Video Summarization using Subtitles

Anindita Sarkar
Department of CSE
SRM Institute of Science and Technology
Chennai, India

S. Megha
Department of CSE
SRM Institute of Science and Technology
Chennai, India

Puppala Srilekha
Department of CSE
SRM Institute of Science and Technology
Chennai, India

S. Niveditha
Asst. Professor, Department of CSE
SRM Institute of Science and Technology
Chennai, India.

***Abstract*— the main objective of this paper is to propose an improvised method to summarize videos in a way such that important and concise information is presented to end users. In our system we aim to present important aspects of a video through summarization of subtitles in a video with the help of text summarization and video mapping algorithms. The summarized video will be played with an audio generated version of the summarized subtitle file. The end result of the system would be a summarized video backed with a summarized audio generated version.**

Keywords—text summarization, video summarization, Latent Semantic Analysis

I. INTRODUCTION

Our time has come to see the fundamental significance of advanced innovation in our day by day lives. It allows us to open a colossal assortment of data. We have an infinite number of media content ranging from pictures to large-scale videos. A large amount of videos are transferred to YouTube, Dailymotion, Vimeo and other video-sharing sites each moment [2]. Due to these large volumes of digital data, it takes a significant amount of manpower and resources to retrieve and process important and relevant information. Various sources of digital media such as documentaries, sports matches and educational videos can be found on the internet. Processing of such large scale media can become tedious, time consuming and also be very heavy on hardware. Therefore summarization strategies are incredibly expected to consume the ever-developing measure of information accessible on the web. In essence, summarization is intended to assist us with consuming important data faster.

Video summarization is a system whose objective is to produce a short summary of a video, which can either be a succession of fixed pictures (key frames) or moving pictures (video skims) [7]. This will later help in better management of memory as a shorter version of the original video with all the important data is stored while taking space much less than the original video. This also helps in browsing, video indexing and retrieval. There are two ways of summarizing videos**:** Key Frame Based Video Summarization and Video Skimming based Video Summarization [3].

Key frames are still pictures chose to outline the video content in a fast and extremely smaller manner, with the goal that clients can get a handle on the general video content more rapidly than by watching a lot of video sequences. [8].

In Video Skimming based video summarization, the first video is fragmented into different parts which is a video cut with shorter span. Each section is joined by either a cut or a slow impact [3]. An example for video skimming is the trailer of movie.

Video summarization is a topic that has a lot of research potential and numerous works have been produced on the same. We would like present a system which focuses on a method of summarization where in the end result is a video which retains important and relevant information for everyone's benefit. This system uses text summarization as the primary method of summarizing videos. Automatic text summarization is the problem in the field of data science of creating a short and accurate summary from a drawn out report. Automatic text summarization can be utilized in an assortment of uses.

This paper focuses on picking out frames by summarizing an auto generated subtitle text file and mapping the summarized text with the frames in the video. The final result of the video is a generated audio playing behind the summarized video.

The next section of the paper discusses the related work that was carried out for the system. The third section describes the proposed system in detail while the IV section describes the stages involved in the system. The V section carries experimental details and results while section VI concludes the paper.

II. RELATED WORK

A number of systems for summarizing videos have been into existence. Srinivas M, Pai M and Pai R [2] proposed effective video summaries based on various aspects of a video such as brightness, contrast, edge distribution, hue count and so on. Key frame extraction was divided into 3 stages: The scores for each frame is computed in the first stage and the second stage involves the selection of key frames. The elimination of duplicate frames is performed in the third stage. The limitation of this system is that it does not take into account the audio aspect of the video. A system developed by Pan G et al [1] uses clustering to over segment frames into clips based on similarity and proximity. The preprocessing of this system however needs improvement. Another system by Kansagara R, Thakore D and Josh M [3] uses various techniques to summarize videos such as skimming and selection of key frames. The paper describes various key frame extraction algorithms to summarize videos such as: Clustering by Euclidean distance, Motion Focusing and so on. But this results in production of ambiguous summarized videos. In a view to overcome these shortcomings we propose a novel system that summarizes the video as well as generates a summarized audio for a holistic and informative experience.

