Automatic Video summarization with Timestamps using natural language processing text fusion

Ahmed Emad, Fady Bassel, Mark Refaat,
Mohamed Abdelhamed, Nada Shorim, Ashraf AbdelRaouf
Faculty of Computer Science
Misr International University, Cairo, Egypt
ahmed1703326, fady1710742, mark1711712,
mohamed1701989, nada.ayman, ashraf.raouf{@miuegypt.edu.eg}

2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC) | 978-1-6654-1490-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/CCWC51732.2021.9376115

***Abstract*—In videos, description and keywords play an important role in the choosing process of the right video to watch. The main idea of the proposed approach is to generate descriptions and timestamps for videos automatically. Our approach plays an essential role in reducing the time consumed searching for the proper video. It aims to save time for users watching wrong unwanted videos and saves their time using timestamps. Timestamps would help to find and watch only the desired part of the video. One of the main goals of our approach is actual keyword extraction. Extracted keywords help finding videos with the significant video's keywords. The summarizing of the video depends on frames, emotions and speech. Firstly the video content appears in the frame and output a summarized text for the video content. Secondly, emotion and how it changes during a specific period merged with the outputted summarization of the frames. Thirdly, the audio transcribing into text occurs and output an abstractive summarization of the audio track. Finally, the fusion happens between all summarizations (audio, video, emotion) using natural language processing techniques. Techniques such as tokenization, sentence segmentation and lemmatization & stemming, and then abstractive summarization. Video summarization occurs to get a meaningful accurate description of the video. Having an accurate description helps finding the inquired content matching the description. The implemented experiment showed that on average 87% of the participants found generated text well representing the video.**

***Index Terms*—Video summarization, Text Fusion, Video Timestamps, Natural language processing, Long Short Term Memory LSTM, Face expressions.**

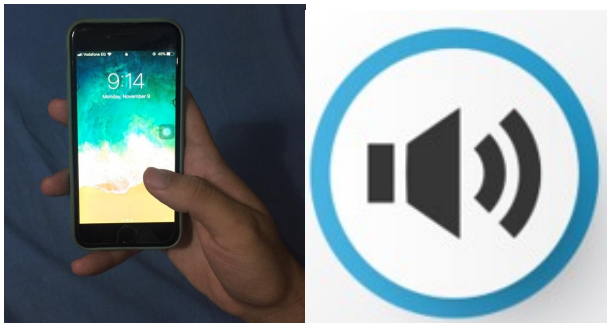
I. Introduction

There are millions of videos all over the internet, and more and more are being uploaded every second. As per the statistics done by YouTube [1], 1 billion hours of videos are being watched by users everyday and more than 500 hours of videos are being uploaded to YouTube every minute. This large collection of videos makes it very hard for the audience to find the desired video. Video platforms such as YouTube [2] introduce video descriptions and tags trying to overcome this problem, but the problem persists and another problem of click baits does appear.

Image processing techniques are usually used for any sort of video editing to enhance the overall outcome. Image processing techniques are now used in many other domains such as medical imaging and treatments [3, 4] and Natural language processing & understanding [5, 6, 7]. Image Steganography is one of the important research topic that uses image processing [8]. Online video chatting is also one of the topics that are used in research and real life [9].

As technology advances, people started to make some applications addressing this problem. Some applications that summarize the video or describe it were introduced lately. They mainly use algorithms such as Convolutional neural network - CNN [10] and Long Short-term Memory - LSTM [11]. They also use natural language processing [12]. Some of the applications uses video frames in generating the summarization and others use the audio from the video to do so. Both generating a video output. But as per our knowledge, none generates a textual output description for videos based on video frames, audio, and emotions. Which will give the most accurate results to what is really happening inside the video.

Fusion is the process of mixing multiple texts with each other by adding linking Words (such as but, and, where). Also by removing repeated words (such as he he), names and replace it with pronouns. Figure 1 shows an example of a text fusion process. In our proposed approach, the system will fuse between three outputted text, the first one is outputted from video (description of frames), the second one is emotions (extracted from video scenes), and the last one is the audio which is extracted from video through audio to text models.



