Video Summarizer For Marathi

Chinmay Desle
Department of Artificial Intelligence &Data Science
Jawahar Education Society's, A.C. Patil College of Engineering, Kharghar
Navi Mumbai, India
chinmaydesle@acpce.ac.in

Kartik Kadam
Department of Artificial Intelligence &Data Science
Jawahar Education Society's, A.C. Patil College of Engineering, Kharghar
Navi Mumbai, India
kartikkadam@acpce.ac.in

Dr. Jayprabha Terdale
Department of Artificial Intelligence &Data Science
Jawahar Education Society's, A.C. Patil College of Engineering, Kharghar
Navi Mumbai, India
jvterdale@acpce.ac.in

Dnyaneshwari Devekar
Department of Artificial Intelligence &Data Science
Jawahar Education Society's, A.C. Patil College of Engineering, Kharghar
Navi Mumbai, India
dnyaneshwaridevekar@acpce.ac.in

Harshal Zanje
Department of Artificial Intelligence &Data Science
Jawahar Education Society's, A.C. Patil College of Engineering, Kharghar
Navi Mumbai, India
harshalzanje@acpce.ac.in

Abstract - Video summarization system is powered by a transformation-based algorithm, the Gemini API for summarization, and Streamlit for an interactive frontend interface. With an ever-increasing number of Marathi video contents, the system fills the gap of efficiently extracting relevant information from long videos without compromising on language and contextual integrity. Audio is extracted from video files, and speech-to-text conversion uses advanced Marathi-optimized speech-to-text technology. Whatever is transcribed forms part of text data that is fed through a transformation-based algorithm that preprocesses it, putting it into a desired format for Gemini API, which gives summary information in abbreviated, contextually appropriate paragraphs. The frontend, implemented with Streamlit, allows for video uploading, summary viewing, and interaction all in one seamless experience. These transformations on textual data yield enhanced quality and coherence of generated summaries.

Keywords— YouTube videos, video content, transcription, summary, translation., transformation-based algorithm and Gemini API.

I. INTRODUCTION

Videos have become the most popular way to communicate and share information in today's fast-paced digital world. From business presentations and security footage to instructional tutorials and entertainment, the amount of video information produced every day is astounding. Since watching these video through to the end might take a lot of time, effective video summary strategies are required to reduce waiting times and increase accessibility.

Several creative approaches have been suggested to solve this problem. For example, M3Sum [1] employs natural language processing methods without requiring pre-labeled data to produce summaries and performs impressively on datasets such as TVSum and SumMe. The other method relies on LLMs to interpret videos by translating voice into text that enables exact long-form video descriptions [2]. Additionally, techniques like the U-shaped Non-local Network [3] are made to identify important parts of videos while ignoring less important information.

In video summarization, techniques based on deep learning have also shown great potential by focusing keyframe selection, event recognition, and activity analysis, which provide clear and precise results [4]. In addition, tools with multi-language support, such as those summarizing in English, Hindi, and Marathi, have made video summaries more accessible to diverse audiences [5][6].

To build upon these advancements, a new system has been designed that uses the advanced Transformer algorithm that is implemented by combination with the Gemini API. Transformers are particularly well-suited for finding patterns and extracting key information and, therefore, ideal for video content summarization. The integration of Transformer with Gemini API makes various possibilities for summarizing videos in terms of accuracy and contextuality.

What sets this system apart is its ability to generate multilingual summaries in languages including English, Hindi, Marathi, French, and more, therefore targeting a worldwide audience. A user can upload a video or enter a URL, plus the preferred language to obtain a brief summary outlining the main ideas of the original material. This function makes the system more accessible and language-neutral, which benefits media outlets, educational institutions and large businesses.

This paper discusses methodology and implementation through this video summarization system based on the contribution of Transformer algorithms and the Gemini API that will boost both efficiency and precision. In addition to analyzing the role the of multilingual summarization, the paper shows how it can democratize information access and meet the demands of a globalized world.

The objective of this research is to develop an efficient video summarization system tailored to Marathi video content, to match the fast-evolving needs for quick access and summaries. Through transformation-based algorithms, advanced NLP techniques, and Gemini API, this system makes it possible for highly accurate contextually

appropriate, as well as multilingual, summaries. The overarching goal is to enhance information accessibility and usability of information for diverse subjects like education, media, and corporate communications targeted at a multilingual, multi-regional population of the country.

