JOURNAL OF INFORMATION SCIENCE AND ENGINEERING XX, XXXX-XXXX (2016)
DOI: XX.XXX.XXXX.XX

VSEM: A Hybrid Model for Video Summarization

PRASHANT GIRIDHAR SHAMBHARKAR, RUCHI GOEL, +
Department of Computer Science and Engineering
Delhi Technological University
Delhi, 110042
E-mail: prashant.shambharkar@dtu.ac.in; ruchigoel@mait.ac.in

In today's fast-moving digital era, video technology plays an important role. An effective video summarising approach is urgently needed to handle a lot of video data due to the ever-growing number of video content. In this paper, the authors have proposed, A hybrid summarization methodology for video summary evaluation using multimedia features (Text, Images, and Audio) that assess how well a video summary can keep the ranking of vital video frames, semantic data, and audio present in the original video. Video summary can be evaluated by ranking text, audio, and semantics of video frames, giving more accurate summarisation results. The proposed methodology works in three phases: The first part takes the text in the video, the second phase takes the audio to the file, and the last phase focuses on the video frames rather than images in the video. TVSum dataset has been used for the experimentation. F1 has been used as the evaluation metric for checking the efficacy and efficiency of the proposed methodology. The result shows that the proposed hybrid model achieves the highest F1 score of 69.9% and saves 75-80% of user time in watching video summaries instead of the whole video.

Keywords: Video Summarization, NLP, Multimedia Features, Computer Vision

1. INTRODUCTION

Technology development has caused a quick increase in multimedia data on the Internet, making it difficult for consumers to access crucial information quickly [1]. Video is the most challenging multimedia (including text, pictures, graphics, and audio), as it incorporates all other media data into a single data stream and is difficult to access effectively due to its unstructured format and changing format length[2]. Video information is a sequential data type that gives unlimited data through its moving content [3]. Think about using YouTube to search for educational or tourist-related content, many individuals prefer not to invest their time in watching or listening to lengthy recordings. Instead, they often seek out concise video clips that provide a condensed and more digestible summary. It is inefficient to browse through the millions of returned results. It would be much simpler to view a brief description of each result. Secondly, because of the limited storage space, it is also necessary to summarise videos without losing much information. These

+ Ruchi Goel

issues can be solved by summarizing the essential information from the vast amount of available content. Video summarization methods pique the viewer's interest by choosing exciting scenes from the original video [4]. By highlighting significant portions of the original video, video summarising techniques can grab the audience's attention [5]. The viewer can comprehend the information without watching the clip. To extract specific critical frames from a video, video summarization creates a representative summary with a smaller file size. Both the identification of the various activity sequences across time and the accurate summary of each series with the next are necessary for adequate video description approaches [6].
Additionally, eliminating redundant and useless video content may have uses in video retrieval, storage, and indexing [7]. It will also increase the effectiveness of associated video analysis tasks, including action recognition and video captioning [8]. Manually summarising and editing videos requires a lot of time and work. An automatic summarizing approach is necessary to identify important events in the original video content. A quality video summary is characterized by its ability to achieve a high level of recall, maintain a high level of precision, and minimize redundancy [9]. Creating a good video summary requires thoroughly comprehending the video's structure and semantic content.

One of the challenging issues associated with the video summarization problem is the decreased computational costs to produce consistent video summaries from vast volumes of data. Another challenging task is the effective fusion of multimedia resources, such as audio, text, image, and video [10]. The significant occurrences can be automatically identified by evaluating the text, audio, and visual elements. Retrieving information from audio or visual content is still tricky because high-level semantic information needs to be recovered from low-level audio or visual data.
Video-based applications are used in various fields, such as security and surveillance, personal entertainment, medicine, sports, news videos, educational programs, movies, etc. A series of images with some timely information make up the video. The textual information represents the information's linguistic form, while the audio consists of speech, music, and numerous distinctive noises. Rich media includes video, frequently combining other media forms, including text and audio [11].
Examining several media modalities, including text, audio, and visual information, is necessary for video representation [12]. The video format can include a variety of components, including audio and textual information (such as closed captions). t serves to elucidate the sequence, organization, and content of the individual frames that collectively form the moving video image. A modality in the multi-modal space depends on how particular media and associated elements are organized inside a conceptual architecture. These modalities involve specific techniques or methods to encode heterogeneous information harmoniously and may include textual, visual, and aural modalities. Multi-modal learning, especially audiovisual learning, has recently garnered a lot of attention and has the potential to make many computer vision tasks [13]. However, current video summarization techniques only consider the visual data and ignore the text and audio data. In this study, we contend that the text and audio modality can help the visual modality comprehend the structure and content of the video more effectively, which will also help the summary process. assumption is that models may generate a better and more comprehensive knowledge of the underlying data, reveal new insights, and allow a wide range of applications by combining information from varied sources such as text, pictures, voice,

and video. In comparison to previous methods using pixel-based or text-based video summary evaluation using multimedia features to choose the most representative or noteworthy, exciting video portions, VSEM simultaneously uses the features of text, audio, and visual information for the video summarising task.
The paper is structured as follows: Section 2 provides information about the literature survey in this field. The problem statement of the paper is discussed in section 3. Section 4 of the article offers the proposed methodology of hybrid video summarization. Section 5 of the paper discusses the dataset used, the methodology's findings, and the analysis.

