Video Summary Generation System

Aishwarya Mansing Phadtare, CSE Department, D. Y. Patil College of Engineering and Technology Kolhapur,
aphadtare83@gmail.com
Mitali Ganesh Raut, CSE Department, D. Y. Patil College of Engineering and Technology Kolhapur,
mitali.raut12@gmail.com

Ashwini Lalit Patil, CSE Department, D. Y. Patil College of Engineering and Technology Kolhapur,
ashpatil2310@gmail.com

Arya Ajit Kulkarni, CSE Department, D. Y. Patil College of Engineering and Technology Kolhapur,
aryakulkarni0502@gmail.com

Spruha Tushar Gadre, CSE Department, D.Y. Patil College of Engineering & Technology Kolhapur
spruhagadre@gmail.com

Pranoti Ashok Kumbhar, CSE Department, D. Y. Patil College of Engineering and Technology Kolhapur,
pranoti.kumbhar55@gmail.com

Abstract: -

This paper presents a video summary generation system that uses AI to create condensed, informative versions of video content. By utilizing machine learning and natural language processing (NLP), the platform identifies and extracts key sections of a video, eliminating redundant information. It generates structured summaries in multiple languages, available for download in text or PDF formats. The system also features AI-generated quizzes to assess comprehension and a chatbot for real-time assistance. Designed for education, corporate training, and research, the platform enhances video consumption by making it more efficient, interactive, and accessible.

I. INTRODUCTION

The Video Summary Generation System is an advanced AI-powered platform designed to optimize how users engage with video content by integrating features such as automated summarization, language translation, summary customization, downloadable summaries, AI-generated quizzes, and an interactive chatbot. Utilizing machine learning algorithms, the system efficiently processes video content by extracting key insights and generating concise summaries, making it easier for users to grasp essential information without watching the full video. This is especially beneficial for individuals with time constraints who still need to understand the main concepts quickly. To further enhance accessibility, the language translation feature enables summaries to be converted into multiple languages, catering to a global audience and eliminating language barriers. The platform also provides a summary download option, allowing users to save summaries in various formats, such as text files and PDFs, making it convenient for offline access, sharing, and incorporation into presentations or reports. In addition to summarization, the platform promotes active learning through its AI-generated quiz feature, which creates interactive assessments based on video content, enabling users to test their knowledge, reinforce learning, and improve retention. This feature is particularly valuable in educational settings, corporate training programs, and professional development courses. To further enhance user engagement, an AI-powered chatbot is integrated, providing real-time assistance, answering questions related to video content, offering additional insights, and guiding users through the platform. The chatbot is designed to understand context and deliver precise responses, ensuring a seamless and intuitive user experience. By leveraging automation, personalization, and AI-driven insights, the platform revolutionizes video consumption by making content more interactive, accessible, and engaging for users in diverse fields such as education, business, research, and training. Its ability to streamline information, facilitate learning, and enhance engagement ensures that users can efficiently extract valuable knowledge from video content, making it a powerful tool for modern digital learning and communication.

II. LITERATURE REVIEW

Video summary generation is the rapid growth of video content on the internet has led to increasing demand for systems that can efficiently condense videos into meaningful summaries. Song et al. [1] introduced TVSum, a technique leveraging web video titles to guide summarization, highlighting the role of textual metadata in identifying key video segments. Their approach integrates user-generated metadata to generate summaries more aligned with human interest.Language translation is essential for making summaries globally accessible. The RNN-based encoder-decoder framework by Cho et al. [6] introduced a powerful architecture for machine translation that learns to map sequences of one language to another. This method underpins many modern multilingual systems and is foundational for integrating translation features into video summary platforms. Their model, capable of handling variable-length sequences and maintaining contextual relevance, supports summary conversion into multiple languages without losing meaning. Integration into Educational and Interactive Systems are not covered directly in the provided references, the advancements in summarization and translation from these works underpin features like AI-generated quizzes and chatbots. The ability to extract semantically rich and contextually aware summaries, as demonstrated in [2] and [4], forms the basis for generating meaningful assessments and interactive learning tools. These features contribute significantly to learner engagement and knowledge retention in educational and professional contexts.

