International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Agentic AI for Automated Video Summaries and Subtitle Creation

Prof. Ravindra Patil[1], Prof. Ekata Shanbhag[2], Sameerahamad Bagalad[3], Vinay Kumbar[4], Basappa Ningappa Vajramatti[5], Satish Gawada[6]

[12]Project Guide, Professor, KLS Vishwanathrao Deshpande Institute of Technology, Haliyal, India. [3456]BE Students, Department Of Computer Science and Engineering (AI & ML), KLS Vishwanathrao Deshpande Institute of Technology, Haliyal, India.

Abstract -This paper introduces an agentic artificial intelligence framework designed for the automated analysis, summarization, and multilingual subtitle generation of video content. The system integrates advanced large language models (LLMs) from Groq, leveraging their capabilities for high-speed transcription, comprehensive insight extraction—including key topics, actionable items, critical questions, and emotion breakdowns—and robust multilingual translation. Complementing these capabilities, the framework incorporates ElevenLabs for high-quality text-to-speech summary playback and a Streamlit-based interactive user interface. This holistic tool efficiently processes videos from diverse sources (uploaded files or YouTube URLs), generating accurate transcripts, insightful summaries, and dynamically translated subtitles in languages such as English, Hindi, and Kannada. The proposed framework holds significant promise for enhancing content accessibility and streamlining information digestion across educational, media, and business sectors.

1. INTRODUCTION

Throughout history, writing has been the most essential way for people to communicate ideas with each other and preserve information. From early humans carving marks on cave walls to modern users typing on digital screens, the practice of writing has transformed alongside every major technological shift. What began with simple pen-and-paper tools later expanded to chalkboards, whiteboards, touchscreens, and digital styluses each offering quicker, cleaner, and more convenient ways to express thoughts. Yet all of these methods still depended on a physical surface. As computers, cameras, and sensing technologies advanced, researchers started exploring ways to interact with machines without actually touching them. Modern systems can now pick up gestures like a wave, a point, or the movement of a finger— opening the door to new, contact-free forms of writing and interaction. This led to a broad area of study called contact-free interaction, whereby, gestures, rather than contact, prompt action through a viewer or vision. Writing in the air is an example of this development. Instead of using a stylus or keyboard, an individual can simply move their hand and the movement is traced on screen. As less physical devices are needed, hygiene is improved in shared environments while providing ease for individuals who have trouble using conventional writing devices. Plus, it will inspire creative experimentation in classrooms, exhibitions, and learning spaces. This, after all, is not a new concept, today, it has moved out of research labst4 With a regular webcam and some common software libraries, anyone can create a simple system to track only finger motion and convert that motion to written characters. This project follows this same premise. It will create a setup to detect a person's hand, and track their finger, to recreate what that individual virtually writes, resulting in another efficient, portable, and interactive way to write without contact. Finally, it simply shows how simple computer vision techniques help make daily tasks a little more flexible and contemporary.

2. SYSTEM ARCHITECTURE AND METHODOLOGY

[2.1] Overall System Design

The system's interactive web interface is developed using Streamlit, facilitating user interaction through video file uploads or YouTube URL submissions. Upon input, the video's audio track undergoes extraction and is subsequently segmented into manageable chunks to optimize processing efficiency for larger files. These audio segments are then submitted to Groq's Whisper API for highly accurate speech-to-text transcription. The resulting English transcript serves as the foundational data for subsequent AI-driven analysis, orchestrated by Groq's LLM (llama-3.1-8b- instant). This analysis yields comprehensive insights, including detailed summaries, identification of key topics, extraction of actionable items, derivation of critical questions, and an emotional tone breakdown. All textual outputs—transcripts and insights—are subject to multilingual translation into user-selected target languages (e.g., Hindi, Kannada, English) utilizing Groq's robust translation capabilities. The framework culminates in dynamic WebVTT subtitle generation, PDF report export of summaries, and audio playback of detailed summaries powered by ElevenLabs.