(a) Someone is holding phone (b) The battery is 3000mah

Fig. 1: Fusion between modalities A and B happens that results in having output: The phone battery is 3000mah

978-0-7381-4394-1/21/$31.00 ©2021 IEEE

The problems facing people while dealing with videos and searching for them, encourage us to conduct a survey. The survey participants were people with different genders, ages, and interests. This survey showed some of the problems that people face while dealing with videos.

95.7% of the survey participants agreed that such a tool would definitely help them dealing with search process. More researches are required to investigate the topic that focuses on videos summarization with a good accuracy. Also the lack of tools that summarize videos with high accuracy. 87.2% of content creators agreed that such a tool could help them and save their time. As per the survey, 66.9% of the respondents assured that the video content does not match video title and description. Also 92.1% of them showed their need to have timestamps for videos. That could save students/researchers so much time finding the exact topic or part that they need to watch.

The paper is organized as follows: Section II describes the related work. The proposed approach is defined in Section III. The experiment is described in Section IV and the conclusions of the research are presented in Section V.

II. Related Work

Based on past work done in the domain of video summarizing, two main approaches were carried out. First approach was carried out by Tran and Hwang [13] As well as Ji and Xiong [14]. Their main target was to output a shorter version of the input video. The other approach was to output a text description of the input video. Rohrbach et al. [15] focused on identifying verbs, objects, and places. They used LSTM for sentence generation. Also Krishnamoorthy and Malkarnenkar [16] focused on extracting subjects, objects, and verbs from video frames.

A. Video to text and feature extraction

Sah and Kulhare [17] proposed a system that summarizes video to text with frames only. They used S2VT (A two-layer LSTM model to learn video representation in the encoder and word representation in the decoder). Also used 152-layer ResNet CNN model pre-trained on ImageNet data to extract to important frames. The dataset used is VideoSet, which consists of eleven long (45 minutes to over 5 hours) videos in three categories. These categories are: Disney, egocentric, and TV episodes. Eight videos are used for training and three for testing. The captioning model was pre-trained on the training split of the MSVD dataset. They suggest that the LexRank, LSA, and SumBasic methods are generally performing better results.

Dilawari and Khan [18] proposed a system that uses CNN and RNN to extract features from video frames. They take the video as an input and start to pass video frames to CNN model and then to RNN to extract visual human features, i.e., age, emotions, and gender. These features are passed to the LSTM model to generate a description. This description passed throw word-level Luong attention distribution to make an abstractive summarization. They used MSR-VTT dataset for CNN and RNN models. They used METEOR score to compare their job with 4 others methods. These methods are: (MP-LSTM, SCN-LSTM, Task Specific Feature encoding, and SCN-LSTM) and proved that they outputted the best results. After that, they build UET Surveillance dataset which contains videos and their textual description.

Moses and Balachandran [19] presented a system that summarizes videos based on frames and has an output summarized video as a result. They first used skimming and feature and key frames extraction. Then, applied classification, clustering and machine learning algorithms. They used these algorithms to solve the problem and get the best result for the summarized video. The datasets used by them to test and train the system are ViSOR, CAVIAR, CUHK, and InHouse datasets. They aim in this project to help a good system and to save storage by having summarized video. However, this could be more efficient and save more storage if this summary were a text that consumes less space than video.

Jin and Liang [20] presented a system that uses LSTM-RNN and CNN algorithms as shown in figure 2. They found lately that the algorithms are used at videos and give very good results. They categorize the videos based on two options, action recognition and end-to-end sequence model. They tested and trained their model using Corpus dataset. This dataset includes nearly 2000 videos from 10 to 25 seconds. METEOR score of 23.70% and 20.21% for video and audio respectively, and scores 26.17% for the combination between them.

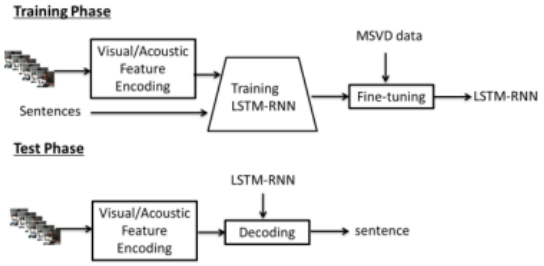


Fig. 2: Training and testing phases used by Jin and Liang [20]

Guang Li et al. [21] proposed a system that treats the video as a sequence of frames. Their input is a video and they output a sentence describing the content of the video. They used a combination of deep CNN and LSTM-RNN algorithms. They output a number of candidate sentences, they then apply a ranking algorithm. Ranking aims to find similar sentences and choose the most repeated sentences as the video description. The dataset used was the Microsoft Research Video Description Corpus.

