.

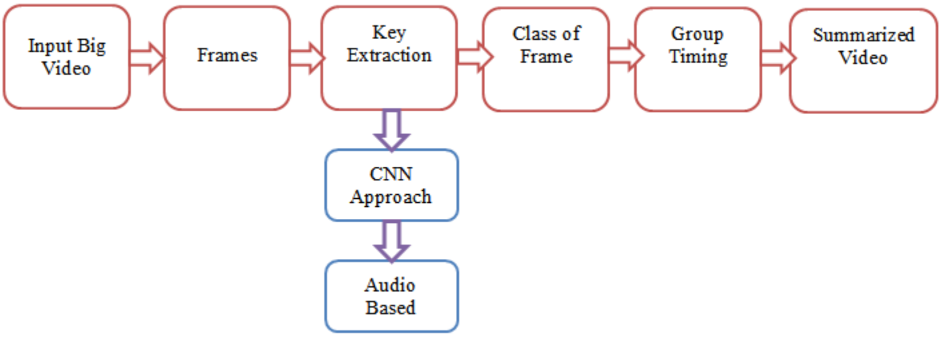

Fig. 2. The steps of keyframe extraction.

2.5. *The proposed algorithm for video summarization*

Video summarization procedures help to create a compressed form of the original video. The Convolutional Neural Network is used to present a different course for keyframes extraction. Frame by frame, the technique works for videos. Reliant on their order, the out-of-work frames are detached first. To extract high-level function vectors, Convolutional Neural Network is used. Frame quality descriptors are further separated into key and non-key frames. Benchmark databases of cricket match images were used to endorse the method. Following Algorithm 1 summarises the process of proposed video summarization.

Algorithm 1: General algorithm for video summarization

Step 1. Start by transferring a picture to the first coevolutionary sheet.
Step 2. Construct a convolutional kernel for image sharpening or smoothing.
Step 3. Convolutional neural network: Use filters to remove appropriate features from the input image before passing it on.
Step 4. Striding and Padding for reduction of number of features.
Step 6. Use Activation Function to understand non-negative linear values from real-world data.
Step 7. Use Pooling layers are also used to drastically condense the number of parameters.
Step 8. Use a completely linked layer to train the network.

3. Experimental Results and Analysis

For output evaluations, we select YouTube cricket sports videos based on varying lengths of time and edited and raw encoding. India and Pakistan cricket match videos are used for the experiment. The data set is 5.88 GB in size. The videos are 3-20 minutes long, and some are 3 hours each. In Fig. 3, the dataset used for training has shown. Fig. 4 shows the extracted frame for training the convolution neural network. In Fig. 5, the timeline of video highlights has depicted. Timings indicate highlights time duration within the video.

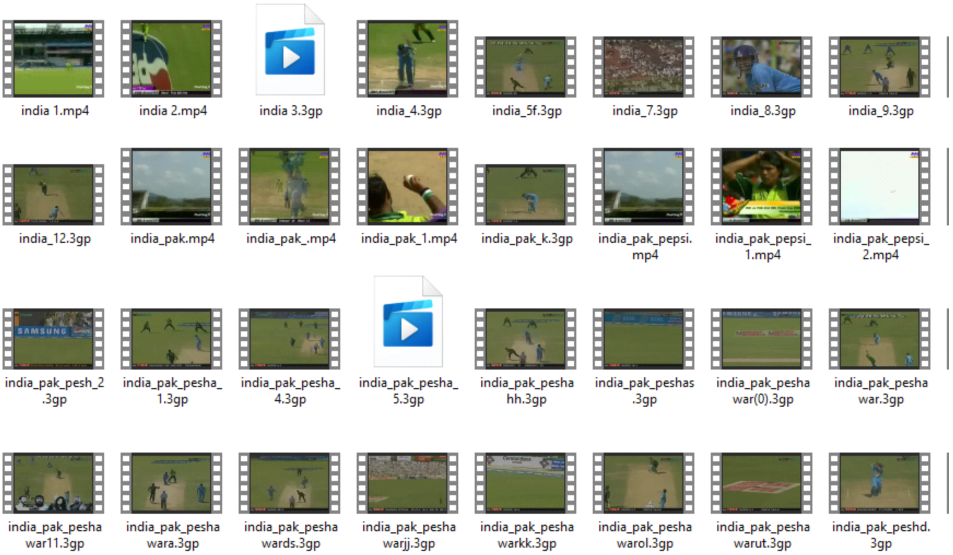

Fig. 3. Dataset view.

