IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11613

VIDEO TO TEXT SUMMARY USING NLP TECHNIQUES

Bhupendra Gupta[1], Deeksha Yadav[2], Abhishek Singh[3], Tamil Arasu V[4], Dr. S. Vijay Kumar[5]

Department of Computer Science and Engineering, Jain University, Bangalore, India[1-5]

Abstract: Long videos captured by consumers are typically tied to some of the most important moments of their lives, yet ironically are often the least frequently watched. The time required to initially retrieve and watch sections can be daunting. In this work we propose novel techniques for summarizing and annotating long videos. Existing video summarization techniques focus exclusively on identifying keyframes and subshots, however evaluating these summarized videos is a challenging task. Our work proposes methods to generate visual summaries of long videos, and in addition proposes techniques to annotate and generate textual summaries of the videos using recurrent networks. Interesting segments of long video are extracted based on image quality as well as cinematographic and consumer preference. Key frames from the most impactful segments are converted to textual annotations using sequential encoding and decoding deep learning models. Summarization technique is benchmarked on the VideoSet dataset, and evaluated by humans for informative and linguistic content. We believe this to be the first fully automatic method capable of simultaneous visual and textual summarization of long

Keywords: Natural Language Processing, Deep Learning, Convolutional Neural Networks, Sequence to Sequence, Word2vector and Deep Speech Package**.**

I. INTRODUCTION

Driven by the exponential growth in the amount of online videos in recent years, research in video summarization has gained increasing attention, leading to various methods proposed to facilitate large-scale video browsing.As it for automatically identifying the important parts of a video and annotating the video. The main idea is to select the key frames from the video feed, and caption them for text summarization, so instead of watching full video we can save our time by reading the important content from the video.

In July 2015, YouTube revealed that it receives over 400 hours of video content every single minute, which translates to 65.7 years' worth of content uploaded every day1. Since then, we are experiencing an even stronger engagement of consumers with both online video platforms and devices (e.g., smart-phones, wearables etc.) that carry powerful video recording sensors and allow instant uploading of the captured video on the Web.

According to newer estimates, YouTube now receives 500 hours of video per minute2; and YouTube is just one of the many video hosting platforms (e.g., DailyMotion, Vimeo), social networks (e.g., Facebook, Twitter, Instagram), and online repositories of media and news organizations that host large volumes of video content. So, how is it possible for someone to efficiently navigate within endless collections of videos, and find the video content that s/he is looking for? The answer to this question comes not only from video retrieval technologies but also from technologies for automatic video summarization.

The latter allows generating a concise synopsis that conveys the important parts of the full length video. Given the plethora of video content on the Web, effective video summarization facilitates viewers' browsing of and navigation in large video collections, thus increasing viewers' engagement and content consumption.

Our proposed method uniquely identifies interesting segments from long videos using image quality and consumer preference. Key frames are extracted from interesting segments whereby deep visual-captioning techniques generate visual and textual summaries. Captions from interesting segments are fed into extractive methods to generate paragraph summaries from the entire video. The paragraph summary is suitable for search and organization of videos, and the individual segment captions are suitable for efficient seeking to proper temporal offset in long videos. Because boundary cuts of interesting segments follow cinematography rules, the concatenation of segments forms a shorter summary of the long video

DOI: 10.17148/IJARCCE.2022.11613

II. LITERATURE SURVEY

Video summarization research has been largely driven by parallel advancements in video processing methods, intelligent selection of video frames, and state-of-the-art text summarization tools. generates story driven summary from long unedited egocentric videos. They start with a static transit procedure to extract subshots from a longer egocentric video and extract entities that appear in each sub shot to maximize an order of k selected subshots while preserving influence over time and individual important events. In contrast, works with any kind of video (static, egocentric or moving), generates superframe cuts based on motion and further estimates the interestingness of each superframe based on attention, aesthetic quality, landmark, person and objects. uses video titles to find the most important video segments. explores a nonparametric supervised learning approach for summarization and transfers summary structure to novel input videos. Determinantal Point Process has also often been used in video summary methods . Using keyframes to identify important or interesting regions of video has proven to be a valuable first step in video summarization. For example, I used temporal motion to define a visual attention score. Similarly, utilized spatial saliency at the frame level. introduced cinematographic rules which pull segment boundaries to locations with minimum motion. favored frames with higher contrast and sharpness, favored more colorful frames, studied people and object content, while studied the role facial content plays in image preference. further tracked objects across a long video to discover story content.