2. LITERATURE SURVEY

A typical computer vision task created for video analysis is video summarization. A decent summary must adhere to at least two goals. First, it should include the most captivating segments of the video; for instance, in a football match, one doesn't want to skip highlights like the kickoff. Second, the information in the video can be effectively condensed by using multiple keyframes or key-shots to show the video material [14].
Summarizing multimedia content has not received as much attention from researchers as text summarization has over the years [15]. In [16], authors provide an extractive summary using two text summarization algorithms and video mapping algorithms. The information in the video can be effectively condensed by using multiple keyframes or key-shots [17]. In contrast to the discipline of computer vision, there has been a significant advancement in the evaluation of text summaries in the NLP community. First, NLP approaches were developed to assess the caliber of text that had been machine-translated from one language to another. Authors [18] employ an existing text summarising evaluation and map a video summary into text. This has the benefit of allowing semantic comparisons to be made between outlines. However, it also means that the judgment does not include visual elements like shaky cameras as long as a specific piece of content is portrayed. By measuring the number of sub-shots that overlap between a given video summary and a ground-truth video summary, authors [19] develop VERT. This system assesses video summaries compared to a provided video summary. The drawback of pixel-based distance is another drawback of this technology. Additionally, individuals frequently struggle to create a video synopsis that accurately reflects the video instead of writing, which they find easier to produce. Asynchronous text, image, audio, and video-based summary of the video was suggested by Haoran Li et al.[12]. After analyzing each asynchronous component separately and using several optimization techniques on the summary, a more accurate final textual summary is generated. Salience matching is also carried out to improve the relevance of the summary.
A temporal and spatially driven method was put out in [20] in which the number of keyframes were automatically determined and extracted using Optimum-Path Forest (OPF) clustering before being utilized to create the final summary. Finding important frames in video summarization is an important and tedious task. A deep learning-based approach for learning video representation was proposed by Michele Merler et al. [21]. Deep learning models like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), etc., are used to process the audio and visual content to learn the representation. A video that serves as the final summary is produced. To understand

the spatiotemporal representations from video, the authors [22] recommended using a ranking-based method to summarise the video in many stages. Ali Javed et al. [23] suggested a technique for enumerating the cricket video's audio-visual components. The final summary for the cricket match videos is prepared by identifying the critical frames for the audio and visual content. To increase the effectiveness of summarization, authors used an Audio-Visual Recurrent Network (AVRN) to include audio and visual information in video summary tasks [24]. Utilizing the latent consistency between audio and visual data is possible with the audio-visual fusion LSTM. The self-attention video encoder can detect global dependencies throughout the entire video stream. In [25], abstractive summaries of narrated instructional are generated on several subjects, including sports, cooking, and gardening using transfer learning. A pre-trained BERT encoder and a transformer decoder with random initialization are used in the transformer design. Auto-generated instructive video scripts using the BertSum abstractive summarization model have a quality level comparable to descriptions chosen randomly from YouTube user submissions.

3. PROBLEM STATEMENT

It is observed that multimedia features (text, audio, and image) play an essential role in a video summary, and combining these can be effective. Video summaries on YouTube are currently based on the relevance of the frames in each video. To accurately summarise a video, we suggest taking a three-pronged approach as shown in Figure 1. Thus,the problem statement into three independent sections. The first part concerns the subtitles in a video. We employ a text summarization tool to convert the subtitles into a shorter version that includes complete sentences. Each line of the subtitles of importance is considered. Thereby, the whole list of sentences in the subtitle file acts as a corpus. Each line in the corpus is already mapped to the timestamp relative to the video. The summarization results are then divided and mapped into a list of sentences based on the timestamp.

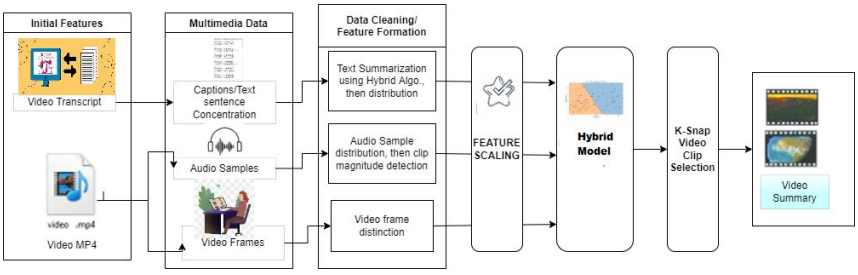

Fig. 1. Video Frame Analysis

To tackle the audio aspect of the multimedia, files are obtained in .wav format, and for 0.1 seconds (Persistence of Hearing) of the video, an array of audio samples is taken. These samples can be considered separate sound waves with their troughs and crests. From these audio chunks MFCC (Mel-frequency cepstral coefficients), Mel Spectrum,

area Under the audio Curve, and audio peak after average cut-off were obtained.
In the second part, to address the audio component of the multimedia, files in the .wav format are used, and a variety of audio samples are taken for each frame of the video. These samples can be considered separate sound waves with their troughs and crests. The amplitude of each such change is compared to obtain the magnitude of that wave. This list of magnitudes is linked to timestamps in the video and is used to determine the frames where the video is silent and where it is lively.
The third part of images or video frames is traversed by considering the changes in each frame. In a video, the fps (frame rate per second) used to be 24 (generally)[26]. Still, with the advancing technology, it is often 120, 240, or even 300. For this high frame rate, the change in a video frame is minute and insignificant upon regular inspection. It is necessary to consider the picture array (pixel arrangement within the image) to determine the precise changes in the next frames. Thus images are treated as an n-dimensional array rather than an image. Mean Absolute Difference is used to track changes over a few n-dimensional arrays with the same dimensions. This allows us to find the crucial frames in the video and spot changes.
The accuracy of our summarized video is compared to the gold standard of the video summarization dataset. The redundancy could affect how accurate the video is.

$$C_{vs} = [\{T_1 + T_2 + T_3 - - + T_n\} \cup \{A_1 + A_2 + - - + A_n\} \cup \{F_1 + F_2 + - - + Fn\}] \quad (1)$$

n=Total no
T_1.....T_n Text in each frame of the video
A_1.....A_n Audio of each frame of the video
F_1.....F_n Number of frames in the video

$$C_{vs} = T_s \cup A_s \cup F_s \quad (2)$$

Where C_{vs} = Total Combined video Summary, which is the combination of T_s (Text Summary), A_s (Audio Summary), and F_s (Image summary). Final summary C_{vs} keeps the length of the summary to a minimum while omitting none of the crucial information from the original data. If we have original video V, then the length of Combined summary LC_{vs} is less than the length of original video summary LV_s.

$$|LC_{vs}| < |LV_s| \quad (3)$$

4. PROPOSED METHODOLOGY

Video summarization has evolved to address the challenges posed by the vast amounts of video data. Its primary objective is to identify and extract the relevant and significant content within a video [27].
For each video, we do the following:
1. Download the video transcript and then video

2. Apply the MAD (Mean Absolute Difference) algorithm on the video frames and find changes in frames and keyframes.
3. Apply the text summarization algorithm to the video transcript.
4. Extract the .wav file from the .mp4 file to extract the audio features.
5. Combine all the 3 types of Scores to get a unified score.
(The clips with the value above cut-off from that score is the summary)
The design of the proposed VSEM system consists of stages as shown in Figure 2.

