An Intelligent System for Effective Knowledge Extraction and Summarization from Videos

Vilas K N
Department of MCA
Ramaiah Institute of Technology
Bangalore, India
vilas147854@gmail.com

Maya A K R
Department of Computer Science and Applications
Bangalore University
Bangalore, India
mayasanjay@gmail.com

***Abstract*—In this age of digital information, organizing and extracting meaningful insights from massive amounts of video content remains a major challenge. The exponential growth of video content has led to information overload, making it challenging to access and retrieve essential information efficiently, particularly for users with limited time or attention spans. This work describes an intelligent system for effective knowledge extraction and representation from videos, a novel solution aimed at addressing this issue by combining modern deep learning techniques with Automatic Speech Recognition (ASR) technology. The core functionality of the system involves transcribing spoken language in videos into accurate text, which serves as the foundation for extracting essential knowledge that provide key insights along with succinct summaries. By implementing powerful algorithms for video analysis, transcribed text generation, summarization, question answering, keyframe extraction, and identifying significant segments, the system facilitates quick content retrieval and deeper analysis without the need to view entire videos. This work is essential as it provides a comprehensive method to enhance the accessibility and usability of video content, enabling efficient summarization and knowledge extraction. The findings show that the system has the potential to transform how people engage with and use video content, making it an invaluable resource for academics, educators, and professionals from a variety of sectors.**

***Index Terms*—Automatic Speech Recognition (ASR), Deep Learning, Natural Language Processing (NLP), Video Summarization**

I. Introduction

The human approach to searching through extensive video content for relevant segments is time-consuming and impracticable, particularly given the current rate of content development. The rapid increase in video content across various platforms necessitates efficient methods for summarization and knowledge extraction. In this fast-growing era, saving time is crucial for both individuals and businesses to stay competitive and productive. Consequently, the need for automated tools in video summarization and knowledge representation has become more pressing. Automated video analysis not only conserves valuable time but also provides deeper insights into the content, enabling users to extract meaningful information without the need to watch the entire video.

The video summarization technology has always been an area of interest for researchers, and the need of it is clearly emphasized and addressed [1] leading to a significant advancement in this field. The earlier literary works mainly focused on providing abstracts using video skimming, keyframe extraction, and techniques for low-level feature extraction [2-4]. Video skimming mainly provides video summary by putting together important clips as a video segment, thereby reducing the video length and saving time. Even though it helps to have a visual summary with images and sounds, it lacks synchronization, just like a movie trailer. Hence, most of the works are focused on summary-oriented skim rather than highlight-oriented. Some of the recent works on video skimming focus on enhanced ways of providing summary by concentrating on positive frames rather than negative frames and using localization methods for precise boundaries of segments [5].

In keyframe extraction, the important frames or images are extracted to provide a quick summary of the video. Initial works on keyframe extraction mainly involved a shot-based approach, but it is not suitable for long videos [6–8]. Various clustering-based techniques were also explored by researchers for effective summarization [9–10]. With the advancement of technology, the field has progressively shifted from the conventional approaches to enhanced deep learning and transfer learning techniques for more robust video analysis [11]. However, each of these approaches has traditionally focused on isolated aspects of video content, limiting the overall understanding and usability of the extracted information. Our proposed work aims to overcome these limitations by providing a comprehensive tool that integrates multiple features, thereby offering deeper insights and extensive knowledge extraction from videos.

Our proposed work aims to summarize extensive video content into brief, informative summaries through advanced video analysis, textual summary generation, and audio extraction. Central to our hybrid approach is the use of Automatic Speech Recognition (ASR) technology, which converts spoken language in videos into accurate transcribed text. The transcribed text from ASR forms the foundation for generating concise summaries and extracting essential insights, allowing for faster material retrieval and deeper analysis without the need to watch complete videos. By employing cutting-edge natural language processing (NLP) and computer vision techniques, the system aims to provide efficient and effective methods for

users to access and retrieve important content from videos quickly.

Additionally, we aim to enable advanced search capabilities where users can search for specific keywords and find related text being discussed. This functionality not only matches sentences with the exact keyword but also extracts sentences containing words similar to the keyword using semantic search techniques, along with generating relevant question-answer pairs. We also plan to explore various summary representations to provide a multi-faceted approach to summarization. The contribution of our work lies in the integration of these objectives into a singular, comprehensive tool for video summarization and knowledge extraction.

The remaining sections of this work are organized as follows: Section II provides a detailed literature survey on video summarization techniques and related works. Section III describes the system design and architecture of the proposed system. Section IV outlines the implementation details, including the tools and technologies used. Section V presents the results and discussion, highlighting the performance and effectiveness of the proposed system and Section VI wraps up the paper with conclusions and future prospects.

II. Literature Survey

This literature survey focuses on key areas relevant to the project's objectives, including ASR for transcription, video summary generation, keyframe extraction and text summarization techniques. Even though key frame extraction techniques and deep learning models such as generative adversarial networks (GANs), autoencoders, and transformers have been applied to enhance video summarization [6–11], our main focus on this paper is ASR-generated transcribed text summarization, where we mainly focus on transcribed text to get the summaries. So, our literature review explores the papers related to ASR, followed by NLP papers on summarization.