Large supervised datasets along with advances in recurrent deep networks have enabled realistic description of still images with natural language text . The extension of this to video can be done by pooling over frames [34] or utilizing a fixed number of frames uses a temporal attention mechanism to understand the global temporal structure of video, in addition they also use appearance and action features through a 3-D Convolutional Neural Network (CNN) which encodes local temporal structure. Most recently,] described a technique, S2VT, to learn a representation of a variable sequence of frames which are decoded into natural text. Recently, demonstrated a hierarchical recurrent neural network to generate paragraph summaries from relatively long videos. These videos were still limited to a few minutes long. We use a variation of the S2VT captioning approach in our work. Given descriptive captions at keyframe locations, we explore extractive methods for summarization. Extractive methods analyze a collection of input text to be summarized, typically sentences. These sentences are selected to be included in the summary using various measurements of sentence importance or centrality. Early seminal summarization research by Luhn used word frequency metrics to rank sentences for inclusion in summaries, while Edmundson expanded this approach to include heuristics based on word position in a sentence, sentence position in a document, and the presence of nearby key phrases. More recent extensions of the word frequency models, including SumBasic and KL-Sum, typically incorporate more sophisticated methods of combining measures of word frequency at the sentence level and using these composite measures to rank candidate sentences. Other approaches, such as LexRank and TextRank focus on centroid-based methods of sentence selection, in which random walks on graphs of words and sentences are used to measure the centrality of those sentences to the text being summarized. A good review of these techniques and others can be found in . The latest research on single document summarization has utilized both dependency based discourse tree trimming as well as compression and anaphoricity constraints.

III. METHODOLOGY

Video To Text

- The input we use is in Video Format.
- Then we need to extract the Audio from the input Video using "Deep Speech Package", which contains two models.

1. Acoustic Model – It helps to convert the audio into text format (Someone talks into text format), it saves the file in " pbmm ".
2. Language Model – It helps to find the correct words (good vocabulary) by considering left and right words. And it saves files in " scorer ".

- After that we will use ffmak, which converts the audio into 16,000 KHz of frame rate.
- Using the number of frames generated by frame rates help to get the buffer size of audio (Buffer size means length of the audio).
- Then I will use the streaming method. Which parallelly converts the audio into text.
- Will read the audio in batch size of 65KB and print the text of 65Kb size of audio.

