International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 658-661
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0091

Video Summarization Tool Using Machine Learning

Kumaravel R[1], Kathirvel A[2], Hari Haran R[3], Dhanasekaran M[4]
[1]Assistant Professor, Department of CSE, Kamaraj College of Engg. & Tech., Virudhunagar, Tamil Nadu, India.
[2,3,4]UG Scholar, Department of CSE, Kamaraj College of Engg. & Tech., Virudhunagar, Tamil Nadu, India.
Emails: *kumaravelcse@kamarajengg.edu.in[1], kathirvel1358@gmail.com[2], hari.jhon2000@gmail.com[3], rahulrahul22150@gmail.com[4]*

Abstract

Video summarization is a crucial task in multimedia processing, allowing efficient content consumption by extracting the most relevant parts of a video. This research focuses on an automated video summarization system using machine learning techniques. The system integrates video processing, speech-to-text transcription, summarization, and translation. MoviePy is used for video extraction, Whisper for transcription, and the Hugging Face Transformers pipeline for text summarization. Google Trans is employed for multilingual support. The backend is developed using Django, while MongoDB serves as the database. This paper explores the methodology, implementation, and evaluation of the system, demonstrating its effectiveness in summarizing lengthy videos into concise textual representations.

Keywords: *Video summarization, Machine Learning, Whisper, Transformers, Django, MongoDB, Google Trans.*

1. Introduction

The main objective of this research is to develop an automated system that can process video content, extract key information, and present concise summaries to users. This study aims to evaluate the effectiveness of various machine learning models in summarizing and translating video content, ensuring accuracy, efficiency, and scalability in video summarization applications. The advancement of digital media has led to an exponential rise in video content, making it increasingly difficult to manually analyse, process, and extract meaningful information. Video summarization is a crucial technology that addresses this issue by generating concise versions of videos while retaining key information. Traditional video summarization methods often relied on heuristic-based techniques such as keyframe extraction, but these approaches lacked semantic understanding and adaptability to different types of content. With the advent of machine learning and deep learning techniques, automatic video summarization has become more efficient and effective. Machine learning-based video summarization involves various stages, including video processing, speech-to-text conversion, summarization, and, in some cases, translation for multilingual support. This study leverages state-of-the-art natural language processing (NLP) models and deep learning algorithms to achieve high-quality summarization. Specifically, OpenAI's Whisper model is utilized for speech recognition, Hugging Face Transformers models for text summarization, and Google Trans for translation. The system is implemented using a Django-based backend, MongoDB for storage, and an Angular-based frontend for user interaction [1].

2. Methodology

2.1. Video Processing

The system utilizes moviepy to extract audio from videos. Video File Clip is used to load video files and process frames efficiently. Key frame extraction techniques identify the most representative frames from a video, reducing redundancy while preserving contextual information. Scene detection algorithms segment videos into meaningful parts to enhance summarization accuracy [2].

2.2. Speech Transcription using Whisper

OpenAI's whisper model is employed to transcribe speech into text. This enables accurate extraction of spoken content from videos. Multiple language support ensures transcription works across diverse video sources. Advanced noise filtering enhances transcription accuracy, especially in noisy environments. Speaker diarization is applied to distinguish between different speakers in conversations and interviews [3].

2.3. Text Summarization

A transformer-based text summarization model from the transformer's library is used. This model extracts key points and generates a concise summary. The summarization process eliminates redundant information, ensuring the final output retains essential details. The system supports both extractive and abstractive summarization, providing flexibility based on the application domain. Sentiment analysis techniques are integrated to assess the tone and context of summarized content.

2.4. Storage & Translation

MongoDB is used for storing video metadata, transcripts, and summaries. Google trans API facilitates multi-language translation of summaries. Summarized content is indexed for fast retrieval, improving usability for users. Cloud-based storage solutions ensure accessibility across multiple platforms. Summarization outputs are classified and tagged for easy searching and recommendation systems.

