Survey on Video Summarization using Extracted Audio

Tintu Alphonsa Thomas
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
tintualphonsathomas@amaljyothi.ac.in

Nandana Rajagopal
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
nandanarajagopal20@gmail.com

Neethu Liz Shaji
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
neethulizshaji@gmail.com

Silby Elza Simon
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
silbyelzasimon2000@gmail.com

Sree Parvathy P
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
parvathypresanthan@gmail.com

***Abstract*— Summarization of videos using extracted audio is an area of natural language processing that aims to summarize videos like lectures, podcasts and meetings that cannot be effectively summarized using just video frames. The task can be divided into four steps: audio extraction, speech-to-text, text preprocessing, and text summarization. This paper discusses about research in these four areas.**

Keywords*— *PEGASUS, FCN, ASR

I. INTRODUCTION

Video Summarization aims to create a brief narrative that highlights the most significant and educational elements of a video. By choosing and presenting the most educational or fascinating content for potential users, video summarising creates a concise summary of the information contained in a larger video document.

There are several reasons why one might want to summarize videos:

• Time-saving: Summarizing a video can save time, especially when the video is long and contains information that is not relevant to the viewer's needs.

• Accessibility: By providing a summary of the video's key points, people who may have hearing or visual impairments can still access the content without having to rely solely on the audio or visual components.

• Learning: By summarizing the key points, viewers can better understand and remember the content.

• Sharing: By providing a concise summary, viewers can quickly share the video's key points with others, making it easier to spread the message or information.

Overall, summarizing videos can help improve accessibility, save time, reinforce learning, and make it easier to share important information with others.

DOI:10.5281/zenodo.8019200

II. LITERATURE SURVEY

A. Audio Extraction

Anil Kale et al. [15] presented an automated audio extraction technique for video retrieval. It is built using a client-server architecture. The specific time the word appears in the video is also obtained via server-side APIs, which are responsible for scanning uploaded video for embedded audio and converting it to text. When the audio is converted to text, the server stores the text as a list of significant words and the relevant time instance in a database. In this manner, each time a user types in a search term, the search query is submitted to the text database, the appropriate time instance of the word's occurrence is obtained, and the chosen movie is played from that point in time.

N. Radha [14] proposed a way for retrieving videos utilising audio and text from the video. It has an audio track and graphic display. To extract each unique slide frame with its temporal transition taken into consideration as the video segment first, identify the slide transition from the visual screen. The textual metadata is first extracted from the slide frames using the video OCR method, and then it is recognised. Based on the OCR results, the matching text in the video data is saved. Second, the speech-to-text analysis is performed using the audio signal. Sphinx speech recognition models are used to identify the voice-to-text conversion process.Using the combined text information from the audio signal and visual slide frame in the proposed work, the video was obtained from the lecture video database. The integrated video retrieval system performs noticeably better than a standalone system developed using audio and text in video systems, respectively.

B. Speech-to-Text

Rohit Ranchal et al. [11] suggested a method for real-time captioning and lecture transcription in the classroom using speech recognition. A variety of classroom settings were used to test speech recognition (SR) technologies to assist students in automatically transcribing oral lectures into text. Two distinct SR-mediated lecture acquisition (SRmLA) strategies, real-time captioning (RTC) and postlecture transcription (PLT), were evaluated in live life and social sciences lecture courses employing conventional classroom technology. Both strategies were compared with regard to their technical viability and dependability of implementation in the classroom, instructors' backgrounds, the accuracy of word recognition, and student class performance. RTC provided students with almost immediate access to the instructor's comments during class. PLT employed a userindependent SR approach to create multimedia class notes with synchronised lecture transcripts, instructor audio, and class PowerPoint slides that students could access online after class. PLT produced better word recognition accuracy than RTC.In a science course, PLT users did higher on optional online assessments and were more inclined to take them in general. The general quality of the lesson increased when multimedia lecture notes were made available. The benefits of SR-mLA for students who have trouble taking accurate and independent notes were emphasized, particularly for non-native English speakers and students with disabilities. Field-tested best practices for maximising SR accuracy were offered for both SR-mLA techniques.

Burhanuddin Lakdawala et al. [10] used CMU Sphinx to suggest a strategy for audio-to-text transcription. This project's transcribing system for healthcare institutions can be used by counselors and non-governmental groups to record survey talks, turn them into text, and then save them. This system includes an open-source application. The CMUSphinx toolkit is used for voice recognition. Multiple languages can be recognised by the system. The CMUSphinx toolbox uses the acoustic model, phonetic dictionary, and language model. The CMUSphinx toolbox does speech recognition and transcription when the user records their voice using a mobile application. The transcription file will be saved as a text file in the device's memory, where the user can access it using the program to download data from the database server.