[2.2] Audio Extraction and Chunking

For both direct video uploads and YouTube content, the initial processing stage involves isolating the audio track. The moviepy library is employed for efficient audio extraction from video files, converting the audio stream into a WAV format. To ensure robust handling of extended video durations and to mitigate potential API size constraints, the extracted audio is further processed using the pydub library. This allows for precise segmentation of the audio into smaller, uniform chunks (e.g., 2-minute intervals). This strategic chunking mechanism is critical for ensuring consistent and reliable processing by

downstream transcription services, particularly for lengthy inputs.

[2.3] Groq-Powered Transcription

The segmented audio chunks are then transmitted to the Groq API for transcription. The system leverages Groq's optimized implementation of the whisper-large- v3 model, renowned for its superior speed and accuracy in automatic speech recognition. Each audio chunk is transcribed independently, and the resultant segments, comprising transcribed text with precise start and end timestamps, are meticulously reassembled. The timestamps associated with individual chunks are carefully adjusted maintain temporal accuracy within the overarching video timeline, thereby producing a cohesive, time-aligned comprehensive transcript.

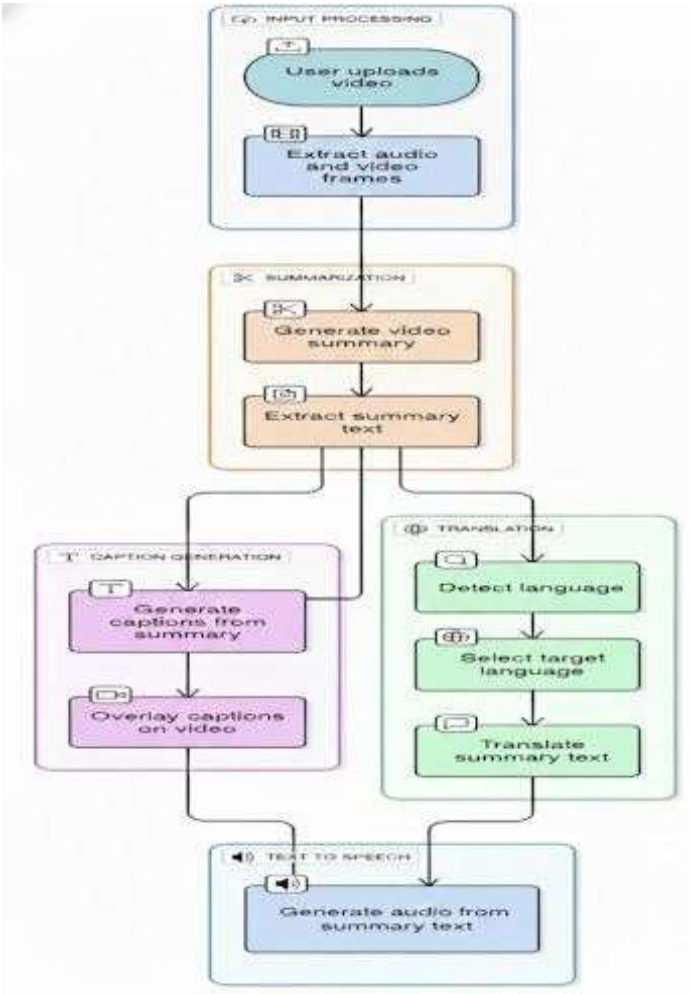

Figure 1: System Architecture Flowchart

[2.4] Agentic AI for Insights Generation

A pivotal feature of this framework is the intelligent generation of comprehensive insights from the preliminary transcript. The Groq API, specifically employing the llama-3.1-8b-instant model, is configured to operate as an "expert analysis assistant." It processes a canonical English version of the transcript (derived from translating the raw input if it contains 'Hinglish' or other languages) and outputs a structured JSON object. This object encapsulates:

- **TL;DR Summary:** A succinct, single-sentence encapsulation of the entire content.
- **Detailed Summary:** A multi-sentence synopsis, the length of which is dynamically adjusted based on user preference (short, medium, or detailed).
- **Key Topics:** A curated list (typically 3-5) of the primary subjects discussed within the video.
- **Action Items:** An explicit enumeration of any tasks, decisions, or recommendations identified within the narrative.
- **Key Questions:** A list of the three most significant questions posed and addressed by the content.
- **Emotion Breakdown:** A quantitative analysis presenting the percentage distribution of detected emotional tones (e.g., Joy, Anger, Sadness, Strict, Neutral).

