IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AUTOMATED VIDEO SUMMARIZATION SYSTEM

Miss. Anushka Babaso Patil [1], Miss. Revati Rajaram Nimbalkar [*2], Mr. Rushikesh Mahesh Hirphode[*3], Mr. Onkar Nitin kadam [*4], Mrs. Trupti H. Mulik [*5]

[*1, 2, 3, 4] Students, [*5]Assistant Professor
Department of CSE (Data Science),
D. Y. Patil College of Engineering and Technology, Kolhapur, India

Abstract: The Automated Video Summarization System uses artificial intelligence, machine learning, and automatic speech recognition (ASR) to transform the way users engage with video content, with the goal of enhancing accessibility, efficiency, and personalization. The goal of proposed system is to automatically analyze video data to generate concise summaries while supporting real-time transcription, language translation, and customizable output formats. By integrating cross-platform support and advanced extension features such as custom summary lengths, voice and language preferences, and adaptive thematic styles, the system ensures a highly personalized user experience. It also introduces interactive capabilities like automated question generation to improve content comprehension and retrieval. The proposed system enables smarter video consumption by offering multilingual, context-aware, and user-driven summarization, thereby redefining how educational, corporate, and entertainment content is consumed across digital platforms.

Index Terms - Video Summarization, Machine Learning, Artificial Intelligence, Automatic Speech Recognition (ASR), Natural Language Processing, Multimedia Processing, Deep Learning, Personalization.

I. INTRODUCTION

The rapid growth of video content across digital platforms has made videos a primary medium for communication, education, and entertainment. However, consuming lengthy videos is time-consuming and often leads to reduced engagement and difficulty in extracting key information. Traditional methods of video consumption are largely passive and inefficient, requiring users to manually search or skim through content to identify relevant parts.

Although some existing video summarization systems at- tempt to address this issue, they still suffer from several limitations. Many systems are platform-dependent and work only for specific sources, limiting their flexibility. In addition, most approaches rely on basic keyword extraction rather than understanding the actual meaning of the content, resulting in less accurate summaries. The lack of multilingual support and interactive features further reduces their effectiveness for diverse users.

To overcome these challenges, the Automated Video Summarization System is proposed as an intelligent and user- centric solution that leverages Artificial Intelligence (AI), Machine Learning (ML), and Automatic Speech Recognition (ASR). The system automatically analyzes and summarizes video content into concise and meaningful outputs, enabling users to quickly understand key information without watching the entire video.

The proposed system supports both video uploads and external links, ensuring cross-platform flexibility. It also includes features such as multilingual translation, customizable summary formats, and interactive capabilities to enhance user engagement. By integrating these functionalities, the system improves accessibility, reduces time consumption, and trans- forms traditional passive video viewing into a more active, efficient, and user-friendly experience across various domains.

II. LITERATURE SURVEY

A survey “Video Summarization Using Deep Neural Net- works” by Apostolidis et al.[1] discusses the development of deep learning techniques for video summarization. It explains how models use CNNs for feature extraction and RNNs, LSTMs, or transformers for understanding temporal sequences to generate summaries. The paper categorizes methods into supervised, weakly-supervised, and unsupervised approaches, and reviews datasets like TVSum and SumMe. It also high- lights challenges such as the need for large labeled datasets, evaluation differences, and difficulty in deployment on low- power devices.

The paper “Attention Over Attention: An Enhanced Supervised Video Summarization Approach [2]” by Puthige et al. proposes a supervised model combining multi-head, spatial, and channel attention to improve frame selection and sequence smoothness. While it shows better performance on benchmark datasets, it requires large labeled data and has high computational complexity.

The paper "Video Transcript Summarizer [3]" by Ilampiray et al. focuses on transcript-based summarization using ASR and NLP techniques like BERT and LDA. It supports multi- lingual processing but faces issues such as ASR errors, unclear speech handling, and lack of visual feature integration.

The work "Abstractive Summarizer for YouTube Videos [4]" by Devi et al. uses transformer models and techniques like LSA/TF-IDF to generate summaries from video transcripts. It also supports multilingual translation, but challenges include large data requirements, sentence flow improvement, and scalability.