4.1 Preparation of Text File

Text summarization systems extract brief information from a document. Using summarization techniques user can determine whether a document is relevant to his or her needs without reading the entire document.
There are two methods of producing automatic text summaries extractive and abstractive [28]. The extractive approaches evaluate each sentence's relevance before choosing the best-scoring ones with the least amount of redundancy. The methods for abstractive text summarization take the original text's location and extract its most important details. Abstractive techniques are more accustomed to the human summary, which is more precise, logical, and expressive. Since the captions dialogues cannot be altered in a video, extractive summarization is used. For the extractive summarization approach, a hybrid of text rank summarization and frequency summarization is implemented.
An unsupervised graph-based content extraction technique called text rank employs the Bag of Words via Word2Vec to give words a numerical value and then uses a cosine similarity matrix, a page rank implementation, and a sentence graph to assess the value of sentences. The disadvantage of text rank is that it excludes relevant keywords while recommending semantically similar phrases and avoids erroneous critical keywords in order to enhance rank [29].
The premise of frequency summarising is straightforward: sentences with high-frequency words in the paragraph, excluding stop words from the nltk toolkit, are rated highest. It is possibly the most straightforward and most often used summarization technique.
In the frequency summarization algorithm, a dictionary with the frequency of occurrence of that word is taken, ignoring the stop words from nltk. In the text rank summarization algorithm, a graph is initialized with the weights corresponding to similarity matrix values. For the hybrid, alongside the similarity matrix values, the occurrence frequency of the words is also considered as shown in Figure 3. In Hybrid summarization, frequency and text rank summarization are considered. Input is a transcript of the video, and then scores of each sentence are taken using text algorithms. After that, scaling is done as shown in Algorithm 1. Minimum and maximum are taken in scaling; after that, iterations are saved. Sentences are chosen with a threshold.

To verify the effectiveness of the above summarization method, it was tested against the Bert-extractive-summarizer from CNN-daily mail news text summarization.
Table1 depicts the average rouge score for the algorithms with Bert-extractive-summarizer for the first thousand instances in cnn-dailymail test data set, where the reference summary is highlighted. Here, it can be seen that the hybrid summary score (rounded up to the 4th decimal place) has the best value here as compared to its components of text rank summarization and frequency summarization, as well as

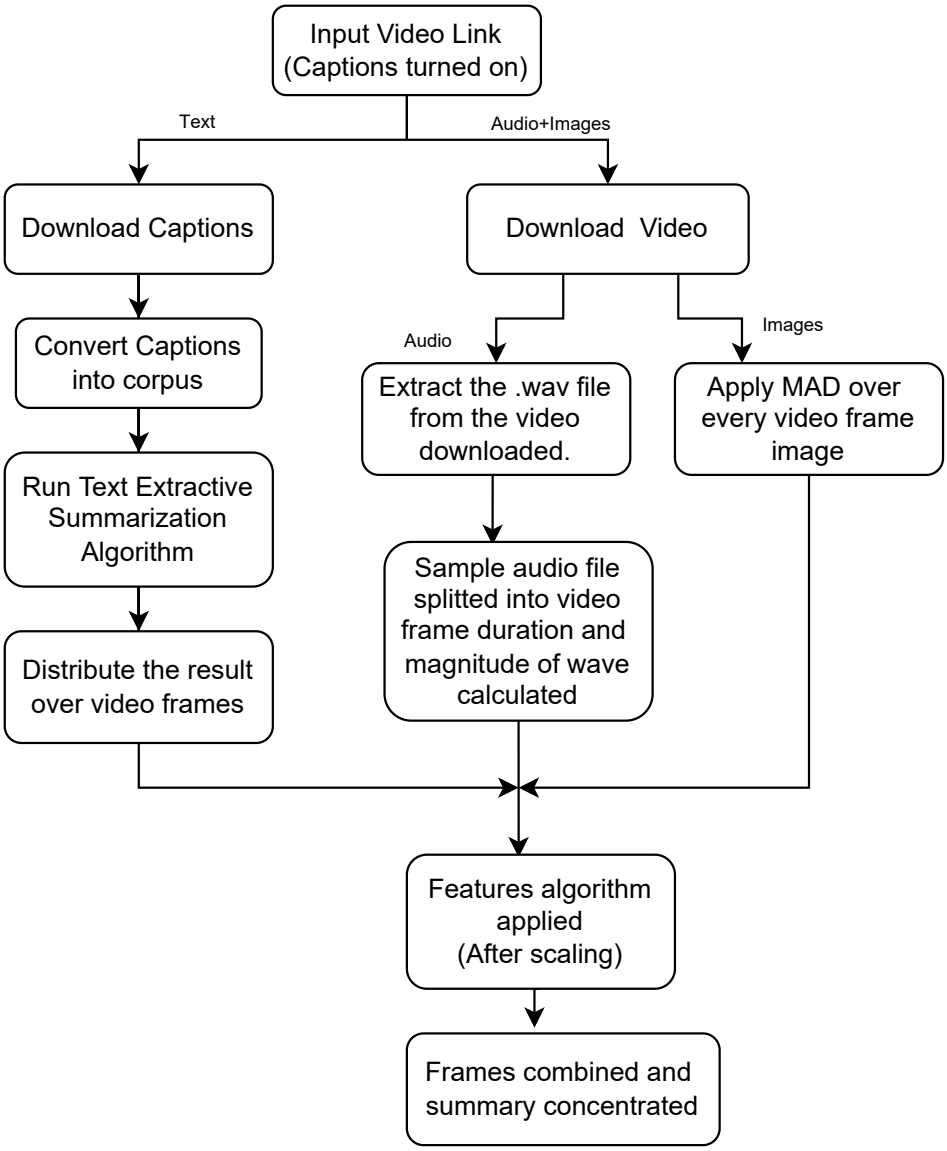

Fig. 2. Hybrid Summarization

BERT Extractive Text summarization with the threshold of 0.80. Figure 4 illustrates how different algorithms will produce different results for a sentence, with some being more focused on one aspect of the paragraph than others.