III. PROPOSED SYSTEM ARCHITECTURE

Depending on the use case and type of documents, text summarization systems can fall into different categories. Abstractive and Extractive; Indicative and Informative; Single Document and Multiple Document Summarization.

Single Document summarization makes for less redundancy whereas Multiple Document summarization makes for more redundancy and irrelevant information.

Extractive Summarization involves excerpts from the text whereas Abstractive Summarization involves writing one's perspective or main points of a topic in his/her own words. The technological advancements on abstractive text summarization is not up to the mark and therefore extractive

summarization is preferred. Indicative summaries are the kind of summaries that require skimming of a document thereby revealing the main highlights of an article or a document. Informative Summaries require a deeper understanding of an article and therefore require scanning an article or document deeply.

Automated Text Summarization provides important and relevant points from a text file or a document. This is one of the major reasons of incorporating text summarization in our system as this makes the process of text summarization more efficient. For our system we require a text summarization algorithm that is Extractive and Informative. And therefore Latent Semantic Analysis (LSA) fulfills the requirements of the system proposed.

The workflow of the proposed system is as shown in figure (1). A video of any size and genre is taken as input to the system. Since this system is suited for videos containing the subtitles alone, it does not produce results for videos without the subtitle file. After checking for the subtitle file, the file is converted to a text file. Preprocessing involves cleaning of the subtitle file and removing redundant and irrelevant data from the text file.

After preprocessing of the generated text file, the text summarization takes place on the text. Text summarization is a Natural Language Processing (NLP) concept and goes through the process of tokenizing, stemming and removal of stop words. Stop words are words which do not add meaning to the text. Examples of stop words are: an, the, and, or and so on. The text summarization algorithm used for summarizing is Latent Semantic Analysis (LSA). After summarizing the text, a text to speech module is incorporated. The next stage of the system involves extracting tokens from the summarized text that can be mapped with the frames in the video. From this process various key frames are extracted which when appended give us a summarized video. Since the frames have been appended, the audio of the summarized video will not be very coherent. Therefore the generated audio will play so that the video and the audio together make for a complete informative experience for the end user.