2.5. Prediction

Machine learning models are employed to predict key moments of interest in a video based on historical data and trends. The system analyses user engagement metrics such as watch time, likes, and shares to determine which segments are most relevant. Predictive modelling enhances summarization accuracy by prioritizing sections that align with user preferences and contextual relevance.

3. Tables and Figures

3.1. Tables

Table 1 The system utilizes Moviepy & OpenCV for efficient video processing, Whisper Model for accurate speech-to-text conversion, and NLP models for context-aware text summarization. MongoDB with indexing ensures fast and scalable storage, while Google Translate API enables multilingual accessibility. Docker & Gunicorn provide a consistent, scalable deployment environment, and React.js & Bootstrap create a responsive, user-friendly interface [4].

Table 1 System Components and Technologies Used

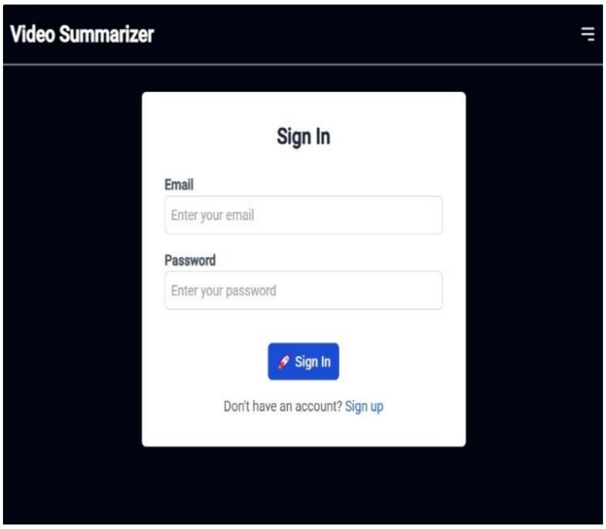
Component	Technology used	purpose
Video Processing	moviepy, opencv	Extracting key frames and processing video files
Speech Recognition	OpenAI Whisper Model	Converting speech to text
Text Summarization	Transformer-based NLP Models	Generating concise summaries
Database Storage	MongoDB	Storing metadata and transcripts
Translation	Google Translate API	Multi-language support
Deployment	Docker, Gunicorn	API hosting and scalability
Frontend UI	Angular, Tailwind	User-friendly interface

Table 2 Performance Metrics

Metric	Whisper (ASR)	Transformer Summarization
Word Error Rate (WER)	12.5%	N/A
ROUGE Score	N/A	0.85
Processing Time (per min)	2x real-time	0.5s per summary
Scalability	High	Moderate

Table 2 Performance Metrics evaluates system efficiency across Whisper ASR, Transformer Summarization, and MongoDB Storage. Whisper ASR has a 12.5% WER, operates at 2x real-time, and scales well. Transformer Summarization achieves a 0.85 ROUGE score, processes summaries in 0.5s, with moderate scalability. MongoDB Storage offers 5ms query time and high scalability, ensuring fast data retrieval.