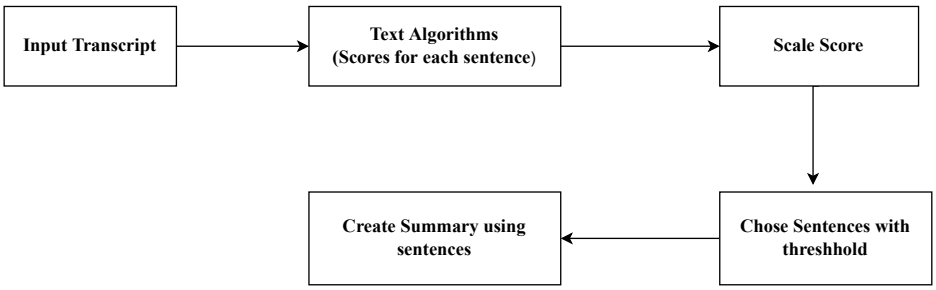

Fig. 3. Hybrid Summarization

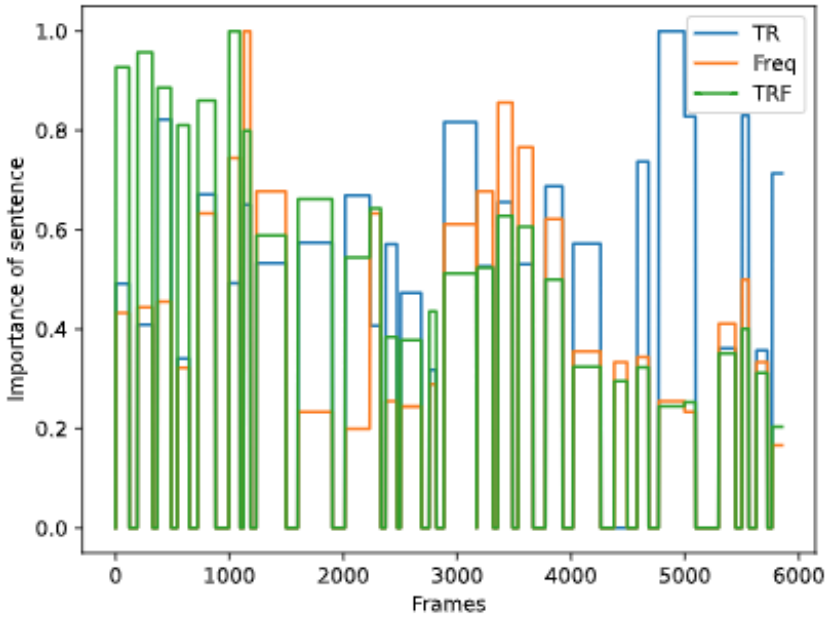

Fig. 4. Text Score by Hybrid Along With its Origin Algorithms

4.2 Audio Separation

Another crucial aspect of a video is its audio, which can be combined with its visual elements to create a powerful summary[30]. According to Coutrot et al. [31], the sound will affect viewer's visual attention while watching videos, and the strength of this influence varies over time. Subjects will glance in different directions with and without the audio, and the eye fixations gathered during the audio-visual test condition are more concentrated. In comparison to just visual features, audio-visual elements can produce

Algorithm 1: Scaling Algorithm

```
Data: Scale_3(lst:list): Input Variable Name - lst , type - list,
      Result: Scaling values
min_val = min(1st):                    //The current minimum value in the list
    if (min_val < 0) then
    | for i in range(len(lst)): do
    |   | lst[i] = lst[i] - min_val
    | end
end
min_val = min(lst)
max_val = max(lst)
if (max_val == min_val) then
    | for i in range(len(lst)): do
    |   | lst[i] = 0.5
    | end
    | return 1st
end
diff_val = max_val - min_val           //Gap between max val and min val
for i in range(len(lst)): do
  | lst[i] = (lst[i]-min_val)/diff_val
end
return 1st
```

Rouge	Frequency	Text Rank	Hybrid	Bert extractive Text Summarizer
Rouge-1	0.2854	0.2114	0.3207	0.2876
Rouge-2	0.1012	0.0563	0.1272	0.0998
Rouge-l	0.2577	0.1924	0.2955	0.2652

Table 1. Rouge Score

greater results.

To isolate the audio data from the video, we employed the MoviePy toolkit. As each presentation's timeline indicates the alignment of the video, audio and slides, the entire audio file is divided into a series of audio segments following the timing of the slide switch, ensuring that each audio clip is correctly aligned to a slides page, originally MP4 speech. After obtaining the wave file Mel Spectrogram, Mel Frequency Cepstral Coefficients, area of audio displayed, and audio amplitude are extracted from an audio chuck of length 0.10 sec (Persistence of Hearing is 0.10 sec.), then the value is distributed over frames. The time vs frequency graph over decibel for the MEL spectrogram of video is shown in Figure 5.

Mel spectrograms are used to align human auditory perception models by converting Hertz values to the Mel scale. The use of 0.10-second audio chunks is motivated largely by a desire to match hearing persistence, ensuring that the analysis catches sound properties that are perceptually meaningful to people. The time versus frequency graph varies