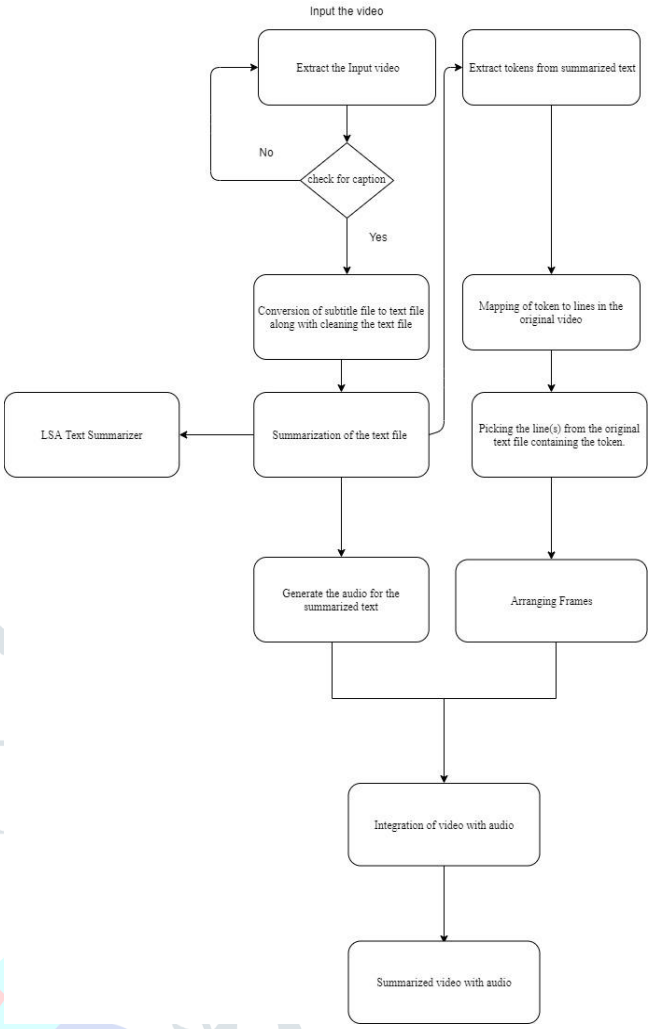

Fig. 1. Flowchart depicting the workflow of the video summarization system

IV. WORK FLOW OF THE SYSTEM

The objective of summarizing videos is to deliver a short synopsis video that passes on the significant and pertinent substance of a given longer video [4]. We propose a system that takes subtitles of the video as an important factor in summarizing videos. A URL of the video to be summarized is given as an input to the portal of the system. If the URL provided by the user is incorrect then the system shows an alert requesting the user to provide a valid URL. The summarizer goes through three stages in order to provide a summarized video with generated audio playing with the video.

A. Subtitle Text Summarization

The system begins by downloading the video as well as the subtitle file from the link given by the user. The subtitle file is converted into text file (.txt) so that summarization algorithms can be applied. There are various generic summarization algorithms in existence such as LexRank, Latent Semantic Analysis(LSA), Maximal Marginal Relevance(MMR), Support Sets to name a few.

Latent Semantic Analysis (LSA) was found to give a better performance for films and documentaries [6]. Latent-Semantic-Analysis (LSA) is a hypothesis and technique for extricating and speaking to the relevant utilization significance of words by measurable calculations applied to a huge corpus of content [5]. LSA induces relevant utilization of content dependent on word frequency. It is a verifiable model of word use that licenses assessments of the semantic closeness between bits of information. It is appropriate for applications for scientists in brain research and instruction

who must evaluate gaining from literary material. By playing out a programmed investigation of the writings that were perused by subjects, the determined semantic space can be utilized for coordinating among pieces of textual data much similarly as a propositional examination. The presence of the exact sentences becomes very important and crucial in these kind of use cases. Significant topics are resolved without the requirement for outer lexical assets: each word's event setting gives data concerning its importance, creating relations among words and sentences that connect with the manner in which people make affiliations [6]. LSA first produces a Document-Term Matrix (DTM) of events of each word in each record (sentences or sections). This DTM is made by giving values to every word in the sentence, since the occurrence of the stop-words is frequent they are ignored. LSA then uses singular value decomposition (SVD), a strategy firmly identified with eigenvector decomposition and the factor examination. We utilize the Singular-Value Decomposition (SVD) to decrease the position of the content network. SVD factors a Y × |S| content matrix say Z and a 3 x 3 grid [9]:

- A |S| × |S| matrix framework with V singular-vectors of A
- A [Y × |S|] corner to corner network E having the solitary qualities in plummeting request along its slanting.

In the event that we hold just the biggest k solitary qualities in the diagonal of the matrix, we get the kth rank framework Ak, which is the best estimate of the first matrix A. The SVD scaling divides the word-by-archive network into a lot of k, ordinarily 100 to 300, symmetrical components from which the first lattice can be approximated by straight blend. Every k value is represented as a continuous sentence instead of unique independent words. The importance of the term is calculated by the formula:

$$\frac{Term\ Frequency}{Document\ Frequency}$$

This algorithm also provides the quality of the summarized essay, the quality was examined and maintained using two approaches. The first analyzed the measure of semantic cover of the expositions with the original text. Each sentence in each summarized text is compared with the original text, and a score is allotted on the premise of the cosine between the summarized sentence and the nearest sentence in the original text. In this manner, if the summarized text is composed with a sentence is actually equivalent to a sentence in the original content, they would get a cosine of 1.0, while if a sentence that had no semantic cover with anything in the original text would get a cosine of 0.0. The subsequent measure is that the mean of cosines of the summarized text is compared to the original text. This also focuses on the coherence of the summarized text to the original text.
In this way an accurate and a clean text summary id obtained by using this LSA algorithm.