The "AI-Powered Real-Time Video Summarization [5]" by Venu Gopal S R et al. presents a real-time system using OpenCV, DNNs, ASR, and NLP integrated with a Streamlit interface. It focuses on speed and usability but faces challenges like limited labeled data, hardware dependency, and balancing speed with accuracy.

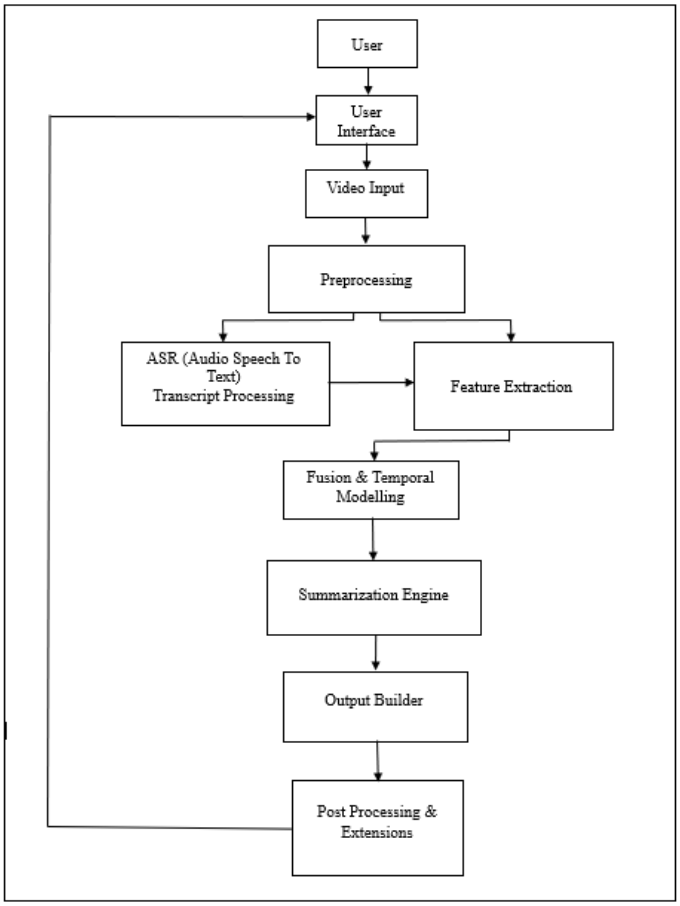
The paper "Query-Driven Video Summarization for Long Video Footage Analysis Using Faster-RCNN and DPP[6]" by Bhute et al. introduces a query-based approach for long videos using object detection and frame selection techniques. It provides personalized summaries but may suffer from low recall, slower performance for long videos, and lack of AR/VR integration.

Recent work in video summarization ranges from foundational surveys of deep learning techniques (Apostolidis et al., 2021) to advanced attention-based models for improved frame selection (Puthige et al., 2023) and transcript-based or abstractive summarization systems with multilingual capabilities (Ilampiray et al., 2023; Devi et al., 2023). Real-time summarization systems for educational applications (Venu Gopal S R et al., 2024) and query-driven approaches for personalized long-video analysis (Bhute et al., 2025) demonstrate the growing practical relevance of this field. Despite these advancements, challenges such as large data requirements, high computational costs, scalability issues, and integration with emerging technologies still remain areas for future research and development.

III. SYSTEM DESIGN

The architecture of the Automated Video Summarization System is illustrated in the given figure. It consists of stages such as video input, preprocessing, speech-to-text conversion, feature extraction, and summarization to generate meaningful outputs from video content.

System Architecture



1. Input and Preprocessing

The architecture of the Automated Video Summarization System begins with the user interacting through the user interface. The user can either upload a video file from the local system or provide an online video link as input. Once the video is received, it enters the preprocessing stage, where the system prepares the video for further analysis. In this stage, important operations such as audio extraction, audio cleaning, and formatting are performed. These preprocessing steps help improve the quality of the data and ensure accurate processing in the later stages of the system.