II. LITERATURE SURVEY

This research [1] proposed a novel approach to video summarization through the application of NLP techniques, pre-trained models for text generation including large language models, a parameter-free method of aligning modalities, and unsupervised multi-modal summarization based on textual descriptions. The method yielded state-of-the-art performance on TVSum and SumMe datasets as compared to other unsupervised and some supervised approaches. Future work includes refinement of alignment algorithms, improvements in integration across modalities, and work on extending the approach to a broader variety of video content models. The method does not require training data, is flexible, and works successfully across different modalities. However, it is possibly limited to videos that are deficient in text and audio content richness.

This paper [2] argues about scalable video summarization through large language models (LLMs). It describes a process for the transcription of long-form videos using speech-to-text models, summaries of transcriptions through LLMs, and the combination of summaries back to video segments in order to create a pseudo-ground truth for training. The approach showed substantial improvements in accuracy and generalization, establishing a new state-of-the-art performance. Future works include making the summarizing of videos without text or narration more robust, exploring cross-modal learning, and improving the pre-training datasets for greater accuracy and generalization. The methodology is scalable, integrates multimodal information to improve performance, but is heavily reliant on the quality of LLM-generated text and may create problems with bad speech-to-video correlation or no narration.

This research [3] utilizes a method from which long-range dependencies of frames in videos can be addressed with U-shaped non-local networks incorporating non-local operations and U-Net. The suggested method gives better performance in extracting relevant frames and thus provides more concise video summaries compared with existing models. Its future use by the authors includes experimenting with a diverse set of video datasets, optimizing computation load and efficiency, implementation for real-time summarization in video. Such a method allows the capture in the video sequence of long-range dependencies and provides a more concise video summary with significantly higher accuracy than the existing methods. However, it is computationally intensive, with the demand for a lot of resource for training and the probability of issues with real-time deployment due to complexity.

The authors [4] review deep learning methods for video summarization-oriented keyframe selection, event detection, and activity feature summarization. They categorize methods as supervised, unsupervised, weakly supervised, and reinforcement-learning-based approaches. Deep learning contributes significantly to increased accuracy and efficiency on video summarization, highlighting their advantages and disadvantages according to CNNs, RNNs, and GANs. Reduced redundancy within the video and increased processing speed are brought to their attention. Provide future directions for improving efficiency in summarization of long-duration videos, enhancing event detection accuracy under complex scenarios, and designing robust datasets and evaluation metrics. While deep learning offers efficient processing of extensive datasets of videos and minimizes the use of storage, it requires high computations and lacks across summarizing videos with low activity and complex scenes.

In this research paper [5] , by employing LSTM and DPP for diversity and NLP for text summarization, performances better than other techniques were recorded for benchmarks. The future work hopes to make much more advanced methods with the support of larger datasets. It supports multi-language summarized video (English, Hindi, Marathi) and has advantages for the generation of diverse and effective summaries. It calls for huge annotated datasets as part of training.

This work [6] tells us about the use of NLP and BERT in summary creation; plus, the use of both the YouTube Transcript API and the Google Translate API for more concise video summaries. Additionally, a browser extension for video platforms saves users time and supports many languages. While the method is used only for videos with transcripts, the time-saving possible advantage and support of multiple languages are its benefits.

This study [7] focuses different approaches that leverage extractive and abstractive summarization through NLP and machine learning techniques. They have reported a compression accuracy of 44.48. Future work aims to achieve higher compression accuracy and further domination on the Indian languages for getting better results of summary generation. Compression has been provided with relatively high accuracy and some specific needs of Marathi language have been accommodated. However, the system is limited to this one language and would have difficulty with complex sentence structures.

This paper [8] proposes an NLP-based approach and applies LSA techniques, MoviePy for video skimming, frame generation, and text recognition to create a summarization of videos along with translated text into Marathi for the benefit of non-English speakers. Future work will focus on assisting with summarization accuracy, expanding to include other regional languages, and stretching into the realm of real-time summarization. Thus, this approach helps with accessibility for non-English speakers and supports many regional languages, but is restricted to Marathi and probably won't translate easily into other languages or contexts.

