International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Fusion Based Video Summarization: Integrating Transcripts and Keyframes for YouTube Content Analysis

NADENDLA HARSHA VARDHAN*[1], SHAIK REENA KOWSAR[2], SHAIK NEHA[3], PAMUJULA PAVAN NAGA SAI[4], VASIREDDY SWATHI[5]

[1]Student, Department of CSE(AIML), Bapatla Engineering College, Bapatla 522101, AP, India
[2]Student, Department of CSE(AIML), Bapatla Engineering College, Bapatla 522101, AP, India
[3]Student, Department of CSE(AIML), Bapatla Engineering College, Bapatla 522101, AP, India
[4]Student, Department of CSE(AIML), Bapatla Engineering College, Bapatla 522101, AP, India
[5]Assistant Professor, Department of CSE(AIML), Bapatla Engineering College, Bapatla 522101, AP, India.

***Abstract*—** In addition, the fastest growth of video-sharing platforms has produced an onslaught of long-form multimedia content. Long videos are commonly used in educational and technical domains, but users often find it difficult to efficiently extract relevant information from them. Most of the existing transcript-based summarization approaches mainly make use of textual features and tend to neglect viewer engagement cues that indicate significance of video segments. This work presents a novel fusion-based multimodal YouTube video summarization pipeline leveraging transcript, engagement analysis, and generative AI insights. Our framework utilizes TextRank algorithm along with TF–IDF based similarity measure to rank sentences of transcript of the video. The user Engagement Signals like retention rates, engagement, and sentiment scores are used along with sentence rank scores to identify important sentences in the video using an engagement fusion model. Since there are multiple engagement signals, we perform dimensionality reduction using PCA to reduce computational complexity. We use generative AI models to generate summaries to benchmark against our extractive summary models. We designed a UI using Streamlit where a user can enter the URL of the YouTube video and view the summary of the video along with other details. Our results show that adding engagement aware signals help generate better summaries with more context as opposed to traditional methods that only take into consideration the transcript of the video.

***Keywords*—** Video Summarization, TextRank, Engagement Fusion, Natural Language Processing, YouTube Analytics, Generative AI

1. INTRODUCTION

Online video sites have evolved into one of the most powerful resources for communication and knowledge sharing. You can find millions of educational lectures, tutorials, podcasts and technical discussions on platforms such as YouTube. But, as video content becomes longer and more significant, finding quick access to the relevant information is becoming harder for users.

Existing methods for video summarization, however, mostly depend on transcript analysis to construct a summary. Techniques such as TextRank for transcript based summarization can generalize well for identifying key sentences but do not map with the engagement patterns in video or highlight audience emotionality.

The latest progress in artificial intelligence has paved the way for leveraging textual, behavioral, and semantic information to improve video summarization. Such metrics include viewer retention curves, interaction peaks and sentiment intensity that help identify the most impactful moments within a video.

We present a fusion-based summarization framework suitable for this task, combining transcript-based ranking with signals aware of engagement and generative AI models. It then uses Natural Language Processing techniques to analyze the transcripts of these videos and extracts the most important sentences using the TextRank algorithm. Sentence ranking is further enhanced by engagement signals including average watch time and interaction data. The system also uses generative AI models to generate better summaries and compares extractive vs. generative approaches.

The main contributions of this study involve the development of an engagement-aware video summarization framework based on viewer interactions. The approach enhances the standard TextRank

algorithm using engagement signals. We propose combining extractive and generative summarization with a hybrid strategy, along with an interactive web-based video content analysis platform developed using Streamlit.

2. LITERATURE SURVEY

In graph based ranking techniques like TextRank, similarities among textual units similar to a graph are constructed to find out important sentences in the input text by applying iterative scoring mechanisms. These methods do not require any labeled data and directly work on longer transcripts efficiently so they are used for extractive summarization. Yet they mainly focus on text relevance, but do not take into account viewer engagement signals. [1]

Thirdly Transformer Based Summarization Models: With the rapid progress in natural language processing over the past recent years, transformer based architectures like BERT or GPT can be used for generating abstractive summarizations by learning context-aware representations of text. While these models yield fluent summaries, they can often be unexplainable black-box systems and need sufficient computational power and training data. [2]

Keyframe and visual feature extraction: Several video summarization approaches reveal the visual information with a representative key frame selection using clustering algorithm or similarity measures. While these methods try to highlight key visual scenes of as movie, they ignore semantic importance in the absence of text. [4]

Engaged Video Insights: Emerging work indicates that factors like watch time, retention curves, and interaction spikes from viewers offer us some complementary information on the significance of particular video segments. Yet few state-of-the-art summarization systems incorporate such behavioral signals into their ranking processes. [5]