A. Related Works

The initial efforts in ASR were primarily experimental, focusing on simple pattern recognition tasks. In the 1970s, researchers began to explore more sophisticated statistical methods, to improve speech recognition accuracy. The works of Rabiner and Juang [12] laid the foundational theories of ASR, focusing on Hidden Markov Models (HMM). The introduction of Hidden Markov Models (HMMs) marked a significant breakthrough in ASR technology. HMMs provided a robust framework for modeling temporal sequences, making them well-suited for speech recognition tasks. Building on this foundation, Hinton et al. [13] introduced deep neural networks (DNN) for acoustic modeling, significantly enhancing recognition accuracy. Graves et al. [14] further advanced the field by incorporating deep recurrent neural networks (RNN), which improved the system's ability to handle sequential data. The research works [15-17] pushed the boundaries further with end-to-end models, which streamlined the ASR process by directly mapping speech to text.

The emergence of machine learning and, later, deep learning revolutionized ASR technology. Artificial Neural Networks (ANNs) and, specifically, Deep Neural Networks (DNNs) were employed to model complex speech patterns more effectively. While DNN and end-to-end (E2E) models show impressive results in controlled settings, their performance often drops in noisy, real-world scenarios. The impact of noise had always been an area of concern for researchers [18, 19]. The works by J. Li and D. Wang et al. provide comprehensive overviews of ASR advancements, from hybrid models to E2E models, establishing a foundation for understanding current ASR systems [20,21]. Studies by L. Sari et al. and S. Feng et al. address the biases present in ASR systems, contributing to the discussion on fairness and equitable performance across different user groups [22,23]. The discrepancies highlight the need for more robust training datasets and improved noise-handling algorithms. Resolving these conflicts involves focusing on practical applications and testing ASR systems in diverse, real-world conditions.

The integration of ASR with text summarization techniques has drawn significant attention in recent years, particularly due to its potential in transforming spoken content into concise and meaningful summaries. The following review explores various methods for summarizing transcribed text generated by ASR systems, emphasizing the importance of these techniques in processing video content such as educational lectures, meetings, and multimedia presentations. Various approaches involved in textual summarization related to both abstractive and extractive summarization are explored by many researchers, emphasizing the need to choose the technique based on the domain [24, 25]. Even the complex tasks involved in textual analysis for grading purposes are a significant area of interest by researchers, where summarization and word embeddings play a major role. Our review thus emphasizes the need for summarizers based on text extracted through ASR from videos using NLP techniques and pre-trained models. Our review of question-answering (QA) systems was found to provide meaningful insights into various information and efficient knowledge extraction from text [26, 27].

The survey thus highlights significant progress in ASR technology and video summarization based on transcribed text, while also identifying challenges such as noise handling and computational complexity. The survey also helps us to conclude that NLP techniques, pre-trained summarizers and question answering method on transcribed text/summary can provide effective ways of knowledge extraction from videos. While significant progress has been made, challenges such as computational demands, subjectivity, and dataset limitations need to be addressed for further improvement. Future research directions may focus on developing more efficient models, creating large-scale annotated datasets, multimodal integration and adaptive learning techniques. The Table I provides a systematic review of the technologies used in ASR along with its limitations.

TABLE I
COMPARISON OF TECHNOLOGIES USED IN DIFFERENT SPEECH RECOGNITION SYSTEMS

System	Technology Used	Strengths	Weaknesses
HMM-based Systems	Hidden Markov Models - Utilizes statistical models to represent the probabilities of sequences of observed events.	Well-established, good for structured tasks	Limited accuracy, poor handling of noise
DNN-based Systems	Deep Neural Networks - Leverages deep neural networks to learn representations of data.	Improved accuracy, better feature learning	Computationally intensive, Vulnerable to overfitting if not properly managed
RNN-based Systems	Recurrent Neural Networks - Employs recurrent neural networks to handle sequential data and maintain context.	Good for sequential data, improved context, effective for continuous speech recognition	Vanishing gradient problem, requires large datasets, performance depends heavily on the quality of the training data
End-to-End Systems	Sequence-to-Sequence Models - Uses sequence-to-sequence models to directly map input speech to output text.	Simplified pipeline, direct speech-to-text	High computational cost, less interpretable
Hybrid Systems	Combination of HMM and DNN (Combines traditional HMM with modern DNN approaches)	Combines strengths of both models	Complexity in implementation, high resource requirement
Transformer Models	Self-attention Mechanisms	Superior in handling long-range dependencies	Requires extensive training data, complex architecture

III. SYSTEM DESIGN

The proposed system framework comprises of several major components, each responsible for different aspects of the video summarization process.

The main components include:

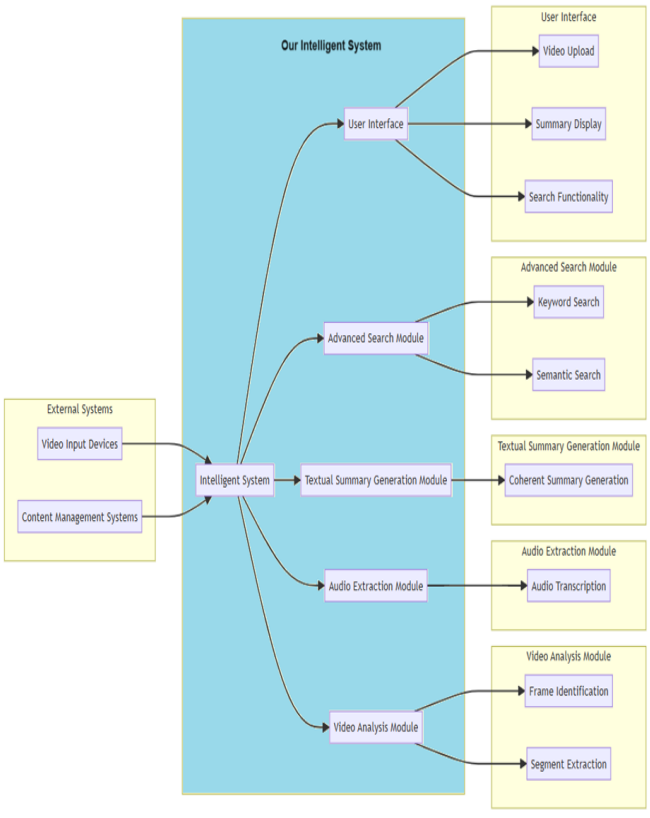

Fig. 1. Overview of the Functionalities of Proposed System.

