INTERNATIONAL JOURNAL OF RESEARCH CULTURE SOCIETY
Monthly Peer-Reviewed, Refereed, Indexed Journal
Volume - 9, Issue - 3, March - 2025
ISSN(O): 2456-6683
[Impact Factor: 9.241]

DOIs:10.2017/IJRCS/202503017 --:-- Research Paper / Article / Review

Video Text Summarizer

[1] Afreen Shah, [2] Dr Rakhi Gupta
[1] MScIT Student, [2] Chairperson HOD Dept. of IT
[1] Dept of Information Technology, Kishinchand Chellaram College, Mumbai, India
[2] Dept of Information Technology, Kishinchand Chellaram College, Mumbai, India
Email - [1] afreenshahafreenshah07@gmail.com , [2] rakhi.gupta@kccollege.edu.in

Abstract: *In today's digital landscape, video has become a dominant medium for communication and information sharing. Extracting key insights from lengthy videos, such as educational lectures, is often challenging and time-consuming. The proposed Video Text Summarization System addresses this challenge by automating the creation of concise and informative text summaries from video content. The system follows a structured, multi-stage process: audio is first extracted from videos, then converted into text using advanced speech-to-text technology. The resulting transcripts are refined through comprehensive text preprocessing to ensure clarity and coherence. By leveraging both extractive and abstractive summarization methods, the system condenses essential information efficiently. Scalable and adaptable, this solution is ideal for applications in education, professional environments, and content creation.*

Key Words: *Video Summarization, Text Summarization, Speech-to-Text, Audio Extraction, Multi-Stage Process, Extractive Summarization, Abstractive Summarization, Educational Videos.*

1. INTRODUCTION:

Thus, the need for efficient and automated methods of extracting data from videos has emerged as a result of the upsurge in videos in areas like education, entertainment, health care, business, and social media, among others. Since YouTube produces millions of videos daily while MOOCs, video conferencing tools, and video-on-demand services generate mass data of videos, it is increasingly becoming cumbersome for the users to scroll through long videos to find relevant information. The improvement of this task can be addressed by video text summarization, which processes and presents a video in short yet meaningful textual form by including audio information, visual signals, and any textual data that may be integrated into the video. Through these summaries, users are able to make a quick overview of important aspects from the videos without having to watch a video. However, it is easier to identify the goals of the VTS compared to TTS because VTS has to process a set of different data at once: speech, images, and text displayed on the screen. [3] AI, NLP, and computer vision have given new approaches for particular lines: video text summarizations, which involve extractive techniques, which include keyframes and dialogues, while there are abstractive techniques that involve synthesizing core infinitives from videos. [2]There are, however, several more issues to be solved, including those related to the handling of large-scale video data, different languages and accents, the consideration of the semantic context of scenes, and the coherence and relevance of the generated summaries. The paper overviews current approaches to identifying the text in videos, reviews critical issues, and considers uses from automated content production to enhancing facilities for viewers with disabilities. Since videos are being generated in large quantities, there is a need for summarization techniques to improve the usability of the videos. The demand for effective and automated extraction of information from videos has become crucial due to the rapid expansion of video content in various fields such as education, healthcare, corporate training, and social media. YouTube, video conferencing tools, and video-on-demand services produce large volumes of video data on a daily basis, making it more and more impractical for users to manually search through lengthy videos to locate pertinent information.

2. LITERATURE REVIEW/ GAP ANALYSIS:

Author	Year	Title	Gap

Habib Khan, Tanveer Hussain	2023	Deep multi-scale pyramidal features network for supervised video summarization	Mention aspects like the need for enhanced generalizability to diverse datasets, or improving scalability and efficiency without sacrificing performance.
Guoqiang Liang, Yanbing Lv	2021	Video summarization with a dual-path attentive network	Dependence on frame-level importance scores may lead to cumulative errors.
Haoran Li, Junnan Zhu	2019	Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video	Our multi-modal summarization method, combining NLP, speech processing, and computer vision, improves the quality of multimedia news summarization by bridging semantic gaps between audio and visual content.

Table: 1

3. METHODOLOGY:

3.1 Overview of the Proposed System:

The Video Text Summarization System automates the summarization of video content. It extracts audio from video files, converts speech to text, and summarizes the text using advanced techniques.

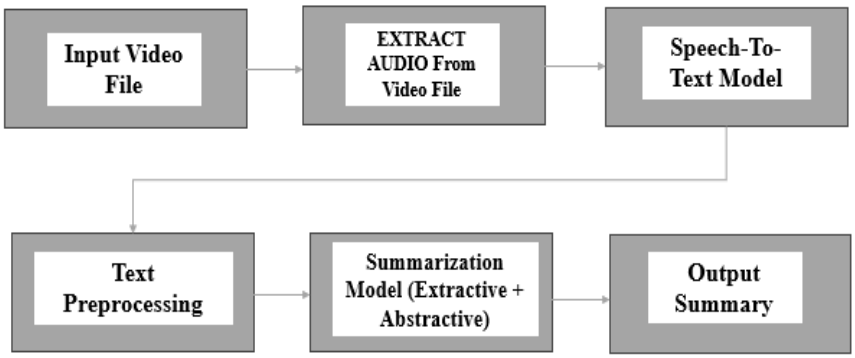

Figure 1: Workflow of the Video Text Summarization System

3.2 Speech-to-Text Component:

This component employed the use of automatic speech recognition (ASR) technology in the conversion of the extracted audio from videos to text. [1] Popular models like Whisper or DeepSpeech will be used to ensure high accuracy of the transcription regardless of the quality of the audio signal. The output produced by the ASR system will be utilized in the next step of the summarization process of the developed architectures.