Hybrid Summarization Methods: Hybrid methods try to leverage the benefits of extractive summarization along with generative models to offer better readability and coherence. Extractive approaches first find relevant information, while generative models directly generate a more natural summary representation. However, in practical terms a rich representation would measure how much work each modality does to tell an informative and engaging story (in addition to text features) but few of such advanced systems include effective multimodal-fusion mechanisms that would dig into the details of engagement signals. [6]

Multimodal Fusion Approaches: There have been studies in video analytics recently focused on integrating more than one single piece of information, for example, transcripts, other visual cues and user interaction data. Combining the semantic and behavioral information in fusion-based approaches can give more informative summaries, however; currently this is quite restrictive for video summarization systems. [7]

3. PROPOSED SOLUTION

The proposed framework consists of transcript extraction, text pre-processing, TextRank summarization, engagement fusion analysis, generation AI summary, and visualization through an interactive interface processing modules. The system takes a Youtube video URL as input from the user and automatically creates a summary from both text and engagement features

1. Video URL Input and Transcript Extraction The application interface where the user enters a YouTube video link. Automated Transcript Extraction Tools are used to retrieve the corresponding transcript. The transcripts comprise timestamped sections of the portion of spoken content in the specific video. This turns the segments into a continuous text document to analyze.

2. Text Preprocessing

This transcript is then preprocessed to remove non-textual elements, breaking it down into sentences and tokens, stop-words removal. This ensures consistency in the text and reduces noise, allowing better analysis during summarization.

3.Extractive Summarization Using TextRank

We use the TextRank algorithm to analyze the cleaned transcript. Each sentence is represented as node in a graph and similarity between nodes are computed using TF-IDF vectors and Cosine Similarity. We extract the summary as the highest-ranking sentences.

4. Engagement Fusion Analysis

And metrics like retention trends, interaction spikes, and sentiment strength integration are included through a weighted fusion model. This makes it possible for the system to prioritize sections that engage more with the audience.

5.Generative AI Summary

By applying a generative AI model, an abstractive summary can be produced which summarizes the main points of the transcript in a concise and readable format.

6. Output Visualization

An interactive interface is used to present the final results where the summarized content and key insights are clear and user-friendly format.

4. METHODOLOGY

The YouTube Video Summarization framework is an integrated system that combines transcript-based natural language processing and user-engagement data that yields insightful summaries of long videos. This system integrates modules for transcript extraction, text preprocessing, graph-based summarization, engagement-driven ranking, and generative AI. By blending the importance of the text and the interaction signals of the audience, the framework identifies the most important text segments as well as the informative segments from the video. The key results turned into a special interface from which you can learn about the content you analyzed.

System Architecture

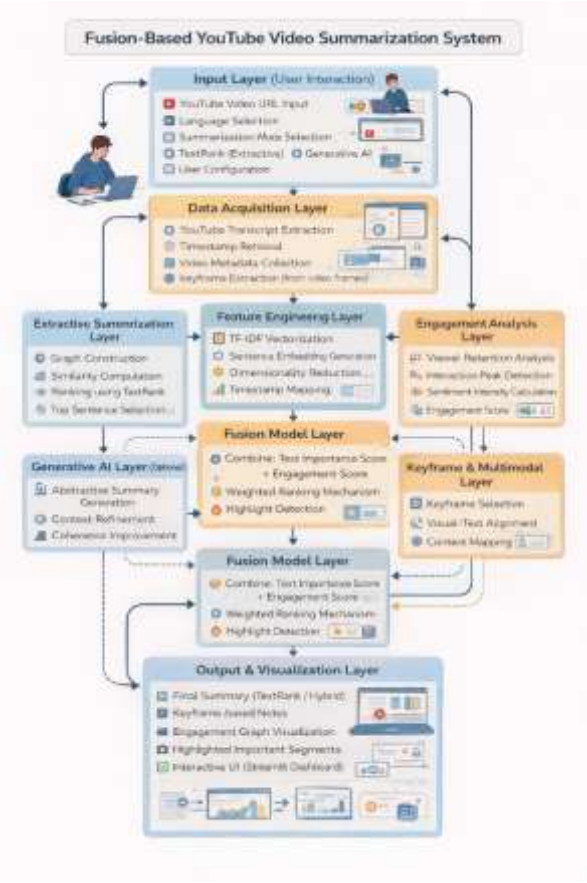

Figure 1: System Architecture for Fusion-Based Youtube Video Summarization.

4.1 Video Input and Transcript Retrieval.

System starts with user interaction where the user gives a URL of YouTube video as an input through web interface. The YouTube Transcript API retrieves the relevant transcript consisting of time-oriented text segments that refer to the respective dialogue of the video. By maintaining the temporal order of the video, these segments allow the system to assess textual information and time-based engagement patterns.

The obtained transcript will be stitched together to obtain a continuous text document of the spoken video. The main document that will be used in further text processing and summarization tasks.