III. PROPOSED SYSTEM

3,1SYSYTEM ARCHUTECTURE

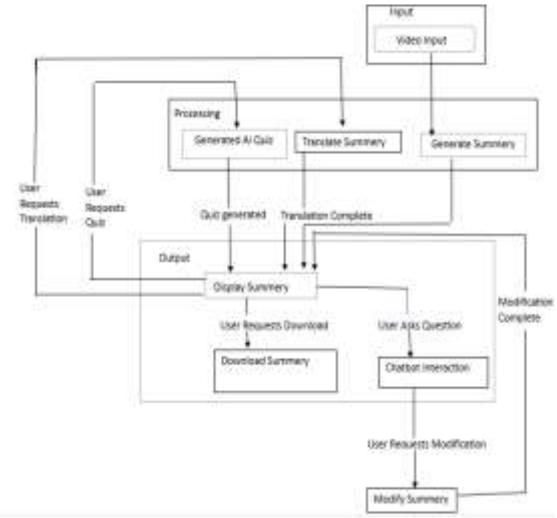

Fig.1 System Architecture

3.1. Video Input: This module is responsible for taking in the video that will be summarized. It acts as the starting point for the entire process.

3.2. Summary Generation and translation: This module processes the input video to create a summary. It extracts the main content from the video and condenses it into a more manageable format. If the user requests, this module translates the generated summary into the desired language.

3.3. Summary Modification and Chatbot: This allows the user to download the summary to their local device. A chatbot is available for user interaction, where the user can ask questions related to the summary.

3.4. Modify Summary: If the user requests any changes or modifications to the summary, this module handles the modification process. It ensures the summary is updated according to the user's preferences or requirements.

Introduction to Video Analysis Using Machine Learning

Video analysis refers to the automated process of extracting valuable insights, patterns, and information from video content. As video data continues to expand across sectors like healthcare, education, media, and surveillance, this field has become increasingly important in both research and practical applications.

Feature Extraction involves identifying and isolating critical visual elements such as objects, motion patterns, and spatial structures. Common techniques include Optical Flow, Histogram of Oriented Gradients (HOG), and Convolutional Neural Networks (CNNs).

Object Detection and Tracking focuses on locating and following objects across video frames. Well-known models for this task include YOLO (You Only Look Once), SSD (Single Shot Detector), and Faster R-CNN.

Action Recognition aims to identify human activities or behaviours captured in video sequences. Techniques such as Long Short-Term Memory 4(LSTM) networks, 3D CNNs, and Transformer models are widely applied in this area.

Video Summarization is used to create condensed versions of video content by selecting keyframes or significant segments. Popular methods include clustering techniques, reinforcement learning, and graph-based models.

Scene Understanding focuses on interpreting environments, contexts, and events present in video data. Techniques like semantic segmentation and attention mechanisms are frequently employed to achieve this.

Summary Generation and Language Translation

Summary generation and language translation are transformative technologies that enhance video accessibility by extracting essential content, condensing it into concise summaries, and ensuring multilingual support. Automatic video summarization leverages advanced deep learning techniques, such as Recurrent Neural Networks (RNNs) and Transformer models, to analyse video transcripts, detect key points, and generate structured summaries while maintaining the context and meaning of the original content. This process is particularly valuable for users who need quick insights without watching lengthy videos, making content more efficient to consume. In parallel, language translation expands accessibility by converting these summaries into multiple languages using Neural Machine Translation (NMT) models, ensuring that information is available to a global audience. By breaking language barriers, this technology allows users from different linguistic backgrounds to engage with video content seamlessly. Additionally, AI-driven translation models continuously improve by learning from large datasets, enhancing accuracy and contextual relevance. The integration of summarization and translation benefits various fields, including education, corporate training, media, research, and global communication, by making information more digestible, inclusive, and widely available. This technology-driven approach saves time, enhances knowledge retention, and improves accessibility, allowing users to interact with video content more efficiently. As AI advancements continue, these features will become even more sophisticated, offering real-time summarization and instant multilingual translations, ultimately revolutionizing the way video-based knowledge is consumed and shared worldwide.