Jeevitha and Hemalatha [22] presented a system that generates a textual description for videos based on frames only. They use 2D and 3D CNN, NetVLAD, and aLSTM. The dataset they use is MSVD, Microsoft video description corpus, it consists of 1970 short video clips, with 80,000 descriptions and 16,000 words. They used BLEU quantitative method to measure their accuracy. According to their tests, the results

said that their new approach (NetVLAD+aLSTM) is always better than traditional approaches such as (C2D, C3D, and C2D+C3D).

Li and Zhu [23] proposed an approach that combines related documents and videos of a topic. The salience score of text was measured, including sentences in documents and transcripts of speech from recordings. Salience score is used to provide information about the importance of the word relative to the entire document. But, according to their conclusion, the inclusion of video and audio does not improve the performance relative to the text-only model. That makes sense, as it would be necessary to summarize a well-written document to get the best results.

B. Audio to text

Zhang and Tian [24] presented a system that uses 3D CNN instead of 2D CNN with LSTM model. They used three datasets: the Microsoft Video Description corpus (MSVD), the MPII Movie Description dataset (MPII), and the Montreal Video Annotation Dataset (MVAD). This research takes video as an input and extracts two types of frames, RGB frames and MHI frames. Each type of frames enter a 3D CNN model to extract features. The output from the two 3D CNN models enter LSTM to summarize the video. Their method achieved the best METEOR score; 31.1% which is 7.2% higher than FGM, 2% higher than Mean-pooling, 1.5% higher than temporal attention model and 1.3% higher than the previous S2VT model.

Ribeiro and Matos [25] introduced an approach that enhances the ordinary speech to text summarization. They used LSA where they give the model background related keywords to train on, so it gives better abstractive summarization. They use the news articles as their dataset. A group of people were used to evaluate and compare the summarizations of the different models. Those models are: (ASR, basic LSA, LSA, human extractive, and human abstractive). Results showed that the best was human extractive. However, the LSA showed better results than the ASR.

C. Multimodal audio and video

Haoran Li et al. [23] introduced a system that aims to bridge the semantic gaps between different multimodal content; NLP, ASR and CV based summarization. They use documents including text and images, videos including their frames, speech transcription, and audio features. They used MSRParaphrase dataset which consists of 5,801 pairs of sentences to train their model. As for the text part, Flickr30K and MSCOCO datasets, Flickr30K contains 31,783 photographs of different activities. MSCOCO dataset consists of 123,000 images. Both contains textual descriptions of the images to train the visual part of their model. To test their model, they created a dataset of news articles, video, and images and tried to summarize it. The results showed that adding the audio and video of the news does not really enhance the results. However, the most effective factor is the text, whether it is from a topic or transcribed from an image or audio. All their ROUGE scores have a 95% confidence interval of at most ±0.25.

Chiori Hori et al. [26] proposed a multimodal fusion for describing videos. Where each modality has its own sequence of feature vectors. The modalities are motion, image and audio features. They use C3D to extract frame features as a vector and MFCC for audio feature extraction. Then these vectors are fused together into one vector. Which is used to predict the next word by encoder-decoder sentence generation using RNN. They evaluated their method on the YouTube2Text and MSR-VTT datasets.According to 3 evaluation metrics (BLEU4, METEOR, CIDEr), their method showed the highest scores. Comparing to LSTM-E, TA, h-RNN and Naive fusion after using YouTube2Text subset without Overdubbed Music.