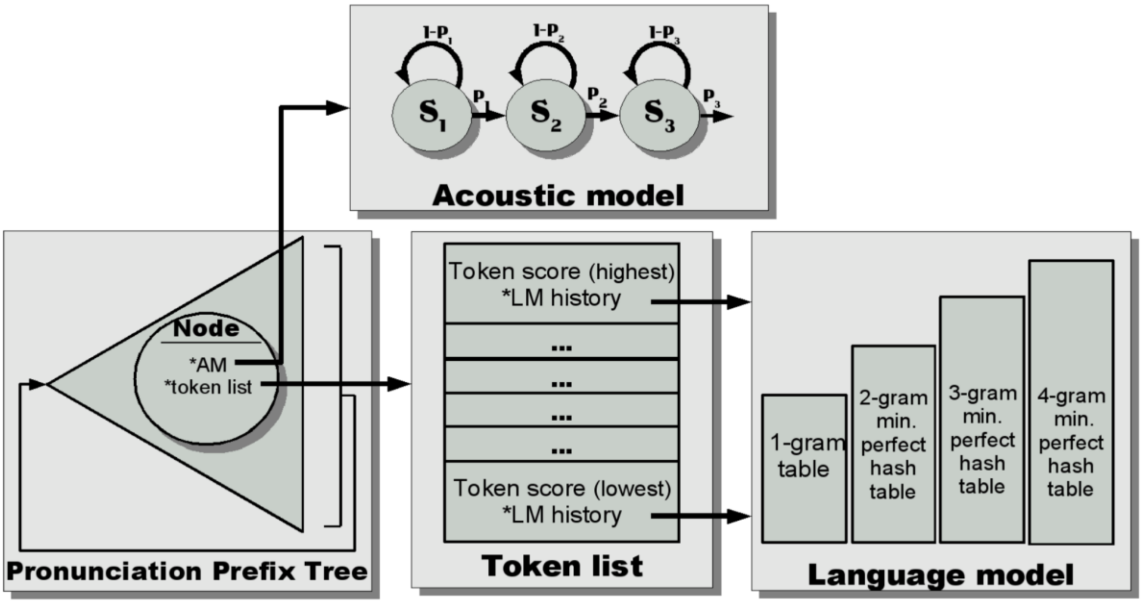
Text To Summary

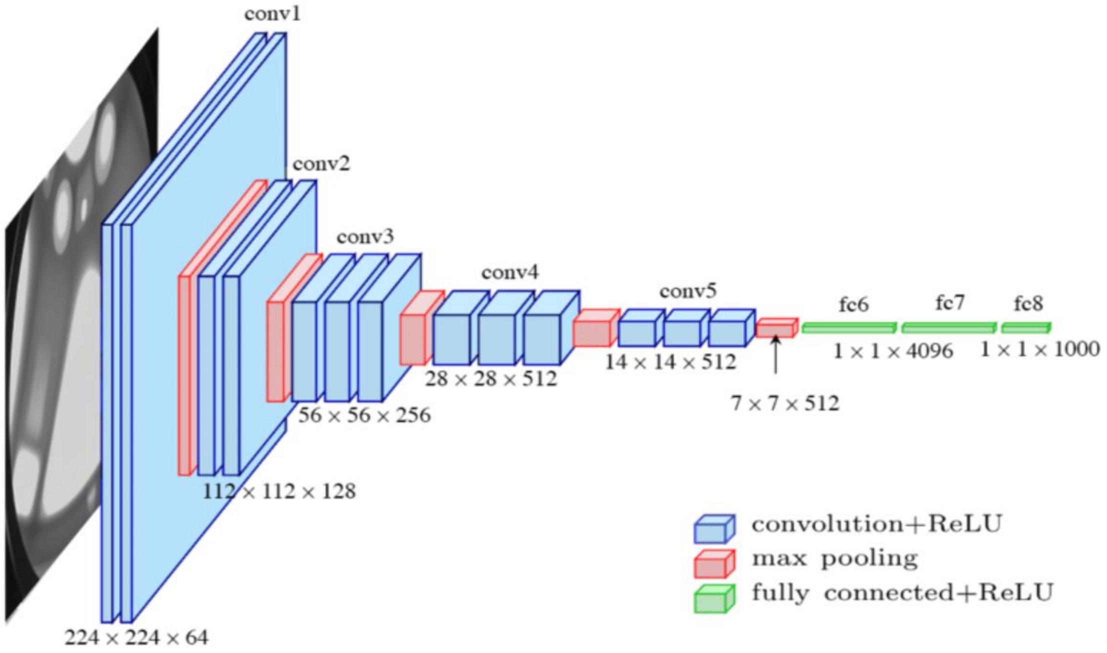
- The output of the previous algorithm will be the input for the next algorithm.
- Text format of the video will be the input for the next algorithm, which will convert into a summary of text using NLP.

DOI: 10.17148/IJARCCE.2022.11613

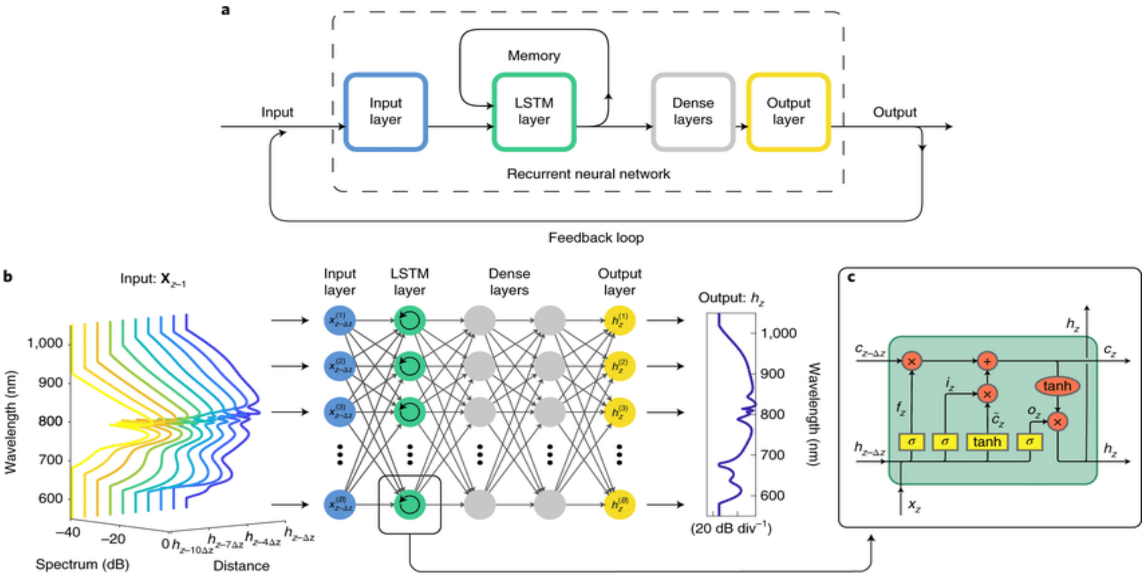
- Firstly, we need to use Tokenization or Split up for the text.
- Then using Neural Network will convert the Tokenize words into Vector using word2vec or Glove.
- After converting into a vector we will use the word_embedding() method, which provides a dense representation of words and their relative meanings.
- Will then create of model using sequence to sequence, which include Encoder and Decoder
- Encoder - It reads the input sequence and summarizes the information (in case of LSTM these are called the hidden state and cell state vectors).
- Decoder - It is a stack of several recurrent units where each predicts an output at a time step.
- And after using all these algorithms our model will give the output as a summary.