Downloadable Summaries

Offering downloadable summaries enhances user convenience by enabling offline access, seamless sharing, and well-structured record-keeping. This feature involves converting text summaries into multiple formats such as **PDF, DOCX (Word), and TXT**, each catering to different user needs—PDF ensures fixed formatting and security, DOCX allows for edits, and TXT provides a lightweight, universally accessible option. The implementation process typically starts with backend processing, where the system generates or retrieves the summary, applies necessary formatting, and converts it into the chosen format using tools like reportlab or pdfkit for PDFs, python-docx for Word documents, and standard file writing for text files. These files can be stored temporarily for instant downloads or saved on cloud platforms like AWS S3 or Firebase for long-term access. The frontend should offer an easy-to-use interface with a format selection option and a **"Download"** button that requests the appropriate file from the backend, which then delivers it through an HTTP response or a cloud storage link. Security aspects such as **authentication, access control, rate limiting, and encryption** must be incorporated to prevent unauthorized access and misuse. Additionally, branding elements like watermarks and headers can be added for a professional touch. Implementing an efficient download feature improves user experience, increases engagement, and provides an accessible and organized way to manage digital content

AI-Powered Quizzes

AI-driven quizzes create an engaging learning experience by generating questions based on video content, helping users reinforce their understanding interactively. Using Natural Language Processing (NLP), these quizzes analyse video transcripts, extract key details, and form different question types such as multiple-choice, fill-in-the-blank, true/false, and short-answer questions. Advanced machine learning models, including Question-Answering (QA) systems, contribute by dynamically generating relevant questions and assessing responses. Techniques like keyword extraction, summarization, and Named Entity Recognition (NER) help pinpoint important concepts, while AI models such as GPT, BERT, or T5 assist in structuring well-formed questions. Additionally, AI can adjust quiz difficulty based on user performance, providing customized learning experiences and detailed explanations to improve retention. Integrating speech-to-text technology allows real-time analysis of spoken content, ensuring that quizzes accurately reflect the video's main points. Automating quiz creation through AI streamlines the process, maintains assessment quality, and offers data-driven insights into learner progress, making education more efficient and interactive.

How AI Generates Quizzes

1. **Transcription & Text Processing:** AI-powered **speech-to-text models** first convert spoken content into text. NLP techniques then clean and structure the text for analysis.

2. **Keyword Extraction & Summarization:** AI scans the transcript to find essential concepts, key terms, and topic summaries using methods like **Named Entity Recognition (NER)** and **text summarization algorithms**.

3. **Question Formation:** Machine learning models such as **GPT, BERT, or T5** generate well-structured questions based on extracted key points. These models use **contextual understanding** to form multiple-choice questions (MCQs), open-ended questions, or even scenario-based problem-solving tasks.

4. **Adaptive Learning Features:** AI can modify quizzes dynamically based on user performance, offering **difficulty adjustments** and **personalized feedback**. If a user struggles with a concept, AI can present additional hints, explanations, or alternative questions to reinforce learning.

5. **Evaluation & Scoring:** AI can **automatically assess** responses, provide instant feedback, and analyse patterns in user answers to identify areas where learners may need improvement.

Chatbot Interactions

AI-driven chatbots revolutionize video-based learning by offering instant support, real-time query resolution, and personalized guidance, making education more interactive and efficient. Utilizing Natural Language Processing (NLP) and machine learning models like GPT-4, these chatbots process video content by converting speech to text, identifying key points, and interpreting context to generate relevant and precise responses. Learners can engage with the chatbot by asking questions, receiving clear explanations, and accessing concise summaries, ensuring a smooth and dynamic learning experience. These chatbots also feature adaptive learning, tailoring their responses to each user's proficiency level—providing basic explanations for beginners and in-depth insights for advanced learners. To enhance comprehension, they can suggest related content, quizzes, and extra study materials, allowing users to strengthen their understanding. Beyond answering queries, chatbots encourage engagement with interactive conversations, gamification features like quizzes and rewards, and multilingual support, making education accessible across different regions. Their seamless integration with e-learning platforms, corporate training programs, and video services ensures context-aware and proactive assistance throughout the learning process. Over time, AI chatbots refine their capabilities by analyzing user interactions, enhancing response accuracy, and recognizing learning patterns, ensuring a continuously improving educational experience. With their ability to deliver intelligent, interactive, and on-demand support, AI chatbots make video-based education more engaging, accessible, and tailored to individual learners worldwide.