- Video Analysis Module: Analyzes video content to identify significant frames and segments using computer vision techniques.
- Audio Extraction Module: Extracts audio from videos for transcription and subsequent summary generation.
- Textual Summary Generation Module: Utilizes NLP techniques to generate coherent and contextually relevant summaries from transcribed audio.
- Advanced Search Module: Allows users to search for specific keywords and related content using semantic search techniques.
- User Interface: Provides an intuitive interface for users to upload videos, view summaries, view keyframes, view question-answers and perform searches.

Fig.1 shows the major components of the overall system, subsystem interconnections, and external interfaces.

The system architecture is designed to handle the complete lifecycle of video content processing, from audio extraction to generating various forms of summaries and providing advanced search capabilities. The system's major functions include:

- Transcribed Text Generation: Extracting audio from videos and generating the corresponding transcribed text using our ASR model based on CNN and LSTM layers.
- Textual Summary Generation: Generating coherent textual summaries from transcribed text using both T5 and Gemini integrated summarizers.
- Video Analysis: Identifying significant keyframes and segments within videos using deep learning model (Caffe model) integrated with K-means clustering.
- Advanced Search Capabilities: Allowing users to search for specific keywords and similar words using NLP and semantic search techniques.
- Multiple Representation Methods: Providing diverse ways of summarizing video content to cater to different user

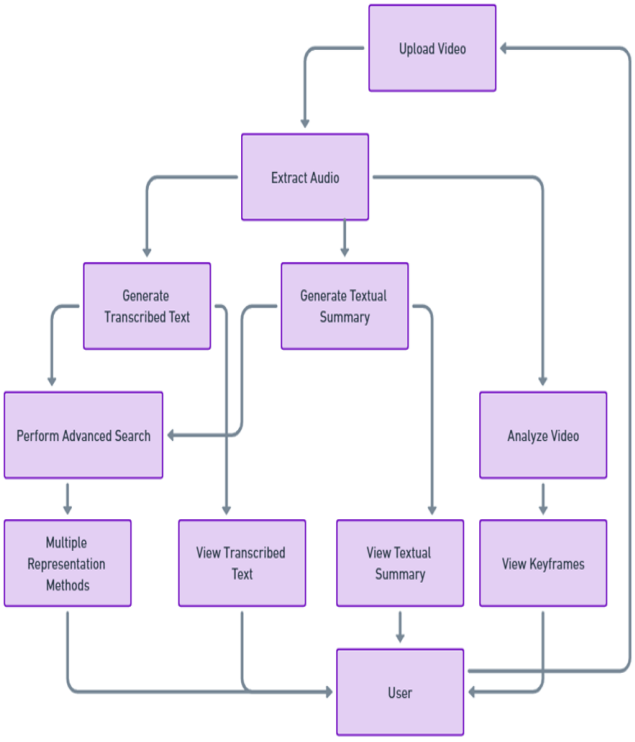

Fig. 2. Functional Diagram with major system functions.

needs including question answering of transcribed text and summaries.

Fig.2 shows the Functional block diagram that depicts the major functions of the system.

IV. Implementation

Key tools include machine learning and deep learning frameworks, along with specialized libraries.

- **Keras**: A high-level neural networks API that serves as an interface for TensorFlow, designed for quick prototyping.
- **TensorFlow**: A robust platform for machine learning, providing a wide array of tools and resources for developing and training models.
- **NumPy**: A crucial library for numerical computations in Python, especially for manipulating arrays and matrices.
- **Librosa**: A specialized library for analyzing audio and music, useful for tasks like loading audio files and extracting characteristics such as Mel Frequency Cepstral Coefficients (MFCCs).
- **Jiwer**: A library that assesses the performance of speech recognition systems, calculating metrics such as Word Error Rate (WER).
- **Matplotlib**: A comprehensive library for creating various types of visualizations, including static, animated, and interactive plots.
- **Scikit-Learn**: Provides a suite of tools for implementing and evaluating machine learning algorithms.
- **OpenCV**: An open-source library for video and image processing, including frame extraction, object detection, and keyframe identification.
- **NLTK and Spacy**: Libraries used for processing natural language, including tasks like tokenization, stemming, lemmatization, and part-of-speech tagging.
- **HuggingFace Transformers**: Offers a collection of pre-trained models like BERT and T5, useful for generating contextually relevant text summaries.
- **FFmpeg**: A powerful multimedia framework used for video and audio manipulation, such as extraction, conversion, and compression.
- **PyTorch**: A flexible deep learning framework known for its dynamic computation graphs, enabling more intuitive model building.
- **Pandas**: An essential library for data manipulation and analysis, providing data structures such as DataFrames.
- **SoundFile**: A library for handling sound files, supporting various formats and offering a simple interface for audio data.