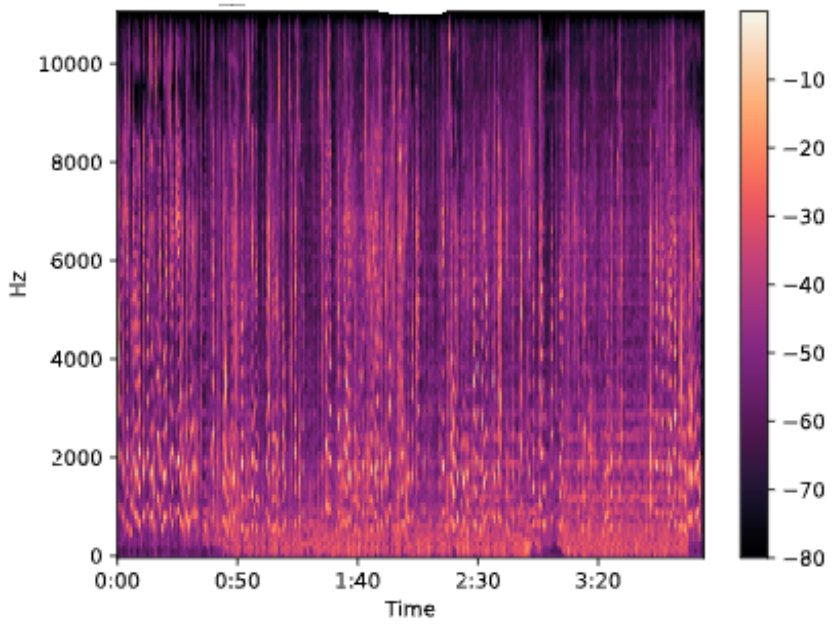

Fig. 5. Mel Power Spectrogram

over MFCC as shown in Figure 6.

For the images (Video Frames), it is to be noted that they can be classified into two types

1) 3D-like colored images which have pixel arrays of form x n x 3'

2) 2D-like grayscale or black-and-white images which only consist of either black pixels or white pixels and have an array of form x n'

In a video, since all the video frames are of the same size, it can be assured that all the video frames are of the same dimensions. Due to this postulate, instead of treating the video frames as an image, they can be treated as an n-dimension array. Finding alterations between two n-dimensional arrays is easier and more precise; For this Mean Absolute Difference (MAD) algorithm is used.

Due to the high frame rate, two adjacent frames are often practically a copy of the other, but on closer inspection, they are not, and every frame is distinct, some more than the other, and to find those distinctions, MAD is applied as shown in Figure 7.

The Mean Absolute Deviation (MAD) is then used to identify crucial frames in the video. Keyframes are distinguished by notable, rapid shifts in picture information. These modifications can include changes to the highlighted object, the introduction of additional objects, or other substantial adjustments. o pinpoint such substantial shifts, a threshold is applied to identify moments when the change experiences a sharp spike. This significant change can be identified by setting a threshold to see when the change spikes. This method

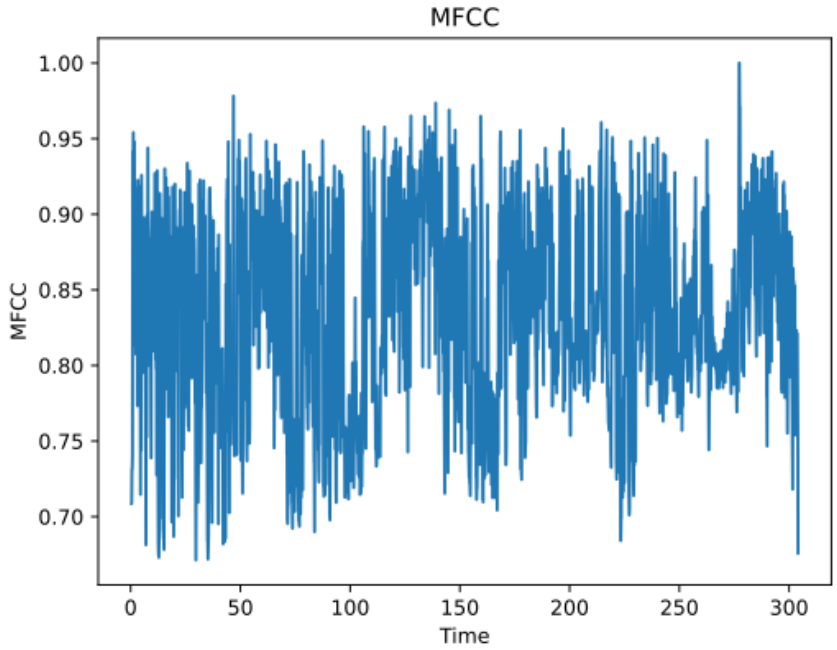

Fig. 6. MFCC Per 0.1 sec

is similar to Jenks natural breaks algorithm [32], which in itself is like a variation of the K-mean algorithm. Here, the Threshold is the sum of mean and standard deviation over MAD, and it is set as a classifier to detect if there is a keyframe. A higher frequency of keyframes denotes a higher degree of change and movement in the video, as shown in Figure 8.

5. Evaluation

5.1 Dataset

The experiment is carried out on the TV SUM dataset [33], which is a benchmark data set of video summarization. Fifty structured videos on ten distinct topics were acquired from YouTube for the TVSUM dataset. Videos are professionally edited about news, cookery, education, and others. All videos are of the length of fewer than 10 minutes. The shots are produced by evenly dividing the video into 2-second chunks, and 20 annotations of shot-level relevance scores are included. Figure 9 dataset shows a partial image of the data set.

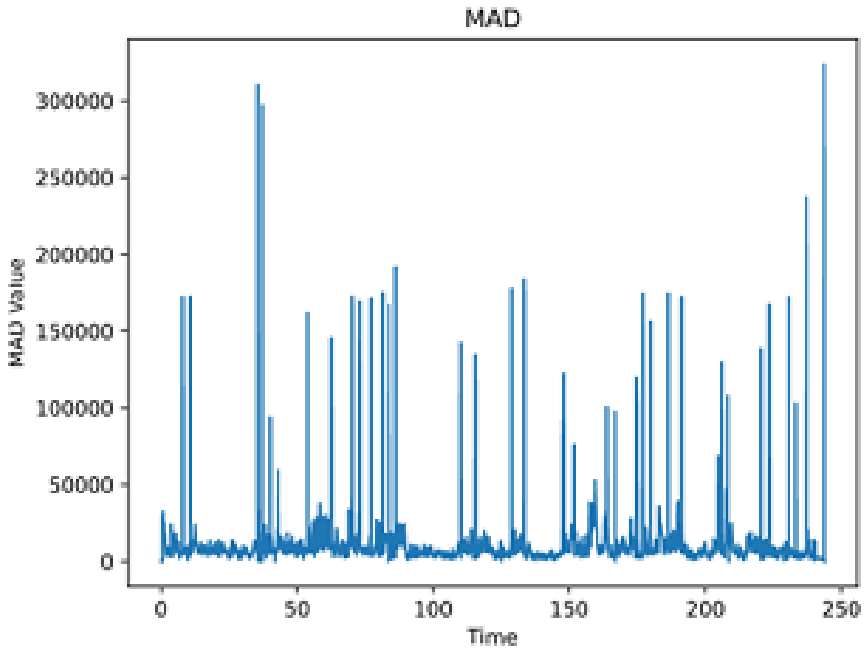

Fig. 7. MAD for Keyframe Identification

5.2 Evaluation Measure and Results

The TVSUM dataset, as previously mentioned, is used to assess the performance of our approach. ROUGE scores are used to evaluate the textual summary.