Fig. 4. Frame extraction from video.

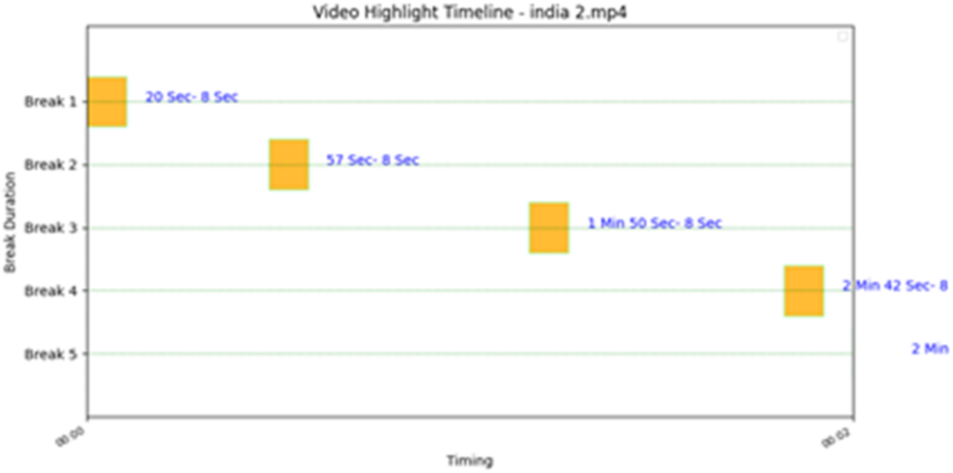

Fig. 5. Video Highlight Timeline.

Fig. 6 is used to depict the determined number of peaks in sound transmission. In Fig. 7, the extracted audio from the video has shown.

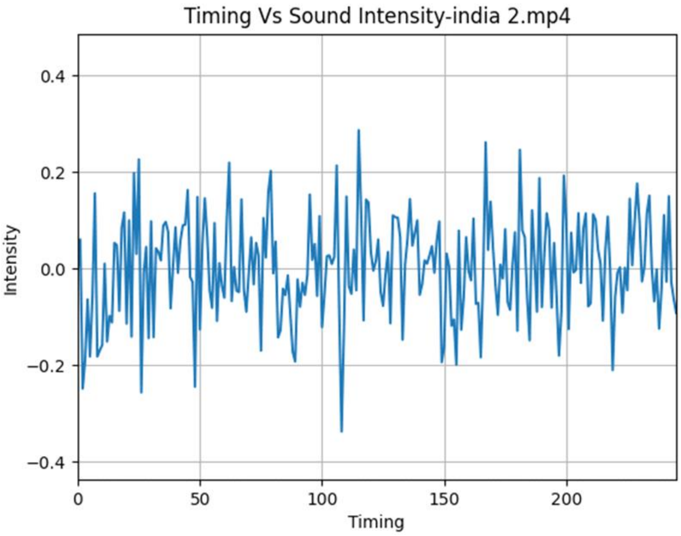

Fig. 6. Timing v/s sound intensity.

india 1.mp4.wav
india 2.mp4.wav
india 3.3gp.wav
india_4.3gp.wav
india_5f.3gp.wav
india_7.3gp.wav
india_8.3gp.wav
india_9.3gp.wav
india_12.3gp.wav
india_pak.mp4.wav
india_pak_.mp4.wav
india_pak_1.mp4.wav
india_pak_k.3gp.wav
india_pak_pepsi.3gp.wav
india_pak_pepsi.mp4.wav
india_pak_pepsi_1.mp4.wav
india_pak_pes.3gp.wav
india_pak_pesh_2.3gp.wav
india_pak_pesha_1.3gp.wav
india_pak_pesha_4.3gp.wav
india_pak_pesha_5.3gp.wav
india_pak_peshahh.3gp.wav
india_pak_peshas.3gp.wav
india_pak_peshawar(0).3gp.wav
india_pak_peshawar6.3gp.wav
india_pak_peshawar11.3gp.wav
india_pak_peshawara.3gp.wav
india_pak_peshawards.3gp.wav
india_pak_peshawarjj.3gp.wav
india_pak_peshawarkk.3gp.wav
india_pak_peshawarol.3gp.wav
india_pak_peshawarut.3gp.wav

Fig. 7. Extracted audio from video.