A. Methodology

The methodology involves a comprehensive approach for processing video inputs, extracting audio, transcribing speech, summarizing text, identifying key frames, and enabling keyword-based searches. Our main model is the ASR model based on CNN and LSTM layers.

1) ASR Model: The implementation of ASR system, the most important module is structured as follows:

1) **Data Preparation**:
   - Audio files are pre-processed to extract relevant features such as MFCCs, which serve as input to the ASR model. The LJSpeech dataset, which contains 13,100 audio files in the form of wav files in the `/wavs/` folder, is used for developing the model. The dataset is split into training, validation, and test data in an 80:10:10 ratio. The label (transcript) for each audio file is a string provided in the `metadata.csv` file.
2) **Model Architecture**:
   - A deep learning model is designed using CNNs and bidirectional LSTMs, to capture both spatial and temporal dependencies in the audio data. A final dense layer with softmax function is used to predict the result.
3) **Connectionist Temporal Classification Loss Function**:
   - The CTC loss function is used to train the model, allowing it to learn the alignment between the input audio features and the target text sequences.
4) **Model Training**:
   - The model is trained on a dataset of audio files and their corresponding transcriptions.
5) **Inference**:
   - The trained model is used to transcribe new audio files, converting speech into text, and the performance of the model is measured using Word Error Rate (WER).

2) Video Input Processing: The algorithm used is given briefly.

Algorithm:

a. Load Video: Load video file using moviepy.
b. Extract Audio: Extract audio track and save it as a WAV file.
c. Resample Audio: Use librosa to resample the audio to 16kHz.

3) Summary Generator: Utilizes NLP libraries along with fine-tuned summarizers, which mainly include T5, that is used to generate coherent summaries in multiple formats. We uses chunks of maxlength as 512 dimensions in T5 to process lengthy texts. We also integrate Gemini to enhance our comprehension of the text and produce summaries. The algorithm used is given briefly.

Algorithm:

a. Preprocess Text: Clean and tokenize the transcribed text.
b. Summarize Text: Use the summarization model to generate summaries.
c. Text-to-Speech: Convert summarized text to speech using gTTS to support even audio summary.

4) Video Analyzer and Keyframe Extraction: The fine-tuned deep learning model (Caffe model) is used to extract features from each frame in the video. The model then processes each frame and generates a feature vector that represents the content of the frame. The extracted feature vectors from all frames are then scaled and clustered using the K-Means algorithm. The K-Means algorithm groups the feature vectors into a specified number of clusters, and identify keyframes that are closest to the cluster centers. The algorithm used is given briefly.

Algorithm:

a. Extract Frames: Capture frames from the video at regular intervals.
b. Feature Extraction: Use a pre-trained caffe neural network to extract features from frames.
c. Clustering: Apply KMeans clustering to identify keyframes.
d. Save Keyframes: Save the selected keyframes as image files.

5) Search and Personalization Module: This module implements NLP based semantic search techniques and keyword-based content retrieval. Cosine similarity is used as the similarity meausure. It also supports Questions and Answer feature integrated with Gemini, to understand the most relevant contents.The algorithm used is given briefly.

Algorithm:

a. Preprocess Text: Clean and tokenize text using NLP toolkit functions.
b. Extract Contexts: Identify contexts around keywords in the text.
c. Semantic Search: Use a sentence transformer model to find semantically similar sentences using cosine similarity measure.
d. Output Results: Return exact and semantic matches for the keywords.

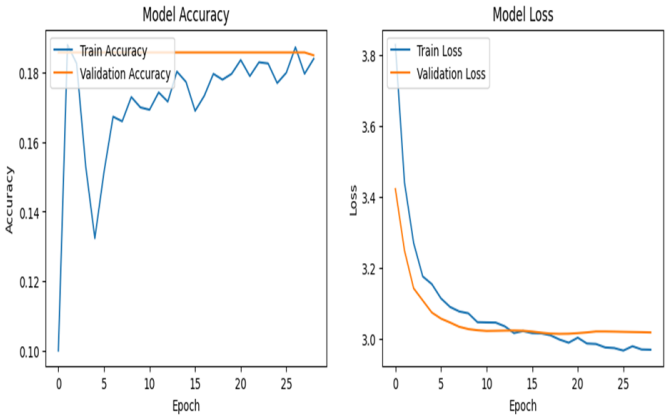

Fig. 3. Accuracy vs Epochs and Loss vs Epochs Plot for Own Dataset.

V. Results and Discussion

The performance of the proposed framework was evaluated and this section provide detailed results, performance analysis, and visualizations to highlight the model's effectiveness.

A. ASR Model

Initially, we tried to train ASR model with our own dataset which was generated by collecting videos meetings, extracting audios and the transcribed text. The dataset mainly consisted of 29 meeting files of various lengths. The loss curve highlights the need of data augmentation, because of underfitting, as the model failed to fetch better accuracy. We couldn't work on more data as it was computationally intensive, and it was difficult to collect and process a large volume of dataset. Fig.3 shows accuracy and loss curve with respect to initial training data.

The ASR model was thus fine-tuned and trained on LJSpeech dataset and the performance of test data was evaluated based on loss and WER.