So, we finally get the out as Text summary format from a Video.





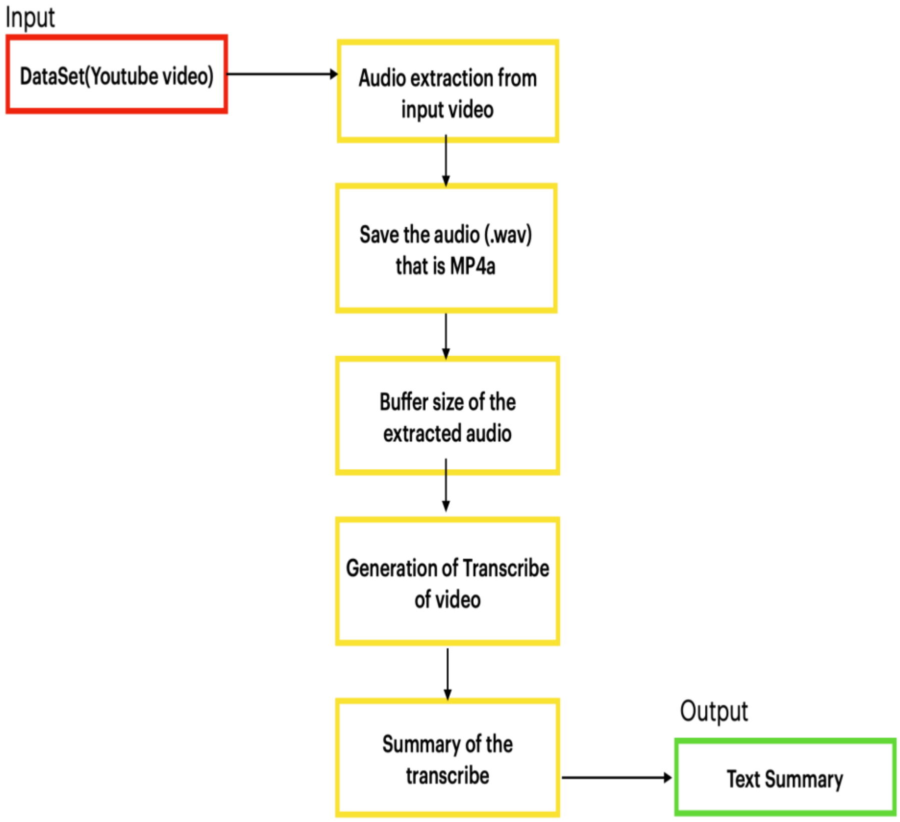
DOI: 10.17148/IJARCCE.2022.11613



IV. IMPLEMENTATION

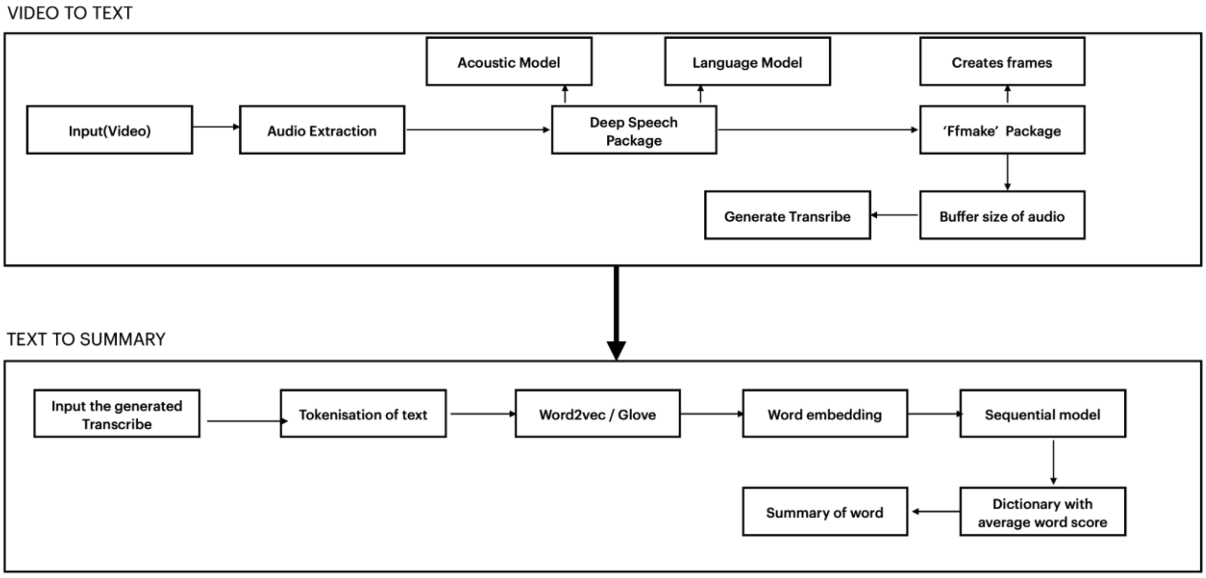
For this project we have chosen YouTube videos, in the present time everyone mostly use to prefer the YouTube videos for there reference. To implement the project, python with deep learning and natural language processing is being used. So, to built the model we have usen Deep Search Package which will help the user to convert the video into summary.