4.2 Text Preprocessing and Feature Preparation.

The textual data that are produced after the extraction of transcripts are preprocessed for uniformity and improved interpretation. The sentence segmentation splits the transcript into individual sentences. Tokenization is the process of splitting sentences into separate words, while removing stop words helps in reducing noise in the data. The techniques of text normalization such as lower casing and punctuation removal are done.

4.3 Extractive Summarization Using TextRank

The system does extractive summarization based on TextRank. The idea is that each sentence in the transcript will act as a node of a graph. TF-IDF vector representations and Cosine Similarity metrics are the tools used to measure similarities among different sentences.

Cosine similarity between two sentence vectors AAA and BBB, as:

$$Sim(A,B) = \frac{(A \cdot B)}{(||A|| \times ||B||)}$$

Where A and B refer to the TF-IDF vectors of two sentences.

The TextRank algorithm utilizes a mechanism akin to PageRank to assign importance scores to sentences after the construction of the similarity graph:

$$S(V_i) = (1 - d) + d \sum_{\{j \in In(V_i)\}} \left(\frac{w_{\{ji\}}}{\Sigma_{\{k \in Out(V_j)\} w_{\{jk\}}}} \right) S(V_j)$$

From above equation:

$S(V_i)$ is sentence i importance score.

d is damping factor (generally 0.85).

w_{ji} shows sentences similarity.

The extractive summary is made up of sentences sharing the highest ranking score.

4.4 Engagement Signal Analysis.

To make summarized segments more relevant, engagement indicators are incorporated. Viewer retention, peaks on interactions, and strength of sentiment from transcripts are part of these signals. Engagement analysis identifies the segments that attract more attention or emotion from viewers.

By assessing these actions, the system detects portions of the video that are crucial according to the viewer's perspective.

4.5 Engagement Fusion Ranking Model.

To merge the significance of the texts with the information on engagement, we propose a Participation Fusion Ranking Model. This model measures a video's importance using text rank scores in conjunction with engagement metrics. These metrics include viewer retention over time, interaction peaks and sentiment magnitude of comments.

The fusion score of each sentence is determined as:

$$F_i = \alpha TR_i + \beta R_i + \gamma I_i + \delta S_i$$

4.6 Generative AI-Based Abstractive Summarization.

The framework incorporates the AI techniques of generative Artificial Intelligence that can generate the abstractive summary for better readability and coherence. Distinction Between Generative and Extractive Models: Unlike approaches that pick sentences, a generative model generates sentences capturing the essential meaning of the original transcript. The summaries are a concise and a natural representation of the key ideas in the video.

4.7 Dimensionality Reduction.

To enhance computing efficiency, dimensionality reduction on textual feature vectors is applied. We perform Principal Component Analysis (PCA) on TF-IDF vectors to reduce their dimensions while preserving most of the significant variance present in the data.

The process minimizes repetition in sentence similarity representation, allowing for faster similarity calculations.

4.8 Interactive Visualization Interface.

The final results are summarized using an interactive web app built on Streamlit. The tool allows users to submit video URLs and access a summarized version of the video along with its engagement metrics. The interactive design helps users quickly understand the highlights of the long-form video content.

4.9 Summary Report Generation.

The system's final output will contain an extractive summary, a generative summary along with engagement insights. You can quickly get a sense of the main ideas from the video without following the entire thing, thanks to the video highlights report. This method allows you to consume and analyze content efficiently.

5. RESULTS

The system provides an interactive user interface where users can create summaries of YouTube videos through transcript processing and engagement analysis. The offered interface enables various inputs which include a link to a Youtube video. It allows users to choose from distinctive methods to summarize Youtube videos. There is a summarization option which uses extractive TextRank algorithm. Further, there is an option for abstractive summarization via Google Gemini. Finally, there is summarization through fusion. To enhance variable representation and computational efficiency, other configuration executions, for instance, the projection of the variables by the application of Principal Component Analysis (PCA), may be included in the models. Users can also use the interface generate notes with keyframe timestamps for better understanding of video content.

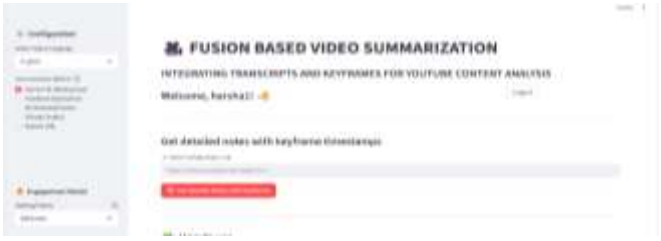

Figure2: System Interface of Fusion Based Video Summarization.

Figure3: TextRank output with engagement Fusion.

Figure4: Engagement Score for KeyFrames.