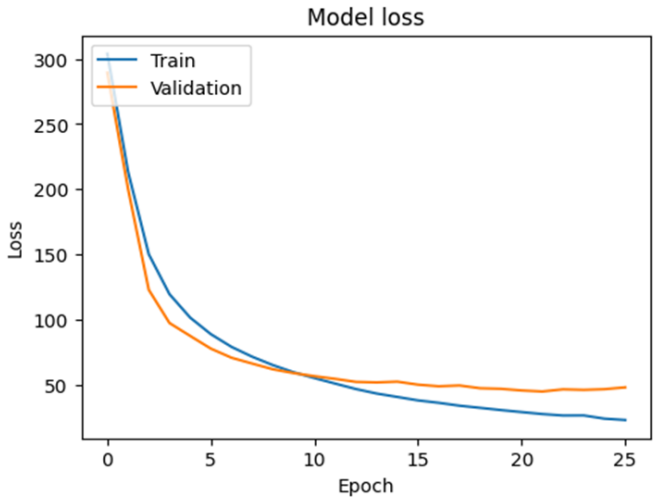

Fig. 4. Loss vs Epochs Plot for Enhanced ASR Model.

```
1/1 [==============================] - 0s 207ms/step
1/1 [==============================] - 0s 209ms/step
1/1 [==============================] - 0s 205ms/step
1/1 [==============================] - 3s 3s/step
--------------------------------------------------------------------------------------------------------
Test Word Error Rate (WER): 0.2651
Test Character Error Rate (CER): 0.0739
Test Match Error Rate (MER): 0.2600
--------------------------------------------------------------------------------------------------------
Target    : perhaps herodotus had never been in babylon
Prediction: perhaps he rodetus had never been in babylon
--------------------------------------------------------------------------------------------------------
Target    : they will be sent to an allotment division composed of representatives of the more important governmental agencies
Prediction: by will be sent to in alotment dovision composed of representatives of the more important governmental agencies
--------------------------------------------------------------------------------------------------------
Target    : the police were at once employed and after infinite pains they discovered the person one moss who had acted as the messenger
Prediction: the police were at once employed and after infinute paines they discovered the person one moss who had acted as the messener
--------------------------------------------------------------------------------------------------------
Target    : by the employment of instruments of precision for the measurement of the physical forces which are at work in the living economy
Prediction: by the employment of insuments of precition for the measurment of the phisital forces which harat ork in the living acconomy
--------------------------------------------------------------------------------------------------------
Target    : made no voluntary movements had his eyes open with the pupils dilated without any reaction to light
Prediction: made novolitary movements had his es opon with the puple's dilated without anuriaction to light
--------------------------------------------------------------------------------------------------------
Target    : the series of boldlyconceived and cleverlyexecuted forgeries in which james townshend saward commonly called jem the penman
Prediction: the series of boldly conceived and cleverly executed forgeries in which james toundsand saward commonly calld a jemd the penan
--------------------------------------------------------------------------------------------------------
Target    : will be spent for actually creating new work and not for building up expensive overhead organizations here in washington
Prediction: will be sped for actually creating new work and not for building up expensive over head orgarnizations here in washing
--------------------------------------------------------------------------------------------------------
Target    : they will be sent to an allotment division composed of representatives of the more important governmental agencies
Prediction: by will be sent to in alotment dovision composed of representatives of the more important governmental agencies
--------------------------------------------------------------------------------------------------------
```

Fig. 5. Predictions obtained using Enhanced Model along with WER, CER and MER values.

The training and validation loss curves with consistent decrease in both training and validation loss in Fig.4 signifies the effective learning progress of the enhanced model over epochs. The performance of the ASR model was evaluated using the Word Error Rate (WER), Character Error Rate (CER) and Matching Error Rate (MER). Fig.5 shows the predictions along with various error rates. The predicted text along with ground text is provided for the sample test data. The lower error rate indicate effective learning and better predictions and ensures the better performance of the model.

B. Text Summarization

We have considered some of the commonly used summarizers for a comparative study to choose the best summarizer for our proposed work. The various summarization techniques using BERT and other sentence transformers are also performed on the extracted text for a comparative study. The similarity measures were then applied to the various summarized text to analyse the effectiveness of summarization. The Fig.6 shows the comparative analysis of various summarization techniques and T5 mainly used by Google was found to be effective. Hence, we have chosen T5 for our work. We also integrated Gemini with T5 to provide better insights on summarized text.

Model	TF/IDF (Score)
Lex_Rank	0.157567719
T5	0.189336193
TextRank	0.158208032
NLTK	0.176043955
LSA	0.139538148

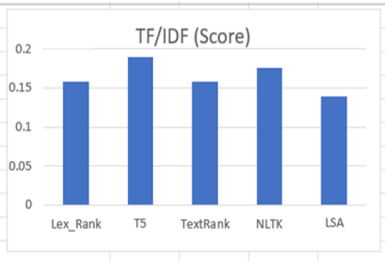

Fig. 6. Comparison of Summarization Techniques.

The Fig.7 shows the results obtained by using our summary generator.

C. Search and Personalization

The Fig.8 shows the results obtained after performing search techniques on the transcribed text. The search techniques provide those sentences which contain the keyword and the threshold is set as 5 to get at least 5 words in a sentence, The search for similar words will fetch results only for those words which have similar meaning as the keyword in the context. Even the question answering module based on Gemini fetched best results thereby enhancing the performance of our proposed work.

D. Keyframe Extraction

The successful keyframe extraction affirmed the efficacy of our proposed framework by retrieving optimal keyframes. This provides a visual impact on summarization using images. The Fig.9 shows the result obtained after performing keyframe extraction on the test video.