This structured analytical output significantly expedites content comprehension and facilitates the derivation of actionable intelligence.

[2.5] Multilingual Translation

The system provides robust multilingual capabilities applicable to both the complete transcript and the generated insights. Groq's llama-3.1-8b-instant model is further utilized for high-quality machine translation. When a target language distinct from English is selected, the framework systematically translates the entire clean English transcript, alongside each component of the generated insights (e.g., detailed summary, TL;DR, key topics, action items, questions), into the specified language (e.g., Hindi, Kannada). This comprehensive linguistic support ensures that users can access all derived information in their native or preferred language, thereby substantially enhancing accessibility and utility. The translation for subtitle generation is performed on a line-by-line basis to optimize contextual accuracy.

[2.6] Subtitle Generation (VTT) and PDF Export

To augment visual accessibility, the system dynamically generates WebVTT (.vtt) subtitle files. These VTT files are constructed from the time-aligned transcription segments, with the textual content translated into the user-designated subtitle language. The generated VTT file is then seamlessly integrated with the video player in the Streamlit interface, enabling synchronized subtitle display. Furthermore, the detailed summary is available for export as a Portable Document Format (PDF) file. The fpdf library is employed for this functionality, incorporating intelligent font selection mechanisms to ensure correct rendering of multilingual characters, including the dynamic downloading and utilization of specialized font such as NotoSansKannada and NotoSansDevanagari for accurate representation of Indian language scripts within the PDF document.

[2.7] Audio Read-Aloud with ElevenLabs

To further enhance content consumption and accessibility, the framework integrates ElevenLabs for advanced text-to-speech (TTS) synthesis. Users are presented with a selection of high-quality synthetic voices (e.g., Rachel,

Adam, Charlotte, Freya) to have the generated detailed summary read aloud. The ElevenLabs API's text_to_speech.stream method, powered by the eleven_multilingual_v2 model, facilitates the real-time streaming of generated audio directly to the user's output device, thus offering an auditory pathway for accessing the summarized content.

3. IMPLEMENTATION DETAILS AND RESULTS

The entire system is deployed as a web application orchestrated within the Streamlit framework, which provides an intuitive and responsive user interface.

[3.1] User Interface

The Streamlit interface features a well-organized sidebar where users can select input methods (video file upload or YouTube URL), specify the desired output language, and adjust the granularity of the generated summary. The main content area dynamically displays the video player, complete with integrated subtitles, the comprehensive transcription, and the structured insights. These insights are compartmentalized into distinct tabs for ease of navigation, presenting the summaries, key topics, actionable items, critical questions, and emotion analysis. Interactive buttons allow users to initiate transcription, trigger insight generation, enable subtitle display, and activate audio playback of summaries, culminating in a highly engaging and efficient user experience.

Figure 2: User Interface

[3.1] Performance and Accuracy

A significant advantage derived from the integration of Groq's inference engine is its exceptional processing speed. Both the transcription phase, leveraging whisper-large-v3, and the subsequent analysis and translation tasks, powered by llama-3.1-8b-instant, are executed with remarkable rapidity. This efficiency renders the system highly effective for processing even extended video durations. The implemented audio chunking mechanism further contributes to robust performance by mitigating the risk of timeouts and processing failures often associated with very large input files.

The accuracy of both transcription and the subsequent summarization is maintained at a high standard, attributable to the advanced capabilities of the underlying LLMs. While providing a useful overview, the emotion analysis, by its nature, offers a generalized approximation of sentiment.