E.Saranyal et al. [5] presented a speech-to-text user assistive agent for an impaired person. The recommended methodology is adaptable, allowing spoken communication to be mastered through a strong discourse recognition process that takes into account loud environmental factors unrelated to the human body. The proposed framework consists of a mobile device with an acoustic bar prior to directed noise reduction, suited for carrying out discourse to content interpretation tasks, and capable of recognising catchphrase spotting system.

Sadaoki Furui et al. [7] proposed a method Speech-toText and Speech-to-Speech summarization of spontaneous speech. Automatic speech-to-text and speech-to-speech summarization are built on the principles of speech unit extraction and concatenation. Extraction is done based on linguistic score, significance score, and confidence score. Sentence compression is the next phase. To compute the score for sentence compression, the transcription that remains is automatically changed into an editorial article style after substantially fewer significant sentences have been removed.

Giovanni Dimauro et al. [9] proposed a method for the assessment of speech intelligibility in Parkinson's disease using a Speech-to-Text system. Systems for text-to-speech translation (STT) enable computers and other electronic devices to recognise spoken languages. Google Cloud Speech API is used by it. Voxester is a software solution that is effective, convenient to use, and reasonably priced for the assessment of digital voice and speech changes in Parkinson's disease.

C. Text Preprocessing

E.Elakiya et al. [13] designed a preprocessing framework for text mining applications. By gathering the most popular preprocessing methods, this study seeks to develop a preprocessing framework for text-mining applications. This method focuses on the three preprocessing steps of expansion, removal, and tokenization (ERT). The ERT creates a list of tokens that can execute all learning algorithms using a corpus as input. Each preprocessing step can be finished more quickly and easily with the help of the ERT framework.

Dino Isa et al. [6] proposed a text document preprocessing with the Bayes formula for classification utilizing the Support Vector Machine. By combining the Bayes vectorizer and SVM classifier, the authors suggest, design, implement, and test a hybrid classification approach that makes use of the simplicity and accuracy of the Bayes technique. Our proposed naive Bayes classifier studies the text document by extracting words that are present in it in order to carry out its classification tasks. This is achieved by vectorizing raw text data using the naive Bayes technique at the front end and classifying the articles using an SVM classifier at the back end.

Andrew Kurbatow et al. [8] proposed the research of text processing effect on text documents classification efficiency methods. The classified texts must be preprocessed in accordance with a variety of requirements. The significance of these demands has been looked at in the context of this work. It is crucial to consider the requirements placed on the outcomes of text document preprocessing, notably the reduction of words to their most basic forms, in order to execute an acceptable classification of text documents. The efficiency requirements were not increased by eliminating stop words; on the contrary, they were little affected. It might be caused by the unique characteristics of the training sample. It is critical to remember that the classification result is strongly influenced by the effectiveness of the training sample.

D. Summarization

A greedy extractive summarization method was proposed by Akhmetov et al. [1] and is used to summarise scientific literature from the arXive and PubMed datasets. The approach focuses on selecting the sentences with the most words having the highest TFIDF (Term Frequency - Inverse Document Frequency) values for the summary and tweaking the minimum document frequency parameter for TFIDF vectorization. The approach yields ROUGE1/ROUGE-2 scores of 0.40/0.13 and 0.43/0.12 using the arXive and PubMed datasets, respectively. These outcomes are comparable to those of state-of-the-art models that employ complex neural network topologies, strong processing capabilities, and enormous training data sets. This strategy uses a straightforward statistical inference methodology.

In their study of an asynchronous MMS (Multi-modal summarization) task, Haoran Li et al. [2] describe how to compile relevant text, audio, and video data into a textual summary. The MMS work is conceptualised as a budgeted submodular function maximisation issue. By judiciously employing audio transcription through guide methodologies, readability is addressed. A unique graphbased methodology is developed to more correctly determine the salience score for each text unit and provide readable and informative summaries. When examining various approaches to establish the relationship between the image and text the image topic model and the image match model are equally effective for the MMS task.

Pre-training Transformers on enormous text corpora with self-supervised objectives has shown great results when tuned for later NLP tasks like text summarization, claim Jingqing Zhang et al [3]. In this study, it was recommended that sizeable Transformer-based encoder-decoder models be pre-trained on sizable text corpora utilising a new selfsupervised goal. In PEGASUS, important sentences from an input document are eliminated or otherwise concealed, and the remaining phrases are then generated as one output sequence, much like an extractive summary. The PEGASUS model exceeds prior state-of-the-art accomplishments in low-resource summarization in six datasets with only 1000 samples. They also confirmed their findings using human assessment, proving that their model summaries outperform human summaries across a range of datasets.

Yen-Chun Chen and Mohit Bansal [4] proposed an accurate and speedy summarization method that first selects significant lines and then rewrites them abstractly (i.e., compresses and paraphrases) to deliver an overall summary in a brief manner. They took their inspiration from the way long articles are summarised by people. Empirically, they achieve the new state-of-the-art on all criteria (including human evaluation) and much higher scores for abstractness using the CNN/Daily Mail dataset. They additionally demonstrate the generalizability of their model on the test-only DUC2002 dataset, where it outperforms a state-of-the-art model.