E. Overall Results

To assess performance in terms of computing time, multiple meeting-based movies of varying lengths (duration) were

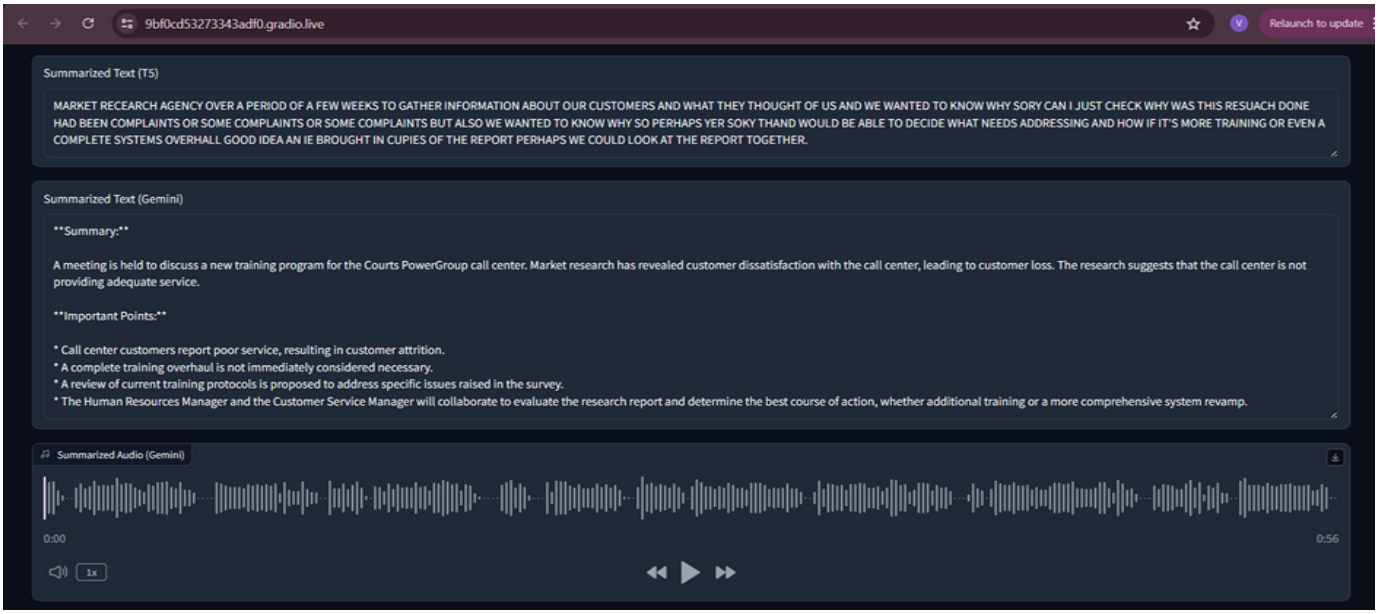

Fig. 7. Results of Summary Generator (Using T5 and Gemini Integrated).

utilized to extract transcripted text using ASR, followed by summarization. Fig.10 shows the video title, video duration, and the time taken for generating transcripted text, summarization, actual word count, and summarized word count in a tabular format.

Even though the computational time was found to vary based on machine configurations, it was observed that the time taken is directly proportional to the video length. The Fig.10

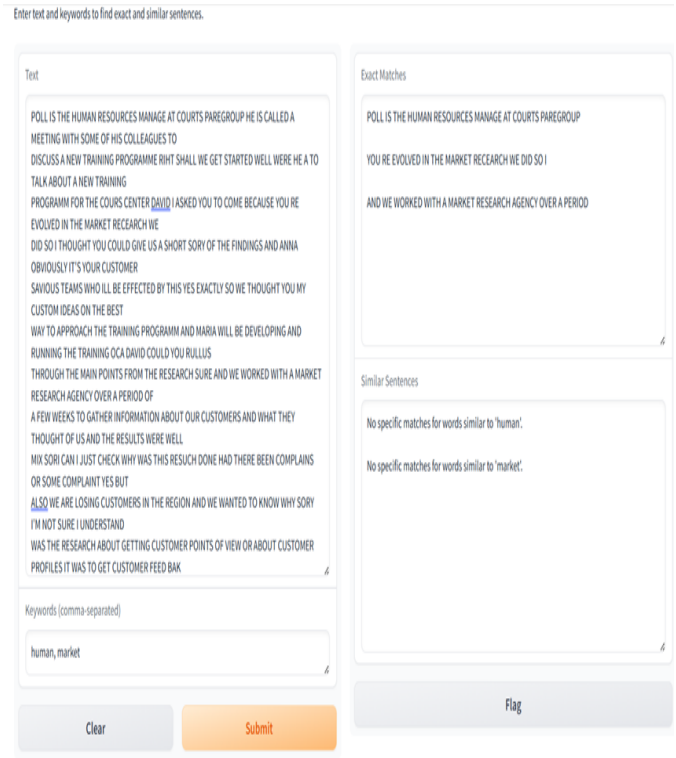

Fig. 8. Search Results based on Keyword and Semantic Search.