2. Speech Recognition and Summarization

After preprocessing, the extracted audio is passed to the Automatic Speech Recognition (ASR) module, where the spoken content of the video is converted into text. Along with speech-to-text conversion, the system also performs feature extraction to identify important patterns, keywords, and contextual information from the video. The extracted transcript and features are then combined in the fusion and temporal modeling stage to understand the sequence and meaning of the content. Finally, the summarization engine processes this information and generates a concise, meaningful, and context-aware summary of the video.

3. Output Generation and User Interaction

Once the summary is generated, the output builder organizes and displays the results in a structured format for the user. The system not only provides the summary text but also generates key points for easier understanding. Additional features such as multilingual translation and text-to-speech functionality improve accessibility by allowing users to read or listen to the summary in different languages. Furthermore, the system includes interactive quiz generation to test user understanding and enhance engagement. These features together make the system more intelligent, user-friendly, and effective for learning and information extraction.

IV. METHODOLOGY

The methodology of the proposed system outlines the sequential steps involved in analyzing video content and generating concise summaries using artificial intelligence and deep learning techniques.

1) User Input: The system begins by accepting input from the user in the form of either a video file upload or a video link. If a link is provided, the video is downloaded using appropriate tools, while uploaded files are stored locally for further processing.

2) Preprocessing: In this stage, the input video is prepared for analysis. The system extracts the audio from the video and converts it into a standard format suitable for speech recognition, ensuring better quality and consistency in further processing.

3) Audio Extraction: The system isolates the audio com- ponent from the video using multimedia processing tools. This step helps reduce computational complexity by focusing only on the relevant audio data required for transcription.

4) Automatic Speech Recognition (ASR): The extracted audio is converted into text using the Whisper model. This process generates a transcript of the spoken content in the video, which serves as the foundation for the summarization task.

5) Text Processing: The generated transcript is cleaned and organized to remove unnecessary elements and improve readability. It is then divided into smaller segments to enable efficient processing of long video content.

6) Text Summarization: The system applies transformer- based models such as BART to generate summaries from the processed text. Each segment is summarized individually and then combined to produce a meaningful overall summary.

7) Summary Refinement: A final refinement step is per- formed to improve the coherence, clarity, and flow of the summary. The system also adjusts the length of the summary based on the user's input requirements.

8) Output Generation: The final summary is presented to the user through a web-based interface. The output is designed to be concise, clear, and easy to understand, enabling quick access to key information from the video.

Implementation Details:

The Automated Video Summarization System is implemented using a combination of web technologies and ma- chine learning models to ensure efficient processing and user interaction. The system follows a modular approach where different components work together to convert video input into meaningful summaries.

A. Frontend and Backend Development

The system is developed using Python with Flask as the backend framework, which handles user requests, file uploads, and processing flow. The frontend is built using HTML, CSS, and JavaScript, providing a simple and interactive interface where users can upload videos or provide links, set summary length, and view the output.

B. Video Processing and Speech Recognition

For handling video inputs, tools like yt-dlp are used to download videos from online platforms, while FFmpeg is used to extract audio and convert it into a standard format. The extracted audio is then processed using the Whisper model, which performs Automatic Speech Recognition (ASR) to convert speech into text.

C. Text Summarization and Output Generation

The generated transcript is processed using transformer- based models such as BART through the Hugging Face Trans- formers library. A chunk-based approach is used to handle long text, where smaller segments are summarized and combined into a final refined summary. The output is then displayed to the user in a clear and readable format, with the option to control the summary length.

V. RESULT ANALYSIS

A. Accuracy and Quality Analysis

Parameter	Observation
Content Coverage	The summary captures the main ideas and key points effectively.
Context Preservation	Maintains the overall meaning and flow of the original content.
Relevance	Most generated sentences are relevant to the video topic.
Clarity	Output is clear, concise, and easy to understand.
Error Sources	Minor inaccuracies may occur due to noise or unclear speech.
Overall Accuracy	Provides satisfactory results for general-purpose summarization.

B. Performance Analysis

The system demonstrates stable and consistent performance across different types of inputs, including both locally up- loaded videos and online video links. It efficiently processes videos through a well-structured pipeline that includes audio extraction, speech-to-text conversion, and summary generation. Each stage is optimized to ensure that the overall processing time remains reasonable and suitable for practical use.