Architecture:


Input
DataSet(Youtube video)
Audio extraction from input video
Save the audio (.wav) that is MP4a
Buffer size of the extracted audio
Generation of Transcribe of video
Summary of the transcribe
Output
Text Summary

DOI: 10.17148/IJARCCE.2022.11613

Sequence diagram:



V. CONCLUSION

This paper introduces a novel method for both video summarization and annotation. Frame to frame motion, frame image quality, as well cinematographic and consumer preference are uniquely fused together to determine interesting segments from long videos. Key frames from the most impactful segments are converted to textual annotations using an encoder-decoder recurrent neural network. Textual annotations are summarized using extractive methods where LSA, LexRank and SumBasic approaches performed best. Human evaluations of video summaries indicate promising results. Independent experiments validate both superframe cuts as well as key frame selection. A key limitation is passing incorrect superframe or key frame information to the captioning framework. A potential solution would be availability of datasets with ground truth on both key segments and associated captions/summaries.

REFERENCES

[1]. D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 190–200. Association for Computational Linguistics, 2011.
[2]. X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. 2015.
[3]. R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In ECCV 2006, pages 288–301. Springer, 2006.
[4]. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE CVPR, pages 2625–2634, 2015.
[5]. G. Durrett, T. Berg-Kirkpatrick, and D. Klein. Learningbased single-document summarization with compression and anaphoricity constraints. arXiv preprint arXiv:1603.08887, 2016.
[6]. H. P. Edmundson. New methods in automatic extracting. Journal of the ACM (JACM), 16(2):264–285, 1969.
[7]. N. Ejaz, I. Mehmood, and S. W. Baik. Efficient visual attention based framework for extracting key frames from videos. Signal Processing: Image Communication, 28(1):34–44, 2013.
[8]. G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, pages 457–479, 2004.
[9]. G. Farneback. Two-frame motion estimation based on polynomial expansion. In Image analysis, pages 363–370. Springer, 2003.
[10]. J. Ghosh, Y. J. Lee, and K. Grauman. Discovering important people and objects for egocentric video summarization. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 1346–1353. IEEE, 2012.
[11]. A. Girgensohn and J. Boreczky. Time-constrained keyframe selection technique. In Multimedia Computing and Systems, 1999. IEEE International Conference on, volume 1, pages 756–761. IEEE, 1999.
[12]. G. Gkioxari and J. Malik. Finding action tubes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 759–768, 2015.

DOI: 10.17148/IJARCCE.2022.11613

[13]. B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In Advances in Neural Information Processing Systems, pages 2069–2077, 2014.

[14]. G. Guan, Z. Wang, S. Lu, J. D. Deng, and D. D. Feng. Keypoint-based keyframe selection. IEEE Tran on Circuits and Systems for Video Technology, 23(4):729–734, 2013.

[15]. M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In the European conference on computer vision, pages 505–520. Springer, 2014.

[16]. M. Gygli, H. Grabner, and L. Van Gool. Video summarization by learning submodular mixtures of objectives. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3090–3098, 2015.

[17]. A. Haghighi and L. Vanderwende. Exploring content models for multi-document summarization. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 362–370. Association for Computational Linguistics, 2009.

[18]. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015.

[19]. T. Hirao, Y. Yoshida, M. Nishino, N. Yasuda, and M. Nagata. Single-document summarization as a tree knapsack problem. In EMNLP, volume 13, pages 1515–1520, 2013.