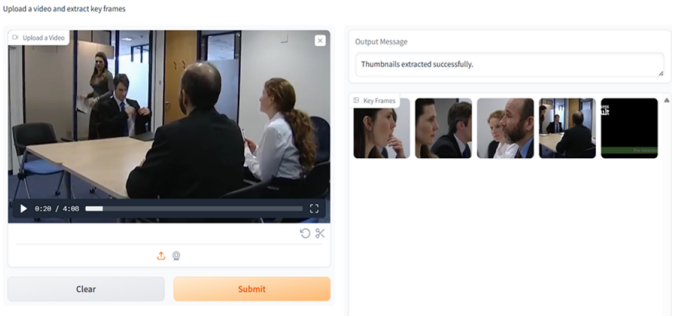

Fig. 9. Keyframes Extracted from a Sample Video.

Video Index	Video Title	Video Length (in seconds)	Audio Length (in samples)	Transcription Time (in seconds) *	Summary Time (in seconds) *	Transcription Word Count	Summary Word Count
0	Business English: Participating in meetings 2	248	3970880	31.84638453	25.83760834	569	99
1	Progress Meeting (Short Comedy Sketch)	174	2776000	14.52206278	15.92221141	342	44
2	Chairing a meeting	371	5929760	50.57343149	31.95389414	869	113
3	Business English B1 - B2: Participating in meetings 1	221	3529760	22.03859329	21.24465895	528	81
4	BUSINESS RESULT INTERMEDIATE UNIT 12 "A formal presentation"	386	6175040	54.4393971	31.00064445	779	117
5	Effective Meetings: Simulated Exercise for Chairing & Minute Taking	857	13706400	210.2041864	91.38141418	2733	383
6	Team meeting updates	268	4288800	30.60452175	28.08300972	631	100

*: depends on machine configurations. The platform we used was Google Colab with T4 GPU.

Fig. 10. Video Analysis Results in terms of Computational Time and Word Count

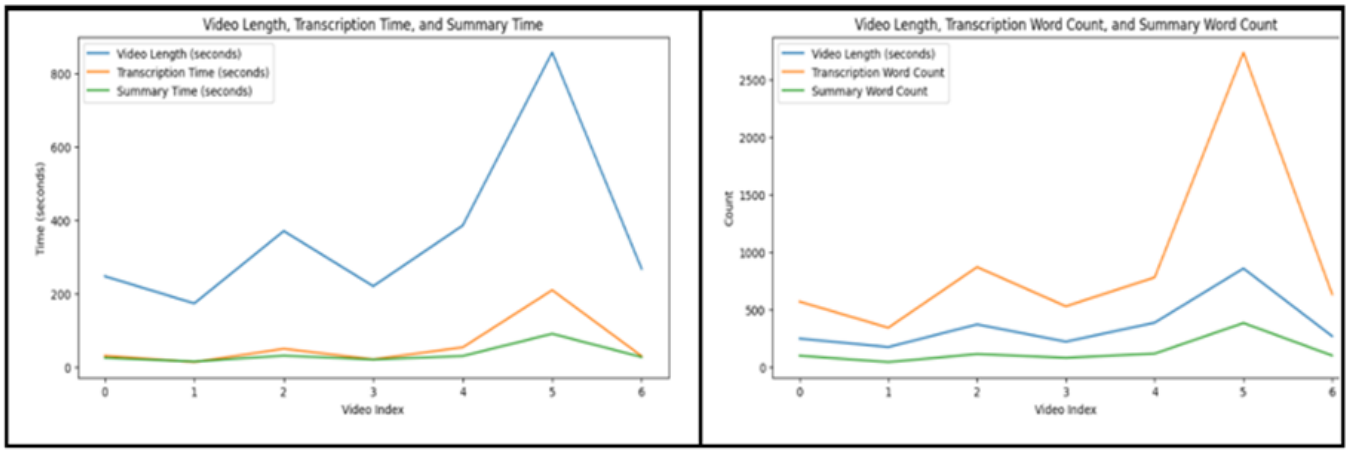

Fig. 11. Computational Time and Word Count Plot

clearly shows that the time taken is significantly less when compared to the actual video length. The plots given in Fig. 11 help us to have a better visual insight into the computational time and word counts.

The findings and discussion show that our suggested framework met all our objectives by displaying robust performance and ensuring high model efficacy. All our results and discussions are based on the GPU available in Google Colab, and the paid GPU was used for training and testing our model due to lack of necessary local hardware reources for development.

VI. Conclusion and Future Work

The proposed framework for an automated tool successfully helped us to develop a system for efficient video summarization and knowledge extraction. Much like ASR systems have evolved through the integration of advanced neural network architectures such as CNN and RNN to capture nuanced audio features, this project utilizes an effective ASR model based on a combination of CNN and LSTM layers to convert audio to transcribed text. The use of transformer-based models with an integration of Gemini helps in generating meaningful succinct summaries from the transcribed text. Computer vision algorithms integrated with K-means clustering helps in keyframe identification. Thus, the system offers robust functionalities such as keyframe identification, coherent summary generation, and advanced search capabilities leveraging advanced techniques in video analysis, NLP, and computer vision. It can tackle the challenges posed by the abundance of online video content by enhancing the accessibility and utility of video content.

Despite these advancements, challenges persist in the development of ASR systems and other areas like real-time processing, enhanced personalization, adaptability to new domains and languages, multimodal integration, and multilingual processing.