IV. GRAPHICAL REPRESENTATION

Stage	CPU Time(s)	GPU Time(s)
Load BART & Whisper	17	15
Audio Extraction	13	13
Transcription (Whisper)	130	30
Summarization (BART)	50	15
Cleanup	10	10

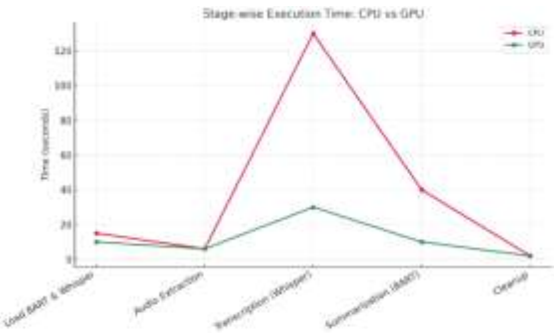

Fig 2 Stage-wise CPU vs GPU Execution Time in Video Summarization

The graph compares CPU and GPU execution times across key stages of a video summarization pipeline: model loading, audio extraction, transcription (Whisper), summarization (BART), and cleanup. Non-intensive stages show minimal difference between devices. GPU significantly outperforms CPU in transcription and summarization. Overall, GPU reduces total processing time by approximately 70%.

V. CONCLUSION:

Video summarization plays a crucial role in applications like video retrieval, personalized recommendations, and surveillance. Methods can be categorized by input type and summary format. This paper reviews current techniques, methodologies, and benchmarks. A key challenge is the lack of datasets supporting multi-modal inputs and outputs across domains, such as education, medical, and news. Interactive summarization, where users control summary length, needs further development. Real-time methods are crucial for applications like autonomous driving and security. Scalable systems to process large volumes of data are essential. Improving context-awareness and semantic understanding will lead to more accurate summaries. Future research should focus on enhancing user interaction, creating diverse datasets, improving real-time capabilities, and advancing semantic understanding.

VI. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to everyone who contributed to the successful completion of our project, titled **"Video Summary Generator."**

Firstly, we are deeply thankful to our institution, D Y Patil college of engineering and technology, Kolhapur for providing us with significant resources and a proper guidance to work on this project.

We extend our sincere gratitude to our project guide, Miss **Aishwarya Phadtare**, for their invaluable support, guidance, constructive feedback, and encouragement throughout the development of the project.

We are also grateful to the faculty members of the **Computer Science and Engineering** for their valuable technical guidance and for encouraging us to explore new horizons in our research and implementation.

We also appreciate the contributions of our peers and colleagues, whose feedback, encouragement, and teamwork were instrumental throughout this project. Lastly, we would like to express our gratitude to our families for their unwavering patience and support during this journey.

This project has been a remarkable learning experience, and we are truly grateful for the chance to work on such an exciting and forward-thinking topic.

VII. REFERENCES:

[1] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, “TVSum: Summarizing web videos using titles,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5179–5187.

[2] S. M. Basha, A. N. Kannadasan, and D. M. Manogaran, “A deep learning-based hybrid model for video summarization and understanding,” Multimedia Tools and Applications, vol. 80, no. 16, pp. 24113–24130, 2021.

[3] R. Subramanian, D. Roth, and R. P. Dick, “Online video summarization via dynamic dictionary learning,” in Proceedings of the 24th ACM International Conference on Multimedia (MM), 2016, pp. 781–790.

[4] Y. Gan, X. Qian, and X. Tang, “A deep generative model for video summarization,” IEEE Transactions on Image Processing, vol. 28, no. 10, pp. 4977–4989, 2019.

[5] M. R. Amer, P. Lei, and S. Todorovic, “Hierarchical video summarization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2012, pp. 3246– 3253.

[6] K. Cho, B. Van Merriënboer, Ç. Gülçehre, and D. Bahdanau, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014,pp.1724–1734.

[7] K. Zhang, W. Lu, and J. Xiao used long short-term memory networks to create temporal summaries of videos. Their method was presented at ECCV 2016 and addresses the challenge of capturing long-term dependencies in video content.

[8] S. Narayan, H. Bui, and L. Cohn proposed a deep learning strategy for summarizing multiple documents, presented at CoNLL 2017. Their method focuses on coherent content selection across sources, which parallels strategies in video summarization.

[9] Y. Liu and M. Lapata explored the use of pretrained language models for extractive and abstractive summarization, as described in EMNLP 2019. Though text-focused, their techniques influence summarization in multimodal contexts.

[10] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid worked on category-specific summarization, enabling targeted summary generation. Their research was presented at ECCV 2014 and emphasizes tailoring outputs based on video content type.