Generated highlights can cover only important sections or highlights of the entire video, devoid of watching the full news and of several hours. The system uses extracted Video Frames for pre-processing. After that, there are these highlighted frames. Convolutional Neural Network is used to analyze the data. Features are identified as a result of this analysis. Extracted and calculated data is used to create summarized videos. The data file contains the cricket videos with the MP4 file format. They contain different classes like Batting shot(N), Catch(L), LBW(R), Sixes(A), Boundary hit(P), Normal(V). For training and testing purposes, it uses 74 videos. The outcomes are divided into six categories. The classes and their labels are listed in Table 1 below.

Table 1. Classes of video summarization.

Sr. No.	Class	Class Label
1	Batting Shot	N
2	Catch	L
3	LBW	R
4	Sixes	A
5	Boundary Hit	P
6	Normal	V

The following matrices have been used to evaluate performances of video summarization in terms of precision, sensitivity, and F-score values. Based on the above TP, TN, FP, and FN values, the Accuracy, Precision is calculated using the formulae below. Equations "Eq. 1" to "Eq. 4" have been used to calculate precision, sensitivity, F-score value, and accuracy, respectively.

$$Precision = \frac{TP}{(TP+TN)} \quad (1)$$

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (2)$$

$$F - Score = \frac{(2*TP}{(2*TP+FP+FN)} \quad (3)$$

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

Table 2 shows the confusion matrix on all classes. In Table 3 and Fig. 8, accuracy and precision values have shown in different classes.

Table 2. Confusion matrix for the six classes**.**

Classes	TP	TN	FP	FN
Class N	1971	9501	37	29
Class L	1940	9532	83	60
Class R	1891	9581	195	109
Class A	1786	9686	137	214
Class P	1958	9514	36	42
Class V	1926	9546	40	74

Table 3. Confusion matrix for the six classes.

Classes	Class Label	Accuracy %	Precision
Batting Shot	Class N	99.42	98.15
Catch	Class L	98.76	95.89
LBW	Class R	97.41	90.65
Sixes	Class A	97.03	92.87
Boundary Hit	Class P	99.32	98.19
Normal	Class V	99.01	97.96

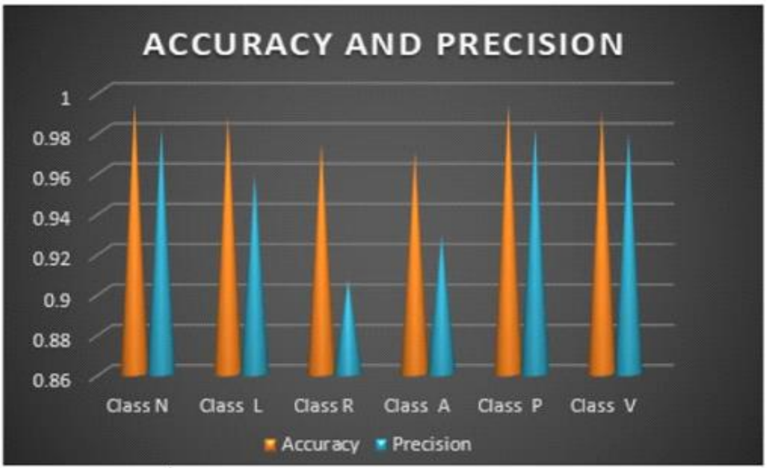

Fig. 8. Performance of proposed method in terms of accuracy and precision.

Overall, the proposed method for video summarization offers 99 % in almost all classes which is more accurate than the proposed summarization system reported by Muhammad *et al*. [8] and Javed *et al*. [9].