- Real-time Video Summarization: Enhance the system to perform real-time video summarization, allowing users to obtain summaries as videos are being streamed or recorded. This could involve optimizing algorithms for speed and efficiency without compromising on accuracy. This optimization may involve streaming optimization, parallel processing and incremental summarization. This advancement would not only improves user experience by providing timely insights but also supports applications in live broadcasts, surveillance, and real-time analysis scenarios.
- Enhanced Personalization: Implement more advanced user profiling and personalization features. This could include learning user preferences over time to tailor summaries based on individual interests or needs, thereby improving user engagement and satisfaction. This personalization could be achieved through advanced algorithms such as reinforcement learning or collaborative filtering, that learn from user interactions and adjust summaries accordingly. Integrating recommendation systems to suggest relevant videos or summaries based on user preferences and previous interactions might also help in enhancing personalization.
- Adaptability to New Domains and Languages: Extend the system's capabilities to handle a broader range of domains and languages. This could involve transfer learning techniques to adapt existing models to new domains or languages with less available training data. Adapting the system to new domains and languages expands its reach, making it accessible to users across diverse fields and linguistic backgrounds.
- Multi-lingual Processing and Language Switching: Implement models capable of processing and summarizing content in multiple languages and language switching involved in a single video. Adapting the system to handle multiple languages used in the same video involves developing robust multilingual models or employing translation mechanisms that preprocess content into a common language for analysis and summarization, thereby handling language-specific nuances and contexts. Recent advancements in multilingual models, such as those based

on transformers, can facilitate this adaptation by enabling models to learn representations that generalize across languages.

In essence, the future scope and further enhancement of the project involve leveraging advancements in AI, deep learning, and multimedia processing to continually refine and expand the capabilities of the intelligent system. By focusing on usability, personalization, and integration with emerging technologies, the project can evolve to meet evolving user needs and expectations in accessing and extracting knowledge from video content.

REFERENCES

[1] S. Santini, "Who needs video summarization anyway?," in *International Conference on Semantic Computing (ICSC 2007)*, pp. 177-184, Sept. 2007.

[2] R. M. Jiang, A. H. Sadka, and D. Crookes, "Advances in video summarization and skimming," in *Recent Advances in Multimedia Signal Processing and Communications*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 27-50, 2009.

[3] C. R. Huang, P. C. J. Chung, D. K. Yang, H. C. Chen, and G. J. Huang, "Maximum a posteriori probability estimation for online surveillance video synopsis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1417-1429, Aug. 2014.

[4] L. Zhang, L. Sun, W. Wang, and Y. Tian, "KaaS: A standard framework proposal on video skimming," *IEEE Internet Computing*, vol. 20, no. 4, pp. 54-59, July-Aug. 2016.

[5] D. Liu and W. Hu, "Skimming, locating, then perusing: A human-like framework for natural language video localization," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4536-4545, Oct. 2022.

[6] M. Fei, W. Jiang, and W. Mao, "Memorable and rich video summarization," *Journal of Visual Communication and Image Representation*, vol. 42, pp. 207-217, Jan. 2017.

[7] S. R. Badre and S. D. Thepade, "Summarization with key frame extraction using Thepade's sorted n-ary block truncation coding applied on Haar wavelet of video frame," in *Proceedings of the IEEE Conference on Advances in Signal Processing (CASP)*, Pune, India, pp. 332-336, June 2016.

[8] M. Sajjad, M. Mehmood, S. Rho, and S. W. Baik, "Divide-and-conquer based summarization framework for extracting affective video content," *Neurocomputing*, vol. 174, pp. 393-403, Jan. 2016.

[9] J. Wu, S. H. Zhong, and J. Jiang, "A novel clustering method for static video summarization," *Multimedia Tools and Applications*, vol. 76, no. 7, pp. 9625-9641, Apr. 2017.

[10] H. Gharbi, S. Bahroun, M. Massaoudi, and E. Zagrouba, "Key frames extraction using graph modularity clustering for efficient video summarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, pp. 1502-1506, Mar. 2017.

[11] S. H. Emon, A. H. M. Annur, A. H. Xian, K. M. Sultana, and S. M. Shahriar, "Automatic video summarization from cricket videos using deep learning," in *Proceedings of the 2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1-6, Dec. 2020.

[12] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, ... and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, Nov. 2012.

[14] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645-6649, May 2013.

[15] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960-4964, Mar. 2016.

[16] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, ... and G. Zweig, "Achieving Human Parity in Conversational Speech Recognition," Microsoft Research Technical Report, 2016.

[17] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, ... and Z. Zhu, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 173-182, June 2016.

[18] J. Li, L. Deng, and Y. Gong, "An Overview of Noise-Robust Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745-777, Apr. 2014.

[19] G. Pundak and T. N. Sainath, "Lower Frame Rate Neural Network Acoustic Models," in *Proceedings of Interspeech*, 2018.

[20] J. Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, 2022.

[21] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, 2019.

[22] L. Sarı and M. Hasegawa-Johnson, "Counterfactually fair automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2021.

[23] S. Feng, O. Kudina, and B. M. Halpern, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[24] N. Giarelis, C. Mastrokostas, and N. Karacapilidis, "Abstractive vs. extractive summarization: An experimental review," *Applied Sciences*, vol. 13, no. 13, p. 7620, 2023.

[25] A. K. R. Maya, J. Nazura, and B. L. Muralidhara, "Recent trends in answer script evaluation–a literature survey," in *3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021)*, Atlantis Press, 2021, pp. 105-112.

[26] T. Vu and A. Moschitti, "AVA: an automatic evaluation approach to question answering systems," *arXiv preprint arXiv:2005.00705*, 2020.

[27] D. Diefenbach, V. Lopez, K. Singh, and P. Maret, "Core techniques of question answering systems over knowledge bases: a survey," *Knowledge and Information Systems*, vol. 55, pp. 529-569, 2018.