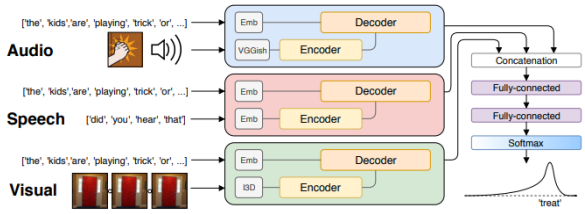


Fig. 3: System overview presented by Lashin and Rahtu [27]

Lashin and Rahtu [27] proposed an approach that takes audio, speech and video as inputs. As well of a sequence of the already generated words as shown in figure 3. The model processes the input in the corresponding decoder encoder block, then all three outputs are fused together. They are fused together in the multimodal generator to produce the next caption word. For audio feature extraction, they used VGGISH. They used I3D for visual feature extraction and word embedding to represent text as vectors. The dataset used was the activity net dataset. This dataset contains 10K videos for training and 5K for validation. They used METEOR metric as it has been shown to be highly correlated with human judgment. They score 10.09 when using video, audio, and speech and make a fusion between them.

III. Proposed Approach

The proposed approach implements a new way for summarizing videos based on their audio, video, and emotions all together. That aims to achieve a good readable and accurate description of video. Also, the fusion of the summarized texts outputted from audio, video, and emotions. To get one summarized text at the end that could be a meaningful description for the video.

Survey results showed that 73.5% of the respondents did not create or upload videos to any video platforms. But for the 26.5% who do upload videos, 33.3% of them do not write a description for their videos. Sometimes, they write only a few words that do not describe the actual video content. 87.2% of them agreed if they had such a tool, they would use it.

The proposed system overview is shown in figure 4 which includes the following steps:

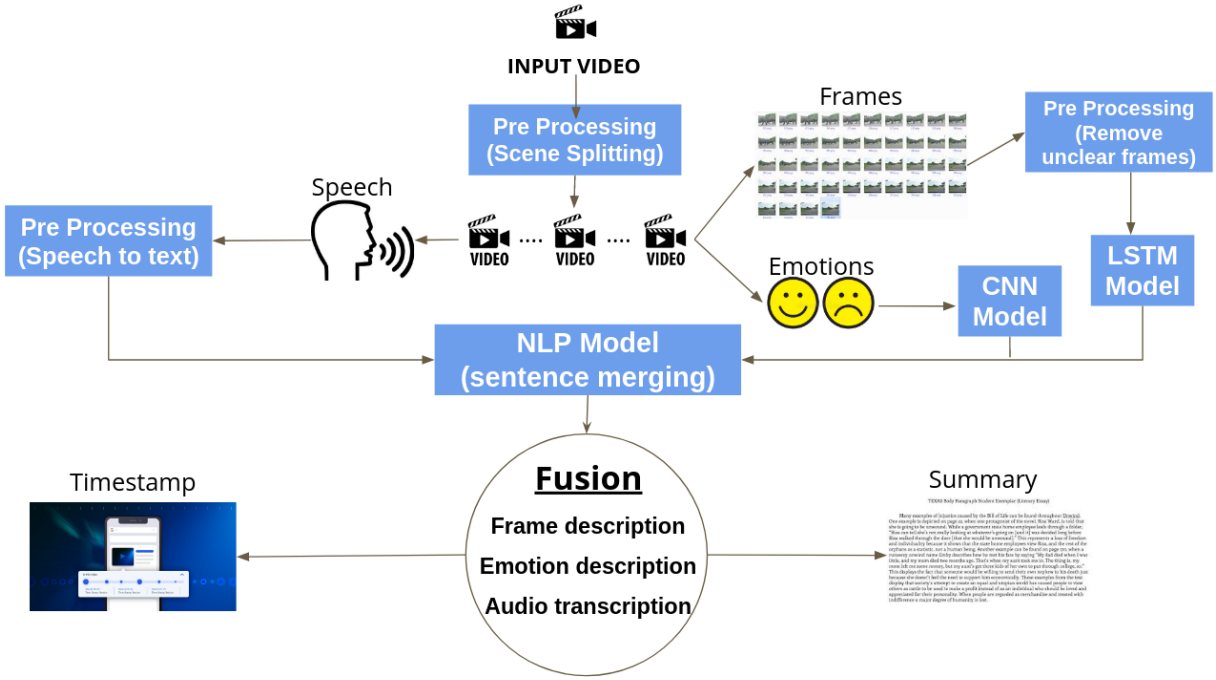


Fig. 4: Proposed System Overview

1) Every couple of seconds, the video is separated into smaller videos. This splitting is done according to scenes hence each video into frames and audio.
2) Each frame of the smaller videos frames is passed through the preprocessing phase. The frames and audio will be preprocessed by removing blurry scenes and noisy audio.
3) The audio is transcribed into text by applying speech to text recognition. The output is a text and then passed to LSTM and CNN models.
4) The three text outputted descriptions are fused together to output a summarized fused meaningful text.
5) Repetition of this process occur to complete the remaining of all smaller videos.
6) An abstractive summarization would be then applied to text using Natural Language Processing (NLP) techniques.
7) NLP is used to produce the video description and timestamps. Having cuts at the scenes and labeling it help finding it later.

A. *Video Summarization*

Every couple of seconds, the video is separated into smaller videos based on scenes. Each smaller video is divided into frames. Then the filtration process for frames occurs. Filtration process occurs by removing the blurry and non-clear frames that will not be used. The filtered frames will pass through LSTM model. LSTM model encodes sequence of frames and then the model decodes their values to have an output text description for the frames. The model output a text description and caption for the video. The text is generated based on the frame content and what appears in the video.

B. *Emotions Summarization*

After the completion of video summarization, the smaller videos are passed by CNN model. The model responsibility is to classify the faces in the frames based on their emotion. Our proposed model is trained to classify 7 different emotions anger, disgust, fear, sadness, surprise, joy, and neutral. Emotions should be extracted from each frame for every period of time to get the major emotion of the scene. The output dominant emotion is later combined with the description in order to result in more accurate and reliable description.

C. *Audio Summarization*

Audio is filtered first by removing unclear voices, noisy and empty parts of the audio. The filtered audio is passed to speech-to text transcription process. This process goal is to get the audio script of the video. The script will then go throw the text ranking algorithm [28] to summarize it. By extracting all sentences from the document and arranging it. The arrangement occurs in descending order of their order in similarity matrix based on the similarity between sentence vectors. The first sentences of the similarity matrix with higher weight are chosen to be part of the summarized text. This abstractive summarization rephrases the words to get a more accurate summary.

D. *Fusion generated text*

Each small video outputs three sentences, one from audio, one from video, and one from emotions. The fusion occurs between these three sentences by adding linking words, removing

duplicate names. All of these are applied using NLP techniques such as tokenization [29], sentence segmentation [30] and stemming & lemmatization [31]. The output will be one sentence for each small video. All these sentences combined making a paragraph that describe the main video. Also extracts the most important keywords for video.

E. Timestamps Generation

While summarization process, timestamp process occur. Every scene is detected at its start and its end to know the main scene. Also make a summarized title for each scene which is used for the navigation process later to have a faster and more accurate search.

IV. Experiment

A. Dataset

Dataset is very important in implementing any machine learning approach and testing the experimental work. We are using two different datasets during our experiments for the training and testing process. FER-2013 [32] (Fig. 5) and MSVD [33] (Fig. 6) are used for their variety of videos that could be used to benefit our approach.

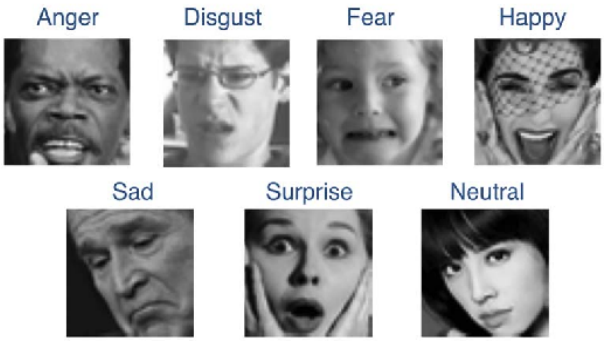


Fig. 5: Samples of FER-2013 Dataset [32]

FER-2013 or facial expression recognition dataset consists of 48*48 grayscale images of cropped faces. The data were categorized into seven categories, each corresponding to a particular emotion, among angry, happy, surprise, disgust, neutral, sad and fear. The training set consists of 28,709 images and the public test set consists of 3,589 images.