[3.2] Multilingual Support

The system provides extensive multilingual support, specifically encompassing English, Hindi, and Kannada, addressing a critical need for diverse linguistic accessibility. This support is manifest across several key functionalities:

- **Transcription Base:** The system intelligently handles mixed-language inputs (e.g., "Hinglish") by first translating them into a standardized clean English base for subsequent analysis, ensuring consistency and accuracy.
- **Output Transcripts:** Complete video transcripts are generated and presented in the user's selected output language.
- **Insights Translation:** All components of the generated insights (TL;DR, detailed summary, topics, action items, questions) are meticulously translated into the target language.
- **Dynamic Subtitles:** WebVTT subtitles are dynamically generated in the chosen language, directly enhancing the video's accessibility for non- native English speakers.
- **PDF Exports:** Summary PDFs are capable of accurately rendering multilingual text, facilitated by the dynamic downloading and utilization of appropriate fonts for specific character sets.

This comprehensive linguistic flexibility significantly enhances the tool's value proposition for global content analysis and consumption.

Figure 3: AI-Generated Summary in Kannada

Figure 3: AI-Generated Summary in English

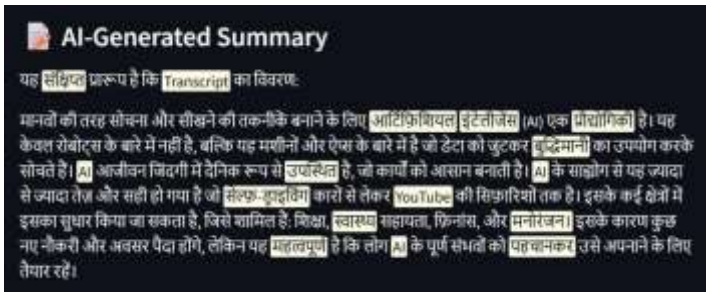

Figure 4: AI-Generated Summary in Hindi

4. DISCUSSION AND FUTURE WORK

[4.1] Strengths and Limitations

The presented framework's primary strengths reside in its unparalleled processing speed, comprehensive feature set, and robust multilingual capabilities. The strategic integration of Groq's high-performance LLMs ensures efficient and timely processing, while the detailed and structured insights provide a profound understanding of video content. The dynamic subtitle generation and audio summary features critically enhance content accessibility for a broad audience. *Authors and Affiliations*

However, certain limitations are inherent to the current implementation. The precision of the emotion analysis, though informative, represents a generalized approximation; more specialized, context-aware models might be requisite for highly nuanced emotional interpretations. Furthermore, while the general quality of translations is high, it can occasionally be impacted by highly idiomatic expressions or culturally specific contexts, especially when translating lists or short segments in isolation. The system's reliance on external API keys necessitates stringent security protocols for deployment and key management.

[4.1] Potential Enhancements

Future work will focus on several avenues for further development and refinement. Implementing speaker diarization would be a significant enhancement, enabling the attribution of transcribed text to individual speakers, which is particularly valuable for analyzing meetings or interviews. Integrating multimodal analysis, wherein visual cues from the video are considered in conjunction with audio information, could yield even richer insights, especially for identifying key events or non-verbal communication. Developing a conversational agent interface to allow users to iteratively refine or query the generated insights would empower a more dynamic content exploration experience.

Expanding the repertoire of supported output languages, potentially through an adaptive mechanism for fetching translation model capabilities, would further broaden the system's global utility. Finally, optimizing the font downloading and rendering mechanism for PDF generation to provide more robust support for an even wider array of Indic scripts or other complex character sets is an area for continuous improvement.

5. APPLICATIONS AND USE CASES

The proposed agentic AI framework demonstrates wide applicability across diverse domains that rely heavily on video-based content dissemination and analysis.