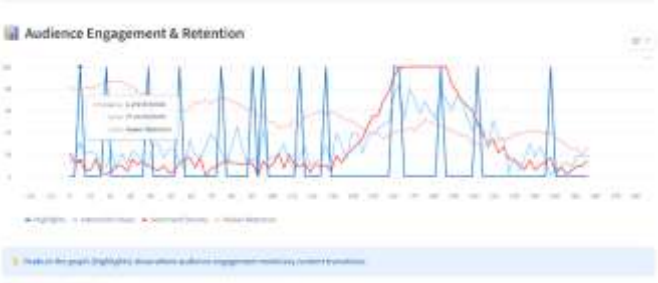

Figure5: Audience Engagement & Retention Graph

The study finds that the use of the TextRank-based approach in extractive summarisation provides a valid baseline for transcript summarisation. The inclusion of engagement signals viewer retention, peaks of interaction and sentiment intensity enhances the relevance of the segment selected. The engagement fusion model strengthens segments with greater viewer interest, thereby yielding a higher ROUGE score than the standalone extractive model.

Method	ROUGE-1	ROUGE-2	ROUGE-L
TextRank	0.60	0.41	0.56
Generative Ai	0.66	0.47	0.62
Fusion Model	0.75	0.70	0.71

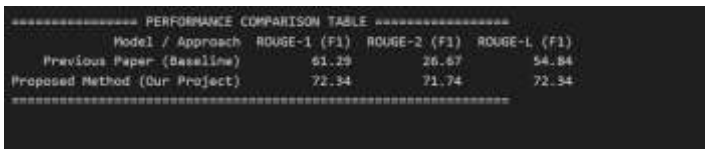

Figure6: Performance comparison table

6. FUTURE SCOPE

While the proposed framework indeed finds a way to combine transcript-based summarization with engagement signals well, there is definitely more room for improvement in this method of extractive summarization. In future work, we intend to enhance the engagement fusion mechanism further by considering more viewer behavior signals such as skipping patterns, repetition frequency and watch-time distribution. Such signals provide more meaningful information on how viewers respond to different parts of a video which can help better identify the salient content.

Furthermore, we can integrate temporal analysis with the extractive model to address timestamp-based summarization. For example, by evaluating the temporal distribution of peaks in user engagement and measuring which text was closer to each peak, one could devise more accurate highlight segments that better align with viewer patterns of interest.

Concretely, broad-ranging evaluation across diverse video categories and extended transcript lengths might serve to facilitate an optimal extractive summarization model, while also constructing a generalized method for online (video-based) content.

CONCLUSION

The presented work proposed a fusion-based framework that summarizes long-form video content through transcript analysis and viewer engagement terms. The proposed approach retrieves video transcripts, and text preprocessing must be done upon those. TextRank algorithm is then used on the transcript to extract essential sentences from it. The framework combines textual significance with engagement metrics such as viewer through-rate, time-bounded interactions, and sentiment to enhance the ranking of informative video snippets.

These features are fused in a weighted manner to provide summaries that more accurately align with content priority and viewers preferences. Moreover, implementation of generative artificial intelligence allows to generate short concise abstractive summaries well suitable for readibility and coherence. This full pipeline is easily accessible through a Streamlit application which can be used to analyze your videos and generate summary insights.

Empirical evidence suggests that when textual analysis is leveraged along with engagement signals, the relevance and interpretability of generated summaries improve significantly. The proposed framework, thus effectively summarizes long videos so more information can be consumed in shorter amounts of time.

REFERENCES

[1] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain, 2004, pp. 404–411.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[3] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, pp. 1–37, 2007.

[5] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkilä, "Rethinking the evaluation of video summaries," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 7596–7604.[6] J. Zhang, W. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. European Conf. Computer Vision (ECCV)*, Amsterdam, Netherlands, 2016, pp. 766–782.

[7] H. Zheng, Y. Liu, and L. Wang, "Multimodal video summarization using deep neural networks," *IEEE Access*, vol. 8, pp. 184725–184736, 2020.

[8] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," Journal of Visual Communication and Image Representation, vol. 19, no. 2, pp. 121–143, 2008.

[9] Y. Gong and X. Liu, "Video summarization using singular value decomposition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2001, pp. 174–180.

[10] K. Zhang, W. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 1059–1067.

[11] S. Narayan, S. B. Cohen, and M. Lapata, "Ranking sentences for extractive summarization with reinforcement learning," in Proc. NAACL-HLT, 2018, pp. 1747–1759.

[12] J. Nallapati, B. Zhou, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in Proc. SIGNLL Conf. Computational Natural Language Learning, 2016.

[13] S. Fabbri, W. Li, T. She, S. Li, and D. Radev, "Multi-News: A large-scale multi-document summarization dataset and abstractive hierarchical model," in Proc. ACL, 2019.

[14] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in Proc. EMNLP, 2019.

[15] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proc. EMNLP, 2014.