The system is also capable of handling longer videos effectively by using a chunk-based processing approach, where large transcripts are divided into smaller segments. This not only prevents memory issues but also improves processing efficiency. Additionally, the integration of optimized tools and models ensures smooth execution without major delays, making the system reliable for real-world applications.

C. Limitations Observed

- Dependency on Audio Quality: The system's performance depends on clear audio input; noise and unclear speech can reduce accuracy.
- Processing Time for Long Videos: Longer videos re- quire more time due to multiple processing stages like transcription and summarization.
- Limited Visual Understanding: The system mainly focuses on audio content and does not deeply analyse visual elements of the video.

D. User Experience

The system provides a simple and user-friendly interface that allows users to easily upload video files or provide video links without any technical difficulty. The design is intuitive, making it easy for users to navigate and understand the workflow, even for first-time users. The overall interaction is smooth and efficient, with a clear process from input to output. Users can also control the summary length based on their needs, and the generated summary is displayed in a clean and readable format, making it easy to understand key information.

Due to its ease of use and accessibility, the system is suitable for a wide range of users, including students, professionals, and researchers, ensuring a convenient and efficient experience in video content consumption.

VI. CONCLUSION

The Automated Video Summarization System effectively utilizes Artificial Intelligence (AI), Automatic Speech Recognition (ASR), and Natural Language Processing (NLP) to generate concise and meaningful summaries from video content. The system is capable of accepting both video files and online video links, extracting audio, converting speech into text, and producing accurate summaries. It also provides flexibility by allowing users to control the length of the summary based on their requirements.

The developed system significantly reduces the time and effort required to understand lengthy videos, making content consumption more efficient and accessible. By integrating advanced models and a user-friendly interface, the system ensures smooth interaction and reliable performance across different types of inputs.

Overall, this project demonstrates the practical application of AI-based techniques in simplifying video understanding. It can be effectively used in domains such as education, corporate training, and research, offering a scalable and efficient solution for managing and summarizing large volumes of video data.

VII. FUTURE SCOPE

In the future, the Automated Video Summarization System can be further enhanced by focusing on scalability, performance optimization, and intelligent personalization. The system can be improved to handle large-scale video data efficiently using cloud-based deployment and distributed processing. Advanced personalization techniques can be introduced to generate user-specific summaries based on preferences, viewing history, and domain requirements. Additionally, integrating more sophisticated multimodal analysis by combining audio, text, and visual features can further improve summary accuracy and contextual understanding. The system can also be extended with real-time processing capabilities for live streaming platforms, along with stronger security, data privacy measures, and seamless integration with enterprise applications such as e-learning systems, corporate training platforms, and content management systems.

REFERENCES

[1] Apostolidis, E. Adamantidou, A. Metsai, V. Mezaris, and I. Patras, "Video Summarization Using Deep Neural Networks: A Survey," Proceedings of the IEEE, vol. 109, no. 11, pp. 1838–1863, 2021.

[2] S. H. Puthige and G. Prabhu, "Attention Over Attention: An Enhanced Supervised Video Summarization Approach," in 2023 International Conference on Computer Communication and Informatics (ICCCI), IEEE, pp. 1–6, 2023.

[3] Ilampiray and D. Aishwarya, "Video Transcript Summarizer," International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), vol. 12, no. 6, pp. 13145–13150, 2023.

[4] Devi, M. Nandini, and P. Kumar, "Abstractive Summarizer for YouTube Videos," International Research Journal of Modernization in Engineering Technology and Science (IRJMETS), vol. 5, no. 4, pp. 1721–1726, 2023.

[5] S. R. Venu Gopal and S. Raghavendra, "AI-Powered Real-Time Video Summarization," International Journal of Creative Research Thoughts (IJCRT), vol. 12, no. 5, pp. 455–463, 2024.

[6] P. Bhute and S. Patel, "Query-Driven Video Summarization for Long Video Footage Analysis Using Faster-RCNN and Determinantal Point Processes," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16, no. 2, pp. 220–229, 2025.