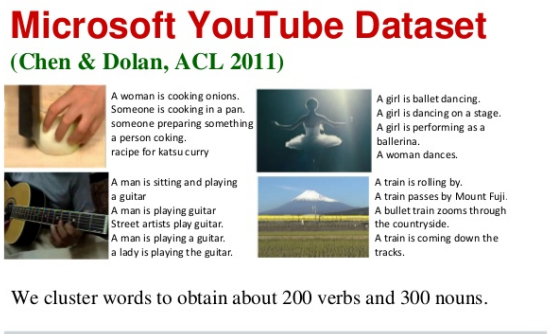


Fig. 6: Samples of MSVD Dataset [33]

MSVD is a benchmark dataset that consists of 1,500 videos with average 10 seconds for each video. Every video has several captions that used for training. These captions are also used in evaluating the output captions in testing. 1,450 videos for training and 50 videos for testing.

B. Experiment setup

Experiment occurred using two machines. The first one is HP core i5-7200U, 12 GB of ram, 240 SSD - running Windows 10 operating system. The second one is HP core i5-8300h, 8 GB of ram 240 SSD - running Linux Ubuntu operating system. The Python Pycharm programming tool is used for developing the approach.

C. Implemented work

1) Video Summarization: The video to text conversion and summarization consists of the following main phases:

- *Frame Extraction:* Video is split into frames.
- *Face and Emotion Detection:* Faces are detected in all extracted frames. The emotion of the extracted faces is used to get the emotional score of every frame.
- *Extract emotion:* Retrieves the most repeated emotion of a given time interval in a video.
- *Speech to text:* Apply a speech-to text API on the video with the start and end time of each word.
- *Combine text and summarize:* Concatenate emotion states with the corresponding speech given a time interval of 20 seconds and then generate a summarization of the resulted text. an external API (AssemblyAI) [34] is used.

2) Emotion detection: A deep learning model (CNN), which is used for the implementation of face Detection and emotion classification. It presents the probability distribution of 7 different emotions as follows, anger, disgust, fear, sadness, surprise, joy, and neutral. This is done based on the work of kousik [35] by applying changes to the output form of the data.

D. Experiment results

The experiment was tested on a three-minute ted talk video conducted by a young speaker Luke Bakic titled ”the importance of reading” [36]. The resultant text summarization in fig 7 was 217 words in total. While the original text before summarization was 485 words. The run time of the experiment took 4 minutes to output the results on the machine mentioned in the experiment setup.

E. Experiment human evaluation

We created a survey to ask people to evaluate our proposed system and its implemented results so far. The survey was conducted by 151 people, 60% females and 40% males. 85% of of the participants age was ranging between 15-25 years. The survey started by asking the participants to watch a three minute ted talk video and then evaluate the performance of the automated description generated by our proposed system by answering three linear scale questions. Figure 7 shows the summary of the video generated by our proposed system. The

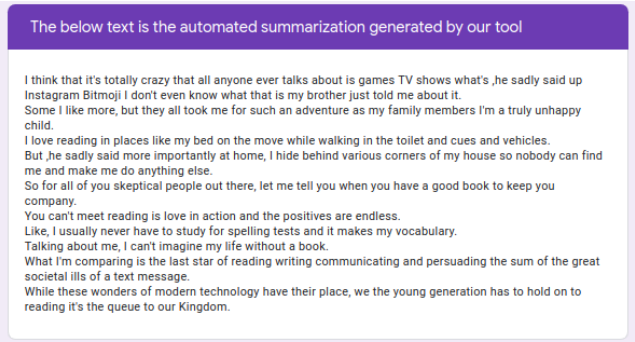


Fig. 7: Output summary of video

three linear scaled questions were all ranged from 1 to 5 and were as follows:

1) How much do you agree that the summary describes the video content? Knowing that 1 is strongly disagree and 5 is strongly agree.

2) How much do you rate the clarity/understanding of the summary sentences? Knowing that 1 is very poor and 5 is excellent.

3) How much do you rate the accuracy of the generated emotions for example:"he sadly said" based on the speaker facial expressions? Knowing that 1 is very poor and 5 is excellent.