4. Conclusion

In this paper, the video summarization system has been studied and proposed a video summarization system using a convolution neural network. A summarization method produces an abstract version of its inputs for user consumption. With so much data on social media, it is critical to explore the text to find information and apply it to a wide variety of users. This research work aims to bridge the gap between the increasing amount of data produced and the number that can be successfully checked manually. This research work proposed a novel methodology for video summarisation have presented in which the automatic selection of important keyframes takes place. It only trims the original video to obtain meaningful keyframes for the entire video story, and after the combination of generated keyframes, the summarizing video is displayed. Examining the most important incidents from large video databases takes time and effort. The following key points were included in the completed work. In this research, excitability estimates in cricket videos are used to create automatic highlights. These featured keyframes are then analyzed using Convolutional Neural Networks. It is a video summarization system that uses extracted keyframes for pre-processing. This analysis extracts and calculates features and summarises them. Videos are generated. In the future, experiments on various forms of footage, such as sporting videos and monitoring device videos, will be conducted. It is also possible to add more activities like different sports, daily activities, etc., and to use for a smart surveillance system.

References

1. E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, Video Summarization Using Deep Neural Networks: A Survey, in Computer Vision and Pattern Recognition (2021). arXiv: 2101.06072v2
2. M. Basavarajaiah and P. Sharma, ACM Computing Surveys **52**, 116 (2020). https://doi.org/10.1145/3355398
3. M. S. Nair and J. Mohan, Signal Image Video Process. **15**, 735 (2021). https://doi.org/10.1007/s11760-020-01791-4
4. J. Mohan and M. Nair, IEEE Access **6**, 59768 (2018). https://doi.org/10.1109/ACCESS.2018.2872685
5. D. Isravel, S. Silas, and E. Rajsingh, Intelligent Data-Centric Syst. 1 (2020). https://doi.org/10.1016/B978-0-12-816385-6.00001-5
6. S. Emon et al. Automatic Video Summarization from Cricket Videos Using Deep Learning, - *Proc. Int. Conf. on Computer and Information Technol.* (ICCIT) (2020) pp.19-21.
7. M. Asim, A. Bouridane, and A. Beghdadi, Color and Visual Computing Symposium (2018).
8. R. Muhammad, R. Agyeman, H. K. Shin, R. Ali, K. –M. Kim, et al., Deep Video Events (DVE), A Deep Learning Approach for Sports Video Summarization – *Proc. of 2018 IEMEK Symp. on Embedded* Technol. (2018).
9. A. Javed, K. B. Bajwa, H. Malik, A. Irtaza and M. T. Mahmood, A Hybrid Approach for Summarization of Cricket Videos - *IEEE Int. Conf. on Consumer Electronics-Asia* (Asia), (2016) pp. 1-4. https://doi.org/10.1109/ICCE-Asia.2016.7804835
10. A. Bhalla, A. Ahuja, P. Pant, and A. Mittal, A Multimodal Approach for Automatic Cricket Video Summarization - *6th Int. Conf. on Signal Processing and Integrated Networks* (SPIN) (2019) pp. 146-150. https://doi.org/10.1109/SPIN.2019.8711625
11. P. Shukla, H. Sadana, A. Bansan, D. Verma, C. Elmadjian, *et al*., Automatic Cricket Highlight Generation Using Event-Driven and Excitement-Based Features, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (CVPRW) (2018), pp. 1881-18818. https://doi.org/10.1109/CVPRW.2018.00233
12. S. Jadon and M. Jasim, Unsupervised video summarization framework using keyframe extraction and video skimming - *IEEE 5th Int. Conf. on Computing Communication and Automation* (ICCCA) (2020) pp. 140-145. https://doi.org/10.1109/ICCCA49541.2020.9250764
13. K. M. Mahmoud, N. M. Ghanem, and M. A. Ismail, VGRAPH: An Effective Approach for Generating Static Video Summaries - *IEEE Int Conf. on Computer Vision Workshops* (2013), pp. 811-818. https://doi.org/10.1109/ICCVW.2013.111
14. Y. Zhang, Michae, X. Liang, and E. P. Xing, Query-Conditioned Three-Player Adversarial Network for Video Summarization, Computer Vision and Pattern Recognition (2018). arXiv:1807.06677
15. S. Gupta and R. Sedamkar, J. Sci. Res. **13**, 901 (2021). https://doi.org/10.3329/jsr.v13i3.53290
16. V. Sowmya and R. Radha, J. Sci. Res. **13**, 809 (2021). https://doi.org/10.3329/jsr.v13i3.52332