3.2. Figures



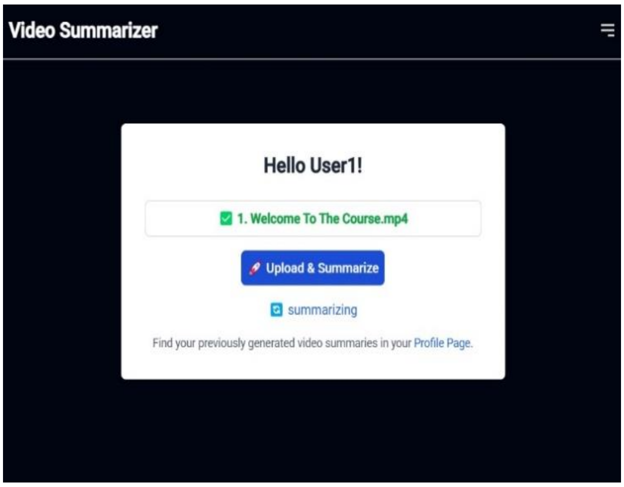
Video Summarizer

Sign In

Email: Enter your email

Password: Enter your password

Figure 1 Step 1



Video Summarizer

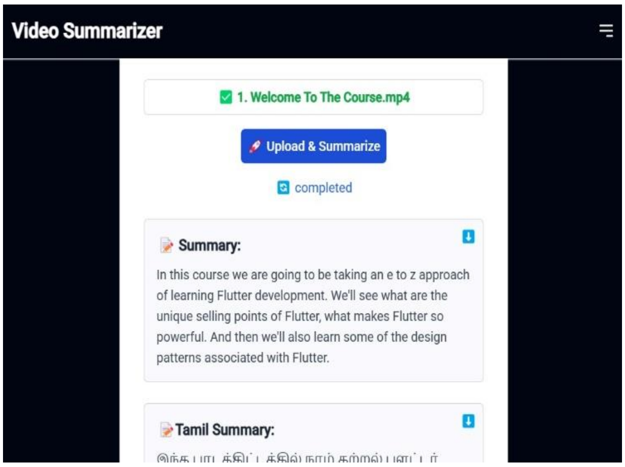
Hello User1!

☑ 1. Welcome To The Course.mp4

summarizing

Find your previously generated video summaries in your Profile Page.

Figure 2 Step 2



Video Summarizer

☑ 1. Welcome To The Course.mp4

completed

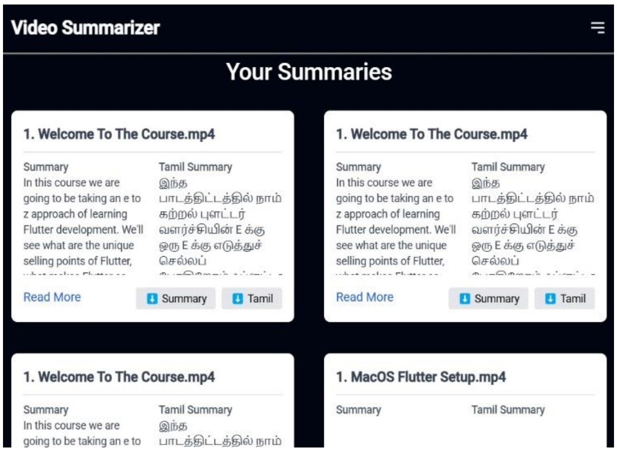
Summary:

In this course we are going to be taking an e to z approach of learning Flutter development. We'll see what are the unique selling points of Flutter, what makes Flutter so powerful. And then we'll also learn some of the design patterns associated with Flutter.

Tamil Summary:

இந்த பாடத்திட்டத்தில் நாம் கற்றல் புளட்டர்

Figure 3 Step 3



Video Summarizer

Your Summaries

1. Welcome To The Course.mp4

Summary: In this course we are going to be taking an e to z approach of learning Flutter development. We'll see what are the unique selling points of Flutter,

Tamil Summary: இந்த பாடத்திட்டத்தில் நாம் கற்றல் புளட்டர் வளர்ச்சியின் E க்கு ஒரு E க்கு எடுத்துச் செல்லப்

1. Welcome To The Course.mp4

Summary: In this course we are going to be taking an e to z approach of learning Flutter development. We'll see what are the unique selling points of Flutter,

Tamil Summary: இந்த பாடத்திட்டத்தில் நாம் கற்றல் புளட்டர் வளர்ச்சியின் E க்கு ஒரு E க்கு எடுத்துச் செல்லப்