The study [9] describes the use of a Python API to fetch YouTube transcripts, HuggingFace Transformers for text summarization, a Chrome extension to allow for user interaction. The method successfully summarized transcripts of YouTube videos and displayed on a Chrome extension. Future work involves working toward expanding the ability to summarize content from other platforms or using a bigger model for more summarization. It has helped the viewers by summarizing the entire video content and has provided a good user interface for it via the Chrome extension, but it's highly dependent on the quality of transcripts provided by YouTube and works only with videos whose transcripts are available.

This paper [10] makes use of NLP, LSA algorithms, Pytube library, MoviePy, and Python video editor to summarize YouTube videos with subtitles, thus generating quick video summaries with a lot less error. Future directions include algorithm enhancement, improving summarization accuracy and reducing processing time, and widening the scope of other video platforms. It's an effective trek for summarizing longer videos into concise summaries, hence saving time, but holds little accuracy due to dependence on subtitle quality and might miss important context.

III. PROPOSED METHODOLOGY

In order to enable users to summarize videos from YouTube URLs or locally uploaded files, this research focuses on creating a Marathi-specific video summarizer. The methodology comprises distinct stages, from input acquisition to summarization.

A. Data Input

The system accepts two forms of video input:

- **YouTube URL Input**: Users provide the URL of a YouTube video.
- **Local Upload**: Users upload a video file directly from their device.

In both cases, users specify the desired output language for the summary, with Marathi as the primary focus.

B. Caption Extraction

The approach for extracting captions depends on the input type:

- **YouTube URL Processing:**
  - The get_youtube_transcript library, which offers timestamped transcriptions of the spoken content of the video (if captions are available), is used to retrieve the captions directly).
  - Future iterations may incorporate fallback mechanisms, like manual transcription, in the event that captions are insufficient or unavailable
- **Local Video Processing:**
  - The video file that was uploaded has its audio extracted using FFmpeg..
  - Using an Automatic Speech Recognition (ASR) model, like OpenAI Whisper, the extracted audio is converted to text.

C. Summarization Using Gemini API

Once the transcription or captions are extracted, they are summarized using the **Gemini API**, which supports summarization in multiple languages, including Marathi. The summarization workflow is consistent for both input types:

- **YouTube URL Input**: The Gemini API receives the captions that get_youtube_transcript produced and summarizes them.
- **Local Upload**: The Gemini API processes the transcription produced from the uploaded video in a similar manner.

  Using sophisticated Natural Language Processing (NLP) techniques, the API produces succinct, readable, and contextually accurate summaries in the language the user has chosen.

D. Translation (If Necessary)

Prior to being summarized, the text is translated into Marathi if the transcription or captions' original language is not Marathi. External APIs or frameworks, such as Hugging Face Transformers for neural machine translation or Google Translate, can be used to incorporate this translation step.

E. Output Delivery

The summarized text is presented to the user in their chosen language. The system prioritizes:

- **Conciseness**: Making sure summaries are succinct but thorough.
- **LinguisticAccuarcy**: Preserving Marathi's grammatical accuracy .

F. Ethical Considerations

To ensure ethical compliance and user privacy:

- Following processing, video files and processed data are immediately erased after being temporarily stored.
- The application respects data privacy laws by not storing or disclosing user information.

G. Tools and Technologies

The following tools and APIs are used for system implementation:

- To extract the captions from YouTube videos, use the get_youtube_transcript function.
- Processing of Audio and Video: FFmpeg is used to manage local video files.
- **S**ummarization**:** Marathi and other language summaries can be produced using the Gemini API.

- Translation: When necessary, use neural machine translation tools or APIs.

IV. WORKFLOW

The Fig: 1 shows a process of summarizing the video content coming from a YouTube URL or a locally uploaded video file. The source of video is selected by the user in the beginning. If a YouTube URL is provided, the captions are fetched directly from the video, summarized, and translated if required. In the case of a local video file uploaded, the audio is extracted from the file, transcribed to text, summarized, with possible translation. In both cases, the summarized content is shown to the user. This allows effective summarization of video content into highlights, and presentation in the user's chosen language. Finally, the workflow ends by displaying the summary.