The dataset TVSUM is split in a ratio of 20:7 for training and testing, the first 20 for training and the latter for testing. In TV-SUM, for each video, there are 20 individual human evaluations given. Only the first human evaluation is picked up for all the summaries to avoid ambiguity in the model. The video compression threshold taken is 0.80.
The outcomes of VSEM were compared to those of the following video summarising techniques, which likewise use the TVSUM data set.

Calculating the F1 measure between the predicted and reference summaries is the most used evaluation strategy. Indicating which frames from the original movie are chosen for the summary, let y_i signify a label with the values 0 or 1 (y_i = 1 if the i-th frame is selected, otherwise 0). To evaluate the summary's quality, we compute the F-score (F1) as follows.

$$F1 = (TP+TN)/(TP+TN+FP+FN) \tag{4}$$

Where
F1 = Accuracy

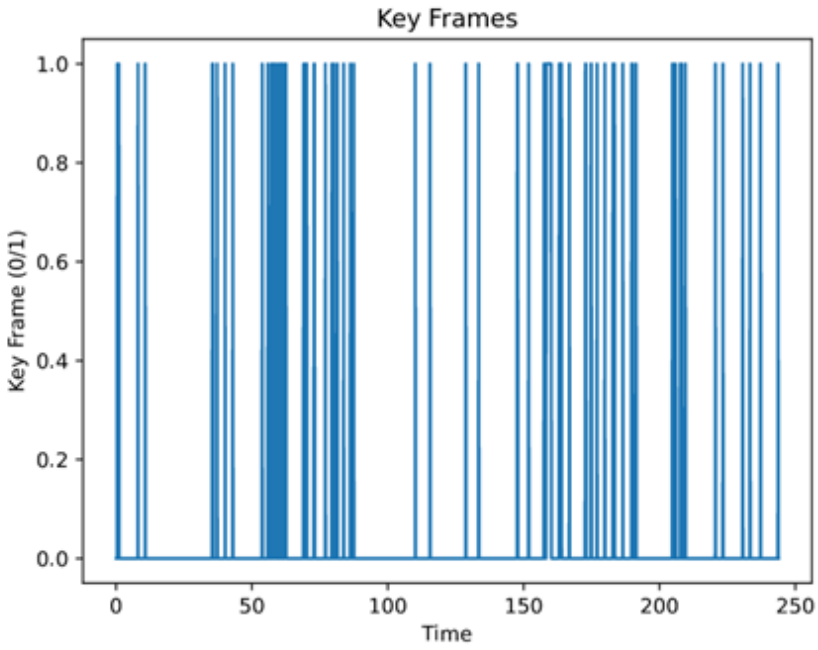

Fig. 8. Frequency of key frames

TP → True Positive, Frames selected by both predicted and human summary
TN → True Negative, Frames rejected by both predicted and human summary
FP → False Positive, Frames rejected by human summary but accepted by predicted summary.
FN → False Negative, Frames rejected by predicted summary but accepted by human summary.

Mean is the F1 score on TVSUM dataset.

Model	Maximum	Mean	Minimum
Linear Regression	0.73175	0.68529	0.65467
Stochastic Gradient Descent Regression	0.78787	0.69136	0.60714
Elastic Net Regression	0.70909	0.67458	0.64285
Ridge Regression	0.73333	0.68946	0.65467
Lasso Regression	0.70909	0.67344	0.64285
Random Forest Regression	0.71794	0.67584	0.62524
Gradient Boosting Regression	0.73214	0.68542	0.64891

Table 2. Average F-measures using different models

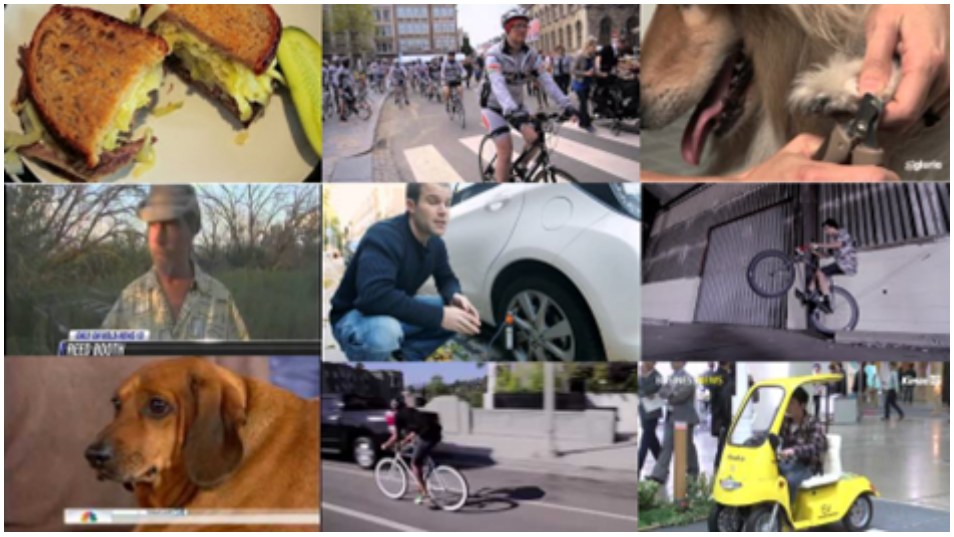

Fig. 9. Partial Image of Dataset

The F1 score using different models is shown in Table 2. F1 score is found on the dataset video and from the results, it is found that Stochastic Gradient Descent Regression is the most suitable Regression available, with the highest mean score of 0.69136 (out of 1.0) and ridge regression with a mean score of 0.68946(out of 1.0). So, we have used a combination of these two regression algorithms in the hybrid model. SVM and decision tree classifiers are not used in this. SVM performs worse when more features than training sets are available; hence, it is not appropriate for large datasets. Due to the abundance of trees, the performance of the summarization approach employing the decision tree classifier is poor. So, even a minor alteration to the decision tree could significantly impact prediction accuracy. The F-measures for each approach using a machine learning model for the video's summary in the database are shown in Table 3. The table shows that VSEM outperforms the assessed methodologies, delivering competitive outcomes while retaining a balance between pace, duration, and quality.