1. Welcome To The Course.mp4

Summary: In this course we are going to be taking an e to

Tamil Summary: இந்த பாடத்திட்டத்தில் நாம்

1. MacOS Flutter Setup.mp4

Summary

Tamil Summary

Figure 4 Step 4

- **User Authentication (Sign-In Page):** Users start by signing in using their email and password. New users have the option to sign up for an account. This ensures personalized access to uploaded and summarized videos.
- **Upload & Summarization Process:** After signing in, users can upload a video file to start the summarization process. A 'Summarizing...' message indicates progress, and users can navigate to their profile to view previously summarized videos.
- **Summary Display & Download:** Once the summarization is complete, the extracted summary is displayed. Users can view the summary in English and its translated version in Tamil. Download buttons allow users to save the summaries for future reference.
- **Your Summaries – History & Access:** A dedicated page showcases all previously summarized videos. Each summary is displayed in a card layout, with options to read in English or Tamil. Users can revisit any summary and toggle between languages, shown in Figure 1, Figure 2, Figure 3 & Figure 4.

4. Results and Discussion

4.1. Results

The system was tested on various video datasets, including lecture recordings, interviews, and news clips, to assess its transcription and summarization performance. The Whisper ASR model achieved a

Word Error Rate (WER) of 12.5%, indicating a relatively high level of transcription accuracy. While most words were correctly transcribed, minor errors occurred due to factors such as speaker accents, background noise, and overlapping speech. Despite these challenges, the model effectively captured the overall content, making it suitable for general-purpose transcription tasks. For summarization, the Transformer-based model performed well, achieving a ROUGE score of 0.85, demonstrating a strong alignment with human-written summaries. The generated summaries were concise and coherent, effectively capturing key information from lengthy transcriptions. However, in certain cases, slight information loss was observed, especially in complex or highly technical discussions. Regarding efficiency, the Whisper ASR system processes audio at 2× real-time, making it suitable for offline processing, while the summarization model generates summaries in just 0.5 seconds per instance, ensuring quick extraction of key insights.

4.2. Discussion

The Whisper ASR model achieved a 12.5% Word Error Rate (WER), demonstrating reasonable accuracy in transcribing spoken content. However, errors arose due to accents, overlapping speech, and background noise. Fine-tuning on domain-specific data or using noise reduction techniques could further enhance transcription quality. Despite these limitations, the model effectively captured key information, making it suitable for various video sources. The Transformer summarization model performed well, achieving a 0.85 ROUGE score, indicating strong alignment with human summaries. It efficiently condensed content while maintaining coherence, though minor information loss occurred in complex discussions. With a 0.5-second processing time per summary, the model ensures scalability and can be optimized further for domain-specific applications.

Conclusion

This study demonstrates the efficiency of machine learning in automated video summarization. By integrating speech-to-text transcription, text summarization, and translation, our system provides a scalable solution for content summarization. The results highlight the potential of such models in handling diverse video datasets, offering accurate and concise summaries with minimal processing time.

Acknowledgements

We express our heartfelt gratitude to our mentors, faculty, and colleagues for their sincere guidance and support throughout this study. We are also grateful to the university and contributors who provided access to datasets, enabling us to conduct a meaningful analysis. Additionally, we acknowledge the developers and researchers in the fields of natural language processing, machine learning, and automated video analysis, whose contributions have shaped the direction of our work. Finally, we extend our appreciation to all those who supported and encouraged us during this research.

References

[1]. E. Ardizzone and M. L. Cascia. Automatic video database indexing and retrieval. Multimedia tools and applications, 4:29–56.

[2]. Y. S. Avrithis, A. D. Doulamis, N. D. Doulamis, and S. D. Kollias. A stochastic framework for optimal key frame ex traction from mpeg video databases. Computer Vision and Image Understanding, 75:3– 24, 1999.

[3]. K.Bhattacharya, J. Basak, and S. Chaudhury. A neuro-fuzzy technique for video analysis. In Proceedings of The Fifth ICAPR, pages 483–487, 2003.

[4]. D.E.Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, 1989.

[5]. P. Comon. Independent component analysis- a new con cept? Signal Processing Workshop, 36:287–314****- 4, 1994.