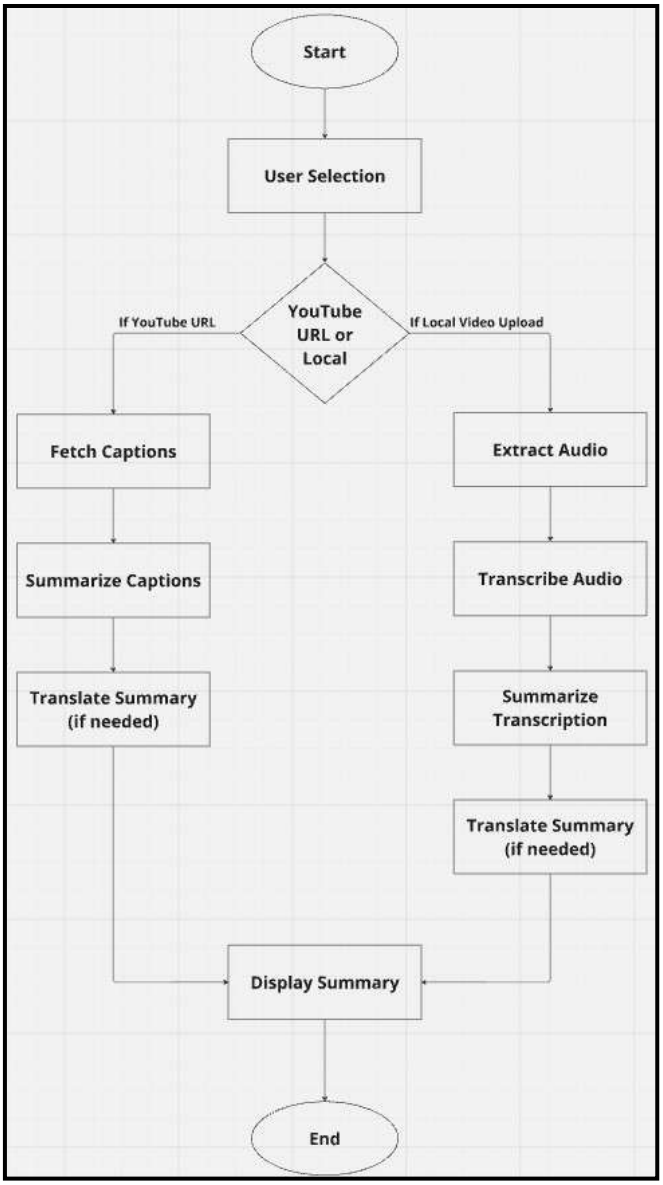

Fig 1 : Workflow

V. RESULTS

The Fig:2 below illustrates the interface for a Video Summarizer Toothat does give a user the possibility of uploading a video or providing a YouTube URL to generate an instant summary in plain language or the language of their choice. Compatible formats for this software are MP4, WebM, and AVI. Users can upload video files through drag-and-drop or browsing

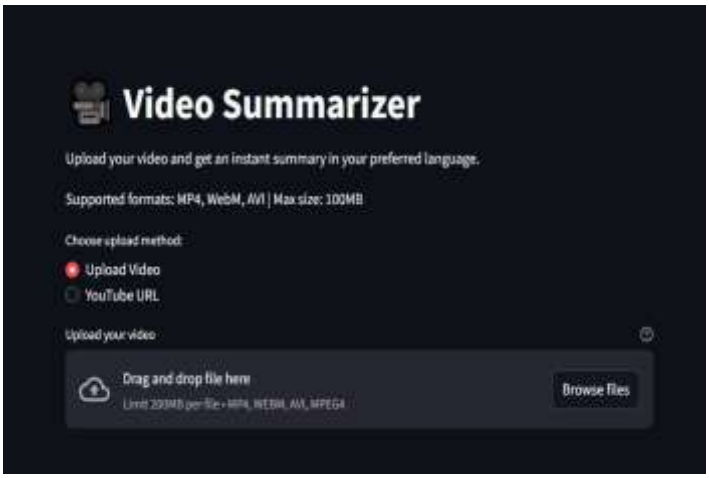

Fig 2 : Home page

The below in Fig. 3, the user selects an output language-Marathi. The user has also to click on the "Generate Summary" button to make the summary of video in the language of choice..