The proposed algorithm is applied to different videos of the TVSUM dataset, and it is

Method	Year	F-Measure
M-AVS [34]	2017	61.0
VASNet [35]	2018	61.42
DSNet [36]	2020	62.1
PGLSUM [37]	2021	61.0
RRSTG [38]	2022	63.0
VSEM	**2022**	**69.6**

Table 3. Average F-measures of the summaries generated by each technique

found that there is a significant difference in the run time of the original video and video summary. The original video was 5 minutes and 54 seconds long, while the summarised

video is 72 seconds long. Instead of watching the entire video, the consumer may see the summary, which saves time. It is found that VSEM saves 75-80 percent of users' time.

Conclusion

To improve the effectiveness of the summary, a hybrid model (VSEM) for the video summarising problem is proposed in this study. Hybrid text summarization is proposed using text and frequency summarization and is compared with the state-of-the-art methods. The proposed hybrid text summarization shows better results. Audio and image features are combined with text, and a hybrid model is proposed. The experimental results on TVsum show that the multimedia components—text, audio, and image—can offer the summary task more information and accuracy than a single visual feature. The simulation results show that the suggested model outperforms the existing techniques in terms of F1 Score (69.6).
In our future work, we will investigate more advanced attention mechanisms to gain more contextual information, like some frames containing textual information in the form of hoardings, boards, etc. that will help to generate a better summary.

REFERENCES

1. F. Amato, A. Castiglione, V. Moscato, A. Picariello, and G. Sperlì, "Multimedia summarization using social media content," *Multimedia Tools and Applications*, Vol. 77, no. 14, Jan. 2018, pp. 17 803–17 827. [Online]. Available: https://doi.org/10.1007/s11042-017-5556-2
2. A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, Vol. 35, no. 2, Apr. 2015, pp. 137–144. [Online]. Available: https://doi.org/10.1016/j.ijinfomgt.2014.10.007
3. P. G. Shambharkar and M. N. Doja, "Movie trailer classification using deer hunting optimization based deep convolutional neural network in video sequences," *Multimedia Tools and Applications*, Vol. 79, no. 29-30, May 2020, pp. 21 197–21 222. [Online]. Available: https://doi.org/10.1007/s11042-020-08922-6
4. M. Kini and K. Pai, "A survey on video summarization techniques," in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Vol. 1. IEEE, Mar. 2019, pp. 1–5. [Online]. Available: https://doi.org/10.1109/i-pact44901.2019.8960003
5. C. Panagiotakis, H. Papadakis, and P. Fragopoulou, "Personalized video summarization based exclusively on user preferences," *Lecture Notes in Computer Science*. Springer International Publishing, 2020, pp. 305–311. [Online]. Available: https://doi.org/10.1007/978-3-030-45442-538
6. A. Bhowmik, S. Kumar, and N. Bhat, "Evolution of automatic visual description techniques-a methodological survey," *Multimedia Tools and Applications*, Vol. 80, no. 18, May 2021, pp. 28 015–28 059. [Online]. Available: https://doi.org/10.1007/s11042-021-10964-3

7. L. Jin, Z. Li, and J. Tang, "Deep semantic multimodal hashing network for scalable image-text and video-text retrievals," *IEEE Transactions on Neural Networks and Learning Systems*, 2020, pp. 1–14. [Online]. Available: https://doi.org/10.1109/tnnls.2020.2997020
8. J. Dong, X. Li, W. Lan, Y. Huo, and C. G. Snoek, "Early embedding and late reranking for video captioning," in *Proceedings of the 24th ACM international conference on Multimedia*, Oct. 2016. [Online]. Available: https://doi.org/10.1145/2964284.2984064
9. A. S. Murugan, K. S. Devi, A. Sivaranjani, and P. Srinivasan, "A study on various methods used for video summarization and moving object detection for video surveillance applications," *Multimedia Tools and Applications*, Vol. 77, no. 18, Jan. 2018, pp. 23 273–23 290. [Online]. Available: https://doi.org/10.1007/s11042-018-5671-8
10. M. V. M. Cirne and H. Pedrini, "VISCOM: A robust video summarization approach using color co-occurrence matrices," *Multimedia Tools and Applications*, Vol. 77, no. 1, Jan. 2017, pp. 857–875. [Online]. Available: https://doi.org/10.1007/s11042-016-4300-7
11. X. Lin, G. Bertasius, J. Wang, S.-F. Chang, D. Parikh, and L. Torresani, "Vx2text: End-to-end learning of video-based text generation from multimodal inputs," 2021, pp. 7005–7015.
12. H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 31, no. 5, May 2019, pp. 996–1009. [Online]. Available: https://doi.org/10.1109/tkde.2018.2848260
13. K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *The Visual Computer*, Vol. 38, no. 8, Jun. 2021, pp. 2939–2970. [Online]. Available: https://doi.org/10.1007/s00371-021-02166-7
14. B. T. Truong and S. Venkatesh, "Generating comprehensible summaries of rushes sequences based on robust feature matching," in *Proceedings of the international workshop on TRECVID video summarization - TVS '07*. ACM Press, 2007. [Online]. Available: https://doi.org/10.1145/1290031.1290036
15. N. Modani, P. Maneriker, G. Hiranandani, A. R. Sinha, Utpal, V. Subramanian, and S. Gupta, "Summarizing multimedia content," *Web Information Systems Engineering – WISE 2016*. Springer International Publishing, 2016, pp. 340–348. [Online]. Available: https://doi.org/10.1007/978-3-319-48743-427
16. P. G. Shambharkar and R. Goel, "Analysis of real time video summarization using subtitles," in *2021 International Conference on Industrial Electronics Research and Applications (ICIERA)*. IEEE, Dec. 2021. [Online]. Available: https://doi.org/10.1109/iciera53202.2021.9726769
17. M. S. Nair and J. Mohan, "VSMCNN-dynamic summarization of videos using salient features from multi-CNN model," *Journal of Ambient Intelligence and Humanized Computing*, Jun. 2022. [Online]. Available: https://doi.org/10.1007/s12652-022-04112-4
18. S. Yeung, A. Fathi, and L. Fei-Fei, "Videoset: Video summary evaluation through text," *arXiv preprint arXiv:1406.5824*, 2014.