3.3 Text Summarization Component:

After the text has been produced, the text summarization component will then engage both methods of extractive and abstractive summarization. [4] Both extractive and abstractive methods will extract major sentences from the transcription of the video while paraphrasing these ideas in the form of new sentences. For this purpose, advanced models such as BERTSUM and GPT-3 will be used, which make it possible to provide complex and cohesive summaries.

3.4 Data Collection and Sample Size:

Primary Data: Video Files: 10-30 files of videos of 5 minutes to 1 hour of content will be chosen for initial assessment from the domains of Internet Educational Lectures, News Reports, and Interviews. Secondary Data: The datasets used in this study include the How2, YouTube-8M, and TED-LIUM datasets, which will be used, and they are pre-existing datasets.

DATA	SOURCES	SAMPLE SIZE
Primary Data	Collected from videos (e.g., YouTube, educational platforms)	10 videos, with transcripts and summaries generated by system

	Social Media platforms (Instagram Reels)	10-20 short videos(1-5 minutes)
	Course Video	5-10 videos (20-60 minutes each)
Secondary Data	Existing video datasets (e.g., YouTube-8M, How2 Dataset)	10-20 videos for benchmarking and evaluation in the first phase
	TED-Lium Dataset	10-20 videos
	AMI Meeting Corpus	10-15 meeting videos

Table: 2

Sample Size: Out of 100 to 200 videos, more will undergo validation and performance assessment where summaries will be compared with ground truth. Key findings: Primary and secondary data

4. ANALYSIS:

4.1 Evaluation Metrics:

To measure the quality of the generated summary, we use:

- ROUGE Scores (Precision, Recall, F1-score) – Measuring content retention.
- BERTScore—Assessing semantic similarity.
- BLEU Score—Evaluating fluency.

4.2 Comparison Metrics:

Metric	**Youtube Video**	**Local Video**
ROUGE-1 (F1)	0.601	0.482
ROUGE-2 (F1)	0.595	0.470
ROUGE-L (F1)	0.601	0.482
BERTScore (F1)	0.739	0.741
BLEU Score	N/A	0.120

Table: 3

4.3 Visual Representation of Performance Metrics: To support the numerical evaluation, the below figures provide a direct visualization of the performance metrics displayed in the system's output.

```
ROUGE-1: Precision: 1.000, Recall: 0.429, F-measure: 0.601
ROUGE-2: Precision: 0.990, Recall: 0.425, F-measure: 0.595
ROUGE-L: Precision: 1.000, Recall: 0.429, F-measure: 0.601
BERT-Scores: Precision: 0.734, Recall: 0.745, F1-score: 0.739
```

Figure 2: Accuracy Of Online Video Summary(YouTube)

```
ROUGE-1: Precision: 1.000, Recall: 0.318, F-measure: 0.482
ROUGE-2: Precision: 0.975, Recall: 0.309, F-measure: 0.470
ROUGE-L: Precision: 1.000, Recall: 0.318, F-measure: 0.482
BERT-Scores: Precision: 0.730, Recall: 0.752, F1-score: 0.741
BLEU Score: 0.120
Readability (Flesch Reading Ease): 33.040
Compression Ratio: 0.318
```

Figure 3: Accuracy Of Local Video Summary(.mp4)

5. CONCLUSION:

This research introduced a video text summarization system dealing with the problem of summarizing video content through speech-to-text transcription and text summarization breakthroughs. The system also improves content consumption efficiency since it cuts the time one has to spend to fish for the most relevant details in rather long videos. The future enhancements will be directed towards the enhancement of the algorithms and the expansion of the possibilities of the use of the system for other platforms and kinds of videos.

6. LIMITATION:

Limited Dataset Diversity: The dataset lacks sufficient variety in video formats and domains.
Small Sample Size: The dataset size is relatively small for comprehensive evaluation.
English-Only Focus: The system primarily supports English, limiting multilingual applicability.
Challenges with Complex Content: Struggles to summarize highly technical or abstract videos accurately.
Noise in Audio: Performance issues with noisy or low-quality audio recordings.

7. RECOMMENDATION:

Expand Dataset Variety: Add diverse content like vlogs, live streams, and cultural videos.
Increase Sample Size: Include hundreds of videos across various domains for validation.
Enable Multilingual Support: Extend capabilities to multiple languages using multilingual datasets.
Enhance Audio Processing: Incorporate noise reduction and robust preprocessing techniques.
Real-Time Summarization: Develop real-time capabilities for live meetings or broadcasts.
Domain-Specific Models: Create specialized models for technical and context-heavy content.
Improved Evaluation: Establish objective evaluation metrics and ground truth summaries.
User Customization: Offer options for customized summaries based on user preference.

REFERENCES:

1. Habib Khan, Tanveer Hussain, Samee Ullah Khan, Zulfiqar Ahmad Khan, Sung Wook Baik, (2024): Deep multi-scale pyramidal features network for supervised video summarization. *Expert Systems with Applications*, 237, Part C.
2. Guoqiang Liang, Yanbing Lv, Shucheng Li, Xiahong Wang, Yanning Zhang, (2022): Video summarization with a dual-path attentive network. *Neurocomputing*, 467. ISSN 0925-2312.
3. Li H., Zhu J., Ma C., Zhang J., Zong C., (2019): Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video. *IEEE Transactions on Knowledge and Data Engineering*, 31(5), 996-1009, 1 May.
4. Tyagi T., Dhari L., Nigam Y., Nagpal R., (2023): Video Summarization using Speech Recognition and Text Summarization. *4th International Conference for Emerging Technology (INCET)*, Belgaum, India.
5. Apostolidis E., Adamantidou E., Metsai A. I., Mezaris V., Patras I., (2021): Video Summarization Using Deep Neural Networks: A Survey. *Proceedings of the IEEE*, 109(11), 1838-1863, November.