According to question 1 89.5% of the participants gave a score between 3-5, question 2 percentage was 86.2% and question 3 percentage was 85.5% for the same score range. the survey results showed that 87% of participants are happy with the generated application.

F. Experiment conclusion

Experiment was conducted on only two modalities of data that are proposed by our system (speech and emotion). Experiment aim was to evaluate our approach so far and gain feedback on future work. Video captioning based on frames will help greatly in nonverbal scenes. Also more work should be applied in text fusion using NLP techniques for using linking words and removing duplicates to give smoother and more accurate sentences.

V. Conclusion

This paper proposes a new state of the art approach for video text summarization based on audio, video content, and emotions. The approach extracts the output text from the video that is done by dividing the video into smaller videos, and then the preprocessing phase occurs. The output is a summarized text from frame description. Another text is extracted from emotions through deciding the major emotion during the scene. For the audio, the audio track is transcribed into another third text and have abstractive summarization. All the three summarized extracted texts are to be fused together using NLP. To get a meaningful description, besides finding the most important keywords in the video. Extracting keywords occurs by finding the most important keywords, that shall help in the searching process of videos. Also, generating automatic timestamps for videos for a faster, easier, and accurate search for the desired content.

References

[1] *YouTube for Press, Stats for youtube*. 2020 (accessed November 7, 2020). URL: https://www.youtube.com/about/press/.

[2] *Youtube*. (accessed November 9, 2020). URL: https://youtube.com.

[3] Taraggy M Ghanim, Mahmoud I Khalil, and Hazem M Abbas. "Multi-stage Off-line Arabic Handwriting Recognition Approach using Advanced Cascading Technique." In: *ICPRAM*. 2019, pp. 532–539.

[4] Salma Sameh, Mostafa Abdel Azim, and Ashraf AbdelRaouf. "Narrowed coronary artery detection and classification using angiographic scans". In: *2017 12th International Conference on Computer Engineering and Systems (ICCES)*. IEEE. 2017, pp. 73–79.

[5] Taraggy M Ghanim, Mahmoud I Khalil, and Hazem M Abbas. "Comparative Study on Deep Convolution Neural Networks DCNN-Based Offline Arabic Handwriting Recognition". In: *IEEE Access* 8 (2020), pp. 95465–95482.

[6] Ashraf AbdelRaouf et al. "Arabic character recognition using a Haar cascade classifier approach (HCC)". In: *Pattern Analysis and Applications* 19.2 (2016), pp. 411–426.

[7] Nada Shorim, Taraggy Ghanim, and Ashraf Abdel-Raouf. "Implementing Arabic Handwritten Recognition Approach using Cloud Computing and Google APIs on a mobile application". In: *2019 14th International Conference on Computer Engineering and Systems (ICCES)*. IEEE. 2019, pp. 88–95.

[8] Ashraf AbdelRaouf. "A new data hiding approach for image steganography based on visual color sensitivity". In: *Multimedia Tools and Applications* (2021). URL: https://doi.org/10.1007/s11042-020-10224-w.

[9] Nada Radwan, MB Abdelhalim, and Ashraf Abdel-Raouf. "Implement 3D video call using cloud computing infrastructure". In: *Ain Shams Engineering Journal* 11.2 (2020), pp. 363–375.

[10] S. Albawi, T. A. Mohammed, and S. Al-Zawi. "Understanding of a convolutional neural network". In: *2017 International Conference on Engineering and Technology (ICET)*. 2017, pp. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[11] Ralf Staudemeyer and Eric Morris. "Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks". In: (Sept. 2019).

[12] James F. Allen. "Natural Language Processing". In: GBR: John Wiley and Sons Ltd., 2003, pp. 1218–1222. ISBN: 0470864125.

[13] Quang Dieu Tran et al. “Exploiting character networks for movie summarization”. In: *Multimedia Tools and Applications* 76.8 (2017), pp. 10357–10369.

[14] Z. Ji et al. “Video Summarization With Attention-Based Encoder–Decoder Networks”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.6 (2020), pp. 1709–1717. DOI: 10.1109/TCSVT.2019.2904996.

[15] Anna Rohrbach, Marcus Rohrbach, and Bernt Schiele. *The Long-Short Story of Movie Description*. 2015. arXiv: 1506.01698 [cs.CV].