Kenny Davila et al. [12] developed a novel technique employing FCN (Fully Convolutional Network) for extracting and summarising whiteboard and chalkboard lecture videos. FCN serves to take lecture video frames and remove the scribbled text from them. After that, stable handwritten material is identified to generate a spatial-temporal index. The estimated timing of the content is then utilised to build the cumulative deleted-content signal, and its peaks are used to produce temporal video segments. The lecture video summary's key-frames are then created, one for each temporal portion, utilising the spatial-temporal index.

J.N. Madhuri et al. [16] advocated employing phrase ranking as a technique for extractive text summarization. It is

demonstrated how to extract text from a single document and summarise it using a ground-breaking statistical technique. Sentence extraction is a technique that condenses the main concept of the provided text. The ranking of sentences is based on the weights that are given to them. Highly scored sentences from the input document are taken in order to build a highquality summary of the input document and store the summary as audio

References

[1] I. Akhmetov, A. Gelbukh and R. Mussabayev, "Greedy Optimization Method for Extractive Summarization of Scientific Articles," in IEEE Access, vol. 9, pp. 168141-168153, 2021, doi: 10.1109/AC CESS.2021.3136302.

[2] H. Li, J. Zhu, C. Ma, J. Zhang and C. Zong, "Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 5, pp. 996-1009, 1 May 2019, doi: 10.1109/TKDE.2018.2848260.

[3] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. "PEGASUS: pre-training with extracted gap-sentences for abstractive summarization". In Proceedings of the 37th International Conference on Machine Learning (ICML'20). JMLR.org, Article 1051, 11328–11339

[4] Chen, Yen-Chun Bansal, Mohit. (2018). "Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting". 675-686. 10.18653/v1/P18-1063.

[5] E. Saranya, B. B. Sam and R. Sethuraman, "Speech to text user assistive agent system for impaired person," 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Chennai, India, 2017, pp. 221-226, doi: 10.1109/ICSTM.2017.8089155

[6] D. Isa, L. H. Lee, V. P. Kallimani and R. RajKumar, "Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine," in IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 9, pp. 1264-1272, Sept. 2008, doi: 10.1109/TKDE.2008.76.

[7] S. Furui, T. Kikuchi, Y. Shinnaka and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," in IEEE Transactions on Speech and Audio Processing, vol. 12, no. 4, pp. 401-408, July 2004, doi: 10.1109/TSA.2004.828699.

[8] A. Kurbatow, "The research of text preprocessing effect on text documents classification efficiency," 2015 International Conference "Stability and Control Processes"in Memory of V.I. Zubov (SCP), St. Petersburg, Russia, 2015, pp. 653-655, doi:10.1109/SCP.2015.7342234.

[9] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano and F. Girardi, "Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System," in IEEE Access, vol. 5, pp. 22199-22208, 2017, doi: 10.1109/ACCESS.2017.2762475.

[10] B. Lakdawala, F. Khan, A. Khan, Y. Tomar, R. Gupta and A. Shaikh, "Voice to Text transcription using CMU Sphinx A mobile application for healthcare organization," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 749-753, doi: 10.1109/ICI CCT.2018.8473305.

[11] R. Ranchal et al., "Using speech recognition for real-time captioning and lecture transcription in the classroom," in IEEE Transactions on Learning Technologies, vol. 6, no. 4,pp. 299-311, Oct.-Dec. 2013, doi: 10.1109/TLT.2013.21.

[12] K. Davila, F. Xu, S. Setlur and V. Govindaraju,"FCN LectureNet: Extractive Summarization of Whiteboard and Chalkboard Lecture Videos," in IEEE Access, vol. 9, pp.104469-104484, 2021, doi: 10.1109/ACCESS.2021.3099427.

[13] E. Elakiya and N. Rajkumar, "Designing preprocessing framework (ERT) for text mining application," 2017 International Conference on IoT and Application (ICIOT), Nagapattinam, India, 2017, pp. 1-8, doi: 10.1109/ICIOTA.2017.8073613.

[14] N. Radha, "Video retrieval using speech and text in video," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2016, pp. 1-6, doi:10.1109/INVENTIVE.2016.7824801.

[15] A. Kale and D. G. Wakde, "Video Retrieval Using Automatically Extracted Audio," 2013 International Conference on Cloud Ubiquitous Computing Emerging Technologies, Pune, India, 2013, pp. 133-136, doi: 10.1109/CUBE.2013.32.

[16] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication(IconDSC), Bangalore, India, 2019, pp. 1-3, doi:10.1109/IconDSC.2019.8817040R.