B. Video Mapping Algorithm

This module of the system deals with converting the large original video into smaller key frames which are to be appended to form a summarized video. So the problem statement is how should the key frames be determined and considered so that a complete summarized video is obtained. This problem is overcome by using the video mapping algorithm. This algorithm helps in mapping the generated text summary to the original video. On the one hand we have the generated summarized subtitle file and on the other we have the original video. After the summary of the subtitle file is generated, the time range of the various summarized sentences is calculated. Time range refers to the number of seconds each sentence takes in the video. Then the start and the end segment of this time range is calculated separately for each sentence. The start and end segments refer to the starting time and ending time of a sentence in seconds. Using this start and end segment time the system then obtains the clips corresponding to the time range respectively resulting in having multiple small chunks of videos in accordance with the summarized subtitle text file. The final video is generated by appending these key frames which were mapped with the summarized subtitle text sentences. Therefore the output of the video mapping algorithm provides us with a summarized and shortened video of the large original video.
This algorithm is optimized in such a way that there is maximum amount of time saved allowing the video to convey the whole information in a nutshell.

C. Audio Generation

Once the video is summarized, a corresponding audio reading the summarized text file is played behind the video summary. Speech-synthesis is the artificial production of human discourse. A framework utilized for this intention is called a speech synthesizer. A Text to speech (TTS) system changes conventional language content into Speech.
This audio is generated to enable a complete understanding of the information presented. This is done by a python dependency called GTTS (Google Text-to-Speech). It is an adjustable speech explicit sentence tokenizer that considers boundless lengths of content to be perused, all while keeping legitimate sound, shortened forms, decimals and more [10]. The audio along with the video facilitates a complete understanding of the video in terms of the information presented. This ensures that the end result of the system is completely unambiguous. Since the audio is generated from the summarized subtitle text, it can be used separately for other purposes as well.

D. Integration

This is the final module which comes into picture after all the above mentioned modules have been processed individually. The complete combination of every module together forms the whole system. In this the extraction of the video and the subtitle file, conversion of the subtitle file to text, the manipulation of the text to form an effective summary, the video mapping, the voice generation is together aggregated to form a single system. After the preprocessing of the voice and text the mapped video and the generated audio is integrated and synced so as to obtain a neat and clean video and audio together as a complete output.
The semantic video record is extracted by dissecting the caption document of any video without compromising on the quality and content of the same.

V. RESULT AND DISCUSSION

The proposed system was tested on a different videos of varying sizes and genres. For the purpose of calculating the accuracy of the system, five different videos of varying durations were picked from the popular video sharing platform YouTube and were made to run on the proposed system. The experimental results of the original video and the summarized video are calculated.

TABLE I.

S.No	*Original video duration (in min)*	*Summarized video duration (in min)*	*Time saved (in min)*
1	17.40	3.50	13.90
2	12.33	2.52	9.81
3	25	5.10	19.90
4	28.30	6.20	22.1
5	13.40	3.10	10.30

It is seen that there is a significant change in the duration of the original video and the summarized video. On an average the system saves 75-80% of the user's time. This system not only saves the time of the viewer but also gives the user relevant content in a concise manner.

Another set of results have been gathered for this system through an experiment as follows. In this experiment a documentary on space and research was taken for testing. Initially the system downloads the given video and the subtitle file of the corresponding video. The subtitle file is then converted into a text file which provides ease in manipulation and summarization of the text file. In this particular video we notice that the initial subtitle file of the original video file consists of about 118 sentences and was later summarized to 23 sentences by the system which is evidently about 20% of the original content. The summarization is done using the Latent Semantic Algorithm (LSA), a type of extractive summarization which takes the key sentences into account.

After the text is summarized the video mapping is done. The video mapping is done by choosing video of the corresponding summarized sentence of the subtitle file.