[16] Niveda Krishnamoorthy et al. “Generating Natural-Language Video Descriptions Using Text-Mined Knowledge”. In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI’13. Bellevue, Washington: AAAI Press, 2013, pp. 541–547.

[17] S. Sah et al. “Semantic Text Summarization of Long Videos”. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017, pp. 989–997. DOI: 10.1109/WACV.2017.115.

[18] A. Dilawari and M. U. G. Khan. “ASoVS: Abstractive Summarization of Video Sequences”. In: *IEEE Access* 7 (2019), pp. 29253–29263. DOI: 10.1109/ACCESS.2019.2902507.

[19] T. M. Moses and K. Balachandran. “A classified study on semantic analysis of video summarization”. In: *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*. 2017, pp. 1–6. DOI: 10.1109/ICAMMAET.2017.8186684.

[20] Qin Jin and Junwei Liang. “Video Description Generation Using Audio and Visual Cues”. In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ICMR ’16. New York, New York, USA: Association for Computing Machinery, 2016, pp. 239–242. ISBN: 9781450343596. DOI: 10.1145/2911996.2912043. URL: https://doi.org/10.1145/2911996.2912043.

[21] Guang Li, Shubo Ma, and Yahong Han. “Summarization-Based Video Caption via Deep Neural Networks”. In: *Proceedings of the 23rd ACM International Conference on Multimedia*. MM ’15. Brisbane, Australia: Association for Computing Machinery, 2015, pp. 1191–1194. ISBN: 9781450334594. DOI: 10.1145/2733373.2806314. URL: https://doi.org/10.1145/2733373.2806314.

[22] V. K. Jeevitha and M. Hemalatha. “Natural Language Description for Videos Using NetVLAD and Attentional LSTM”. In: *2020 International Conference for Emerging Technology (INCET)*. 2020, pp. 1–6. DOI: 10.1109/INCET49848.2020.9154103.

[23] H. Li et al. “Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video”. In: *IEEE Transactions on Knowledge and Data Engineering* 31.5 (2019), pp. 996–1009. DOI: 10.1109/TKDE.2018.2848260.

[24] Chenyang Zhang and Yingli Tian. “Automatic video description generation via LSTM with joint two-stream encoding”. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. 2016, pp. 2924–2929. DOI: 10.1109/ICPR.2016.7900081.

[25] Ricardo Ribeiro and David Martins De Matos. “Mixed-source multi-document speech-to-text summarization”. In: *Coling 2008: Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization*. 2008, pp. 33–40.

[26] C. Hori et al. “Attention-Based Multimodal Fusion for Video Description”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 4203–4212. DOI: 10.1109/ICCV.2017.450.

[27] V. lashin and E. Rahtu. “Multi-modal Dense Video Captioning”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 4117–4126. DOI: 10.1109/CVPRW50498.2020.00487.

[28] David Ten. *Keyword and Sentence Extraction with TextRank (pytextrank)*. 2018 (accessed November 7, 2020). URL: https://xang1234.github.io/textrank/.

[29] Srinivas Chakravarthy. *Tokenization for Natural Language Processing*. (accessed November 10, 2020). URL: https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4.

[30] David D Palmer. “Tokenisation and sentence segmentation”. In: *Handbook of natural language processing* (2000), pp. 11–35.

[31] Julie Beth Lovins. “Development of a stemming algorithm”. In: *Translation and Computational Linguistics* 11.1 (1968), pp. 22–31.

[32] *Facial expression recognition dataset*. URL: https://www.kaggle.com/msambare/fer2013.

[33] David Chen and William Dolan. “Collecting Highly Parallel Data for Paraphrase Evaluation”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 190–200. URL: https://www.aclweb.org/anthology/P11-1020.

[34] *#1 Most Accurate Speech-to-Text API*. (accessed November 10, 2020). URL: https://www.assemblyai.com/.

[35] *Video expression recognition implementation by kousik97*. URL: https://github.com/kousik97/Video-Expression-Recognition.

[36] *The experiment video ’The Power and Importance of...READING!’* URL: https://www.youtube.com/watch?v=rW2r5uStgG0&ab_channel=TEDxTalks.