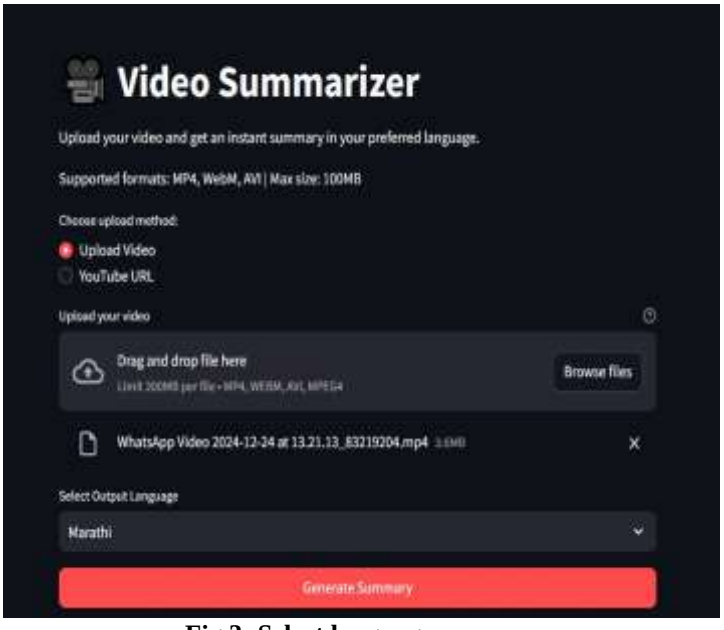

Fig 3: Select language

The below fig: 4 shows the summary of the video which is upload by the user.

Fig 4 : Summary generated

VI CONCLUSION

The video summarization tool combines transformation-based algorithms, the Gemini API for summarization, and Streamlit for an interactive User Interface. It seems fresh, & this gradient of need in the use of summarization tools becomes apparent, particularly when they are turned towards regional languages like Marathi. Through the transformational techniques, the system does what needs to be done-& that is ensure that the text to be transcribed is appropriate for summarization, producing concise and meaningful summaries that capture the context.

The Streamlit makes it user-friendly-the video can be uploaded, and summaries may be fetched relatively straightaway. This system is helpful for the summary of any long-length Marathi into educational lectures or news reports or documentaries, thus saving a good amount of time and energy for the user.

To conclude, the project provides a working framework for video summarization in Marathi, leaving space for future prospects like real-time summarization, support in multiple languages, and coordination with various AI-based tools to come up with a wider reach.

REFERENCES

1. H . Wang, B.Zhou, Z . Zhang, D. Yiming , D. Ho, and K.Wong, “M3Sum: A Novel Unsupervised Language-Guided Video Summarization”. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024
2. 2 D . Argaw1, S . Yoon, H . Deilamsalehy, ” Scaling Up Video Summarization Pretraining with Large Language Models ”. April 4, 2024
3. S . Zang ,H . Jin, Q . Yu, S . Zhang, and H . Yu, “Video Summarization Using U-shaped Non-local Network”. International Journal of Network Dynamics and Intelligence, 2023.
4. P . Saini, K . Kumar,S . Kashid, A . Saini, A . Negi,” Video summarization using deep learning techniques: a detailed analysis and investigation”. Published online: 15 March 2023
5. Prof. K . Hande , H . Karlekar , P . Yeole , A . Likhar and H . Rangari, “NLP based Video Summarisation using Machine Learning” 2023 International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET),India,2023.
6. P . Ilampiray , A. Thilagavathy, A . Nithin, ,I . Raj, ”Video Transcript Summarizer.” E3S Web of Conferences, vol. 399, p. 04015. EDP Sciences, 2023.
7. V . Kadam, N . Mahender,”A text summarization system for marathi language”, Dr. Babasaheb Ambedkar Marathwada University, April 2023.
8. S . Patil , S . Yadav , S . Shinde , D . Waghmare , R . Patil Prof. S. A. Babar,“Video Transcript Summarization in Marathi”,International Journal of Advanced Research in Science, Communication and Technology (IJARSCT),Volume 2, Issue 6, June 2022.
9. G . Begum, M . Sultana, D .Ashritha ,”Youtube Transcript Summarizer” International Creative Research Thought,2022.
10. S . Rai, S . Gagana , K . Vedhavathi ,N . Kiran ”Video Summarization using NLP”, International Research Journal of Engineering and Technology (IRJET) ,Volume8, Issue 8,,Karnataka, India,2021.