19. Y. Li and B. Merialdo, "VERT," in *Proceedings of the international conference on Multimedia - MM '10*. ACM Press, 2010. [Online]. Available: https://doi.org/10.1145/1873951.1874095
20. G. B. Martins, J. P. Papa, and J. Almeida, "Temporal-and spatial-driven video summarization using optimum-path forest," in *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, Oct. 2016. [Online]. Available: https://doi.org/10.1109/sibgrapi.2016.053
21. M. Merler, K.-n. C. Mac, D. Joshi, Q.-b. 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. PP, no. c, 2018, p. 1.
22. S. Huang, X. Li, Z. Zhang, F. Wu, and J. Han, "User-Ranking Video Summarization with Multi-Stage Spatio-Temporal Representation," Vol. XX, no. X, 2018, pp. 1–11.
23. A. Javed, A. Irtaza, H. Malik, M. T. Mahmood, and S. Adnan, "Multimodal framework based on audio-visual features for summarisation of cricket videos," Vol. 13, 2019, pp. 615–622.
24. B. Zhao, M. Gong, and X. Li, "AudioVisual video summarization," *IEEE Transactions on Neural Networks and Learning Systems*, 2021, pp. 1–8. [Online]. Available: https://doi.org/10.1109/tnnls.2021.3119969
25. A. Savelieva, B. Au-Yeung, and V. Ramani, "Abstractive summarization of spoken and written instructions with bert," *arXiv preprint arXiv:2008.09676*, 2020.
26. P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, Vol. 9, 2021, pp. 108 069–108 082. [Online]. Available: https://doi.org/10.1109/access.2021.3100462
27. K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *Journal of Big Data*, Vol. 6, no. 1, Oct. 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0254-8
28. R. Rani and D. K. Lobiyal, "An extractive text summarization approach using tagged-LDA based topic modeling," *Multimedia Tools and Applications*, Vol. 80, no. 3, Sep. 2020, pp. 3275–3305. [Online]. Available: https://doi.org/10.1007/s11042-020-09549-3
29. S. Sumana, "Towards automatically generating release notes using extractive summarization technique," in *Proceedings of the 33rd International Conference on Software Engineering and Knowledge Engineering*. KSI Research Inc., Jul. 2021. [Online]. Available: https://doi.org/10.18293/seke2021-119
30. G. Wu, S. Wang, and L. Liu, "Fast video summary generation based on low rank tensor decomposition," *IEEE Access*, Vol. 9, 2021, pp. 127 917–127 926. [Online]. Available: https://doi.org/10.1109/access.2021.3112695
31. A. Coutrot and N. Guyader, "Toward the introduction of auditory information in dynamic visual attention models," in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, Jul. 2013.
32. N. Khamis, T. C. Sin, and G. C. Hock, "Segmentation of residential customer load profile in peninsular malaysia using jenks natural breaks," in *2018 IEEE 7th International Conference on Power and Energy (PECon)*. IEEE, Dec. 2018. [Online]. Available: https://doi.org/10.1109/pecon.2018.8684113

33. Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2015. [Online]. Available: https://doi.org/10.1109/cvpr.2015.7299154
34. Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder–decoder networks," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 30, no. 6, Jun. 2020, pp. 1709–1717. [Online]. Available: https://doi.org/10.1109/tcsvt.2019.2904996
35. J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," *Computer Vision – ACCV 2018 Workshops*. Springer International Publishing, 2019, pp. 39–54. [Online]. Available: https://doi.org/10.1007/978-3-030-21074-8-4
36. W. Zhu, J. Lu, J. Li, and J. Zhou, "DSNet: A flexible detect-to-summarize network for video summarization," *IEEE Transactions on Image Processing*, Vol. 30, 2021, pp. 948–962. [Online]. Available: https://doi.org/10.1109/tip.2020.3039886
37. E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras, "Combining global and local attention with positional encoding for video summarization," in *2021 IEEE International Symposium on Multimedia (ISM)*. IEEE, Nov. 2021. [Online]. Available: https://doi.org/10.1109/ism52913.2021.00045
38. W. Zhu, Y. Han, J. Lu, and J. Zhou, "Relational reasoning over spatial-temporal graphs for video summarization," *IEEE Transactions on Image Processing*, Vol. 31, 2022, pp. 3017–3031. [Online]. Available: https://doi.org/10.1109/tip.2022.3163855

Dr. Prashant Giridhar Shambharkar has completed his B.E. From Amravati University, M.Tech From RGPV, Bhopal and Ph. D. from Jamia Millia Islamia, New Delhi, and is Working as an Assistant Professor in Department of Computer Science & Engineering having 15+ years teaching experience of various computer science and IT related subjects, worked in various committee at responsible position, member of ISO 9001-2008 member at the institution level, Dy. Chairperson Admission for B.Tech Programme under Continuing Education for Working Professionals, Dy. Coordinator Ph.D. Admissions, Coordinator M.Tech Admissions. Area of research includes Data Mining, Real-Time Systems, and Mobile health monitoring.

Ms. Ruchi Goel has done B.Tech in Computer Science and engineering in 2003 from MDU Rohtak and done master's degree in Computer science from Delhi College of Engineering in 2011. Currently pursuing a Ph.D. from DTU (Formerly Delhi College of Engineering). 18+ years of teaching Experience. Area of interest includes Software Testing, Web Mining, Artificial Intelligence, and Computer Vision.