We now obtain small clips of various sentences present in the summarized text (as seen in the image). Later these small bits of video frames containing various key segments of the original videos are finally clipped together to from a full summarized video. Various key frames that have been extracted can be played

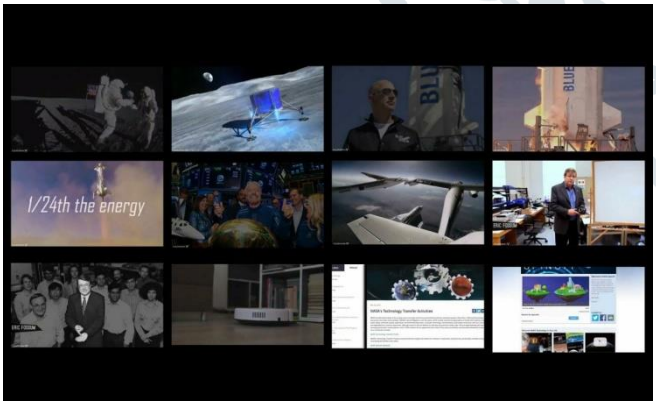

Fig. 2. Various frames selected from a documentary

The video is clipped in such a way that there is no lag or delay between two bits and is made sure that the video looks continuous. This Summarized video is then given a voice over using the GTTS algorithm of the python. The text which was earlier summarized using the LSA algorithm is then read along the corresponding video frames. The video frames and the speech is exactly synced because only the corresponding video frames of the summarized text is chopped from the original video. Finally a complete video of about 3 minutes is generated which was initially of 17 minutes and 40 sec. Therefore we can strongly conclude that this system saves almost 80% of the user's time and gives a better understanding of the long documentaries.

VI. Conclusion and Future Scope

Video summarization is an indispensable piece of numerous video applications, including video ordering, perusing and video retrieval. Briefly and astutely produced video edited compositions will encourage client access to huge volumes of video content in a compelling and effective way. Video summarization has been an important research topic and various systems and algorithms have been presented on the same. In this paper, we have presented a better system of summarizing videos where audio and video are taken into consideration. At present, the methodologies available for summarizing videos do not take the audio of the videos into consideration. With our system, a complete understanding of important information in a long video can be obtained. Videos with different durations can be summarized with better efficiency. Major constraint of the system is the type of video where videos such as short-films, movies or CCTV footages were not considered for summarizing in this system. Such videos require abstractive summarization whereas the proposed system is makes use of extractive summarization. Therefore this system is best suited for documentaries and public speeches. The summarized videos can be used for storing important information which hard to obtain for a rather lengthy video. Storing of shorter video helps in better memory management as many summarized videos containing relevant data can be stored.

Although this system has shown a different method on summarizing videos, there are various features that can be upgraded as an extension of the system. This system can be upgraded by including abstractive summaries which gives the capacity of summarizing any genre of video. Summarization of videos in different languages can be added in the system to extend the use of the system. Passive voice generation functionality can be upgraded in the system.

References

[1] Pan G, Zheng Y, Zhang R, Han Z, Sun D and Qu X: A bottom-up summarization algorithm for videos in the wild (2019)

[2] Srinivas M, Pai M and Pai R: An Improved Algorithm for Video Summarization – A Rank Based Approach (2016)

[3] Kansagara R, Thakore D and Josh M: A Study on Video Summarization Techniques(2014)

[4] Rochan M and Wang Y: Video Summarization by Learning from Unpaired Data (2019)

[5] Landauer and Dumais: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge(1997)

[6] Aparıcioa M, Figueiredo P, Raposo F, Martins de Matosa D, Ribeiro R: Summarization of Films and Documentaries Based on Subtitles and Scripts (2015)

[7] Truong B. T. and Venkatesh S: Video abstraction: A systematic review and classification (2007)

[8] Ciocca G, Schettini R :Dynamic Key-frame Extraction for Video Summarization

[9] Yi H, Rajan D, and Chia L: Semantic Video Indexing and Summarization

[10] https://pypi.org/project/gTTS/