- **Educational Sector:**
  In academic institutions, lecture videos and online course materials can be automatically summarized into concise, topic-wise insights, enabling students to quickly revise large volumes of content. The multilingual subtitle generation enhances inclusivity by supporting learners from various linguistic backgrounds. Moreover, emotion analysis can help instructors assess engagement and emotional tone during lectures, leading to improved teaching methods.

- **Media and Journalism:**
  For media houses and news broadcasters, the framework can be employed to automatically transcribe and summarize interviews, news reports, and press conferences. Journalists can quickly identify key statements, extract quotes, and generate region-specific subtitles, reducing manual labor and improving publication turnaround time.

- **Corporate and Business Applications:**
  Organizations can utilize the system for meeting recordings, webinars, and training sessions to produce detailed summaries and actionable insights. The emotion breakdown component aids in understanding communication dynamics during meetings, supporting decision-making and performance assessment.

- **Accessibility and Public Services:**
  Government and public service institutions can leverage this framework to enhance accessibility for citizens with hearing impairments or language barriers. Automatic multilingual subtitling ensures that official announcements and educational content reach a broader audience efficiently.

6. CONCLUSIONS

We have successfully developed and presented an agentic AI framework for automated video analysis that profoundly streamlines the processes of information extraction, summarization, and translation from video content. By synergistically combining Groq's high- performance LLMs for transcription, summarization, and multilingual translation with ElevenLabs for advanced audio playback and Streamlit for an intuitive user interface, this system emerges as a powerful and highly accessible tool. Its distinctive ability to furnish detailed insights, generate multilingual transcripts, and provide dynamic subtitles caters effectively to a diverse user base, thereby rendering video content more digestible and globally accessible. This work underscores the practical and transformative application of cutting-edge AI technologies in enhancing content comprehension and accessibility across a multitude of domains.

7. ACKNOWLEDGMENT

We are indebted to our Principal Dr. V. A. Kulkarni and management of KLS VDIT for providing an environment

with all facilities that helped us in completing the major project. We are extremely grateful to Dr. Poornima Raikar, HoD of the Computer Science and Engineering(AI & ML) Department for her moral support and encouragement.We wish to express our sincere gratitude to our guide Prof. Ravindra Patil from the Computer Science & Engineering (AI & ML) Department, for their guidance and suggestions. We thank all the teaching and non-teaching staff of the Department of Computer Science and Engineering for their kind help. Last but not the least, We would like to add some personal notes. If there is a driving force that keeps us going, and what has not changed, it is the constant support and blessing of our parents, family and friends. There is no doubt, in spite of our strenuous efforts, error might remain in the major-project report. Naturally, We take full responsibility for any lack of clarity, occasional

8. REFERENCES

[1] Hsu, W., Garg, N., & Zhang, C. (2021). Whisper: Robust Speech Recognition via Weak Supervision. OpenAI. Retrieved from https://openai.com/research/whisper

[2] Nallanthighal, V. S., Hegde, R. M., & Murthy, H. A. (2020). Speech Summarization Using Hidden Markov Model Based Content Selection and Prosodic Analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2647–2659.

[3] Xu, Y., Bai, S., Zhang, M., & Chen, J. (2021). Video Summarization with Attention-based Encoder-Decoder Networks. Multimedia Tools and Applications, 80, 13379–13395.

[4] Lin, C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop.

[5] Potapov, D., Douze, M., Harchaoui, Z., & Schmid, C. (2014). Category-specific Video Summarization. In European Conference on Computer Vision (ECCV), 540–555.

[6] Zhang, K., Chao, W. L., Sha, F., & Grauman, K. (2016). Summary Transfer: Exemplar-based Subset Selection for Video Summarization. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1059–1067.

[7] Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed. draft). Stanford University. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

[8] Chollet, F. (2017). Deep Learning with Python. Manning Publications.

[9] Brownlee, J. (2020). Deep Learning for Natural Language Processing: Develop Deep Learning Models for NLP. Machine Learning Mastery.

[10] OpenAI. (2022). OpenAI Whisper Documentation. GitHub Repository. Retrieved from https://github.com/openai/whisper