Video and Text Summarization Using VDAN and RNN

Joys Princia A[1], Ms. J Sangeetha Priya[2], Kalai Selvi J[3], Rithi Afra J[4], Rukshana S[5]
[1,3,4,5]Saranathan College of Engineering
[2]Assistant Professor, Saranathan College of Engineering

***Abstract* - The main intent of this project is to develop a video and text summarizer. The videos in these days are so quite long. It is difficult for people with hectic work schedule to find time to watch the long videos. Thus, a summarizer will help people in getting the gist immediately. The video summarization is done with the help of Visually Guided Document Attention Network (VDAN). The motive of this network is to extract the textual and visual features. The extraction of visual features is done with the help of Convolutional Neural Network (CNN). The extraction of textual features is done with the help of document level encoding. It also contains Gated Recurrent Unit (GRU). Based on the visual and textual features extracted, the agent decides the corresponding action. The three sets of actions are accelerate, decelerate and do nothing. The text summarization part is done with the help of Recurrent Neural Network. It follows an encoder-decoder architecture. It also makes use of Long Short-Term Memory (LSTM) to keep track of the previous observations. Thus, at the end the summarized video and text are available.**

I.INTRODUCTION

Recently, the rate of production of videos has increased due to the increased ability of individuals to capture or/and create digital videos. But despite numerous technological advancements, there is a pitfall in this advent. In order to solve this problem and extract meaningful information from the video content, video summarization techniques are implemented. Video summarization is the process of reducing a video sequence to a small number of still images known as keyframes, also known as storyboard or thumbnail extraction, or a shorter video sequence made up of keyshots.

An ideal video summarization is that can provide users the maximum information of the target video with the shortest time. Its goal is to produce a compact yet comprehensive summary to enable an efficient browsing experience. The video summary need to convey most of key information contained in the original video.[3]

- It propose an improved three-dimensional (3D) action recognition Convolutional Neural Network (CNN) based on Residual Network (ResNet) to be used as a feature extractor for required video.
- Here, it train a Long Short Term Memory (LSTM) network on a video extracted by the proposed 3D-CNN. It use the combined 3D-CNN and LSTM network as the highlight recognition framework.
- Our project implements basic but effective method to produce summarized videos based on the 3D-CNN and LSTM highlight recognition framework.

Text Summarization has always been an area of active interest in the academia. In recent times, even though several techniques have being developed for automatic text summarization, efficiency is still a concern. Given the increase in size and number of documents available online, an efficient automatic news summarizer is the need of the hour. So, as a solution text summarization is used here. A technique of text summarization is that it focuses on the problem of identifying the most important portions of the text and producing coherent summaries. This text summaries reduces the time and the make the process easier. The standard way of doing text summarization is using seq2seq model with attention.[10]

Deep learning is an artificial intelligence subset that uses neural networks to learn unsupervised from unstructured or unlabeled data. Deep neural learning or a deep neural network are other terms for the same thing. Deep learning techniques on the other hand learn feature representations automatically. It has been observed that one of the most recent impressive deep learning strategies capable of human action discrimination has been discovered. Deep-learning architectures like deep neural network, deep belief networks, recurrent neural networks and convolutional neural networks are applied to fields including text summarization, video summarization, speech recognition, etc.

A Convolutionary Neural Network (CNN) can be a class of Deep Neural Networks in Deep learning, most generally applied to visual imagery analysis. Recurrent Neural Network (RNN) are a category of neural networks that are helpful in modeling sequence data.[3]

II. LITERATURE REVIEW

Chen et al. proposed that video summarization and video captioning are considered two separate practices in current research. For extended videos, it will allow for a richer and more succinct condensation of the video to automatically recognize the important sections of the video content and annotate them with captions. In the training process, they have proposed a general neural network configuration that jointly considers two monitoring signals (i.e. an image- based video summary and text-based video captions) and produces both a video summary and corresponding captions within the test phase for a given video. Main idea is that the summary signals can help a video captioning model learn to concentrate on important frames. On the opposite hand, caption signals can help a video summarization model to hunt out better semantic representations. Modeling both the video summary and therefore the video captioning tasks together provides a completely unique end-to-end solution that creates a captioned video summary that allows users to index and navigate during a video through the highlights. In addition, in both individual tasks, their studies show that the joint model is able to perform better than the state-of-the-art approaches. The limitation here is short term dependencies of simple RNN can’t remember and understand the context behind the input they give.[3]

Zhou et al. has given that by providing short, descriptive summaries that are diverse and reflective of original videos, video summarization aims to promote large- scale video browsing. In this paper, they formulate video summarization as a sequential decision-making process and develop a deep summarization network (DSN) to summarize videos. DSN predicts for every video frame a probability, which indicates how likely a frame is chosen, then takes actions supported the probability distributions to pick frames, forming video summaries. The authors have proposed an end-to-end reinforcement learning based framework, where they design a totally unique reward function that jointly accounts for diversity and representativeness of generated summaries and doesn’t believe labels or user interactions within the smallest amount. The reward role judges during training how diverse and representative the summaries produced are, while DSN aims to receive higher rewards by learning to provide more diverse and more representative summaries. Extensive experiments on two benchmark datasets show that their unsupervised method not only outperforms other state of unsupervised methods, but is additionally like or even superior than most of published supervised approaches. Disadvantage is frames generated as output will not help people to understand clearly on the actual content of the video delivered.[7]

Agyeman et al. introduces a deep learning approach to summarizing long soccer videos by leveraging the three-dimensional Convolutional Neural Network (3D-CNN) and Long STM (LSTM). Their suggested solution includes 1) the step-by-step creation of a 3D-CNN based Residual Network (ResNet) recognizing soccer behavior, 2) the manual annotation of 744 soccer clips from 5 training soccer action groups, and 3) the training of an LSTM network on soccer features extracted by the proposed 3D-CNN based ResNet. To detect soccer highlights, they combined the 3D-CNN and LSTM models. To summarize a soccer match video, they modelled the video input as a sequential concatenation of video segments whose inclusion during a summary video production is predicated on its validated relevance. Since recurrent networks has been used, computing the frames for highlights is a slow process and this acts as a demerit for this paper.[13]

Luthra et al. used a semi-supervised learning algorithm to generate the summaries. Manually generated summaries (ideal summaries) is that the labeled samples for the semi- supervised learning. Both visual and aural feature vectors appropriated a gaggle of videos and they are clustered and thus the individual video sequences are represented as vector-quantized statistic. Then state transition machine-based representation has been generated for both the entire class of videos and thus the samples are labeled. A replacement of information theoretic measure has been proposed for the goodness of a generated summary, which reduces the summarization process to finding the sequence of frames that the price of goodness measure is maximum. Visual gaps and breaks are created between the frames or skims generated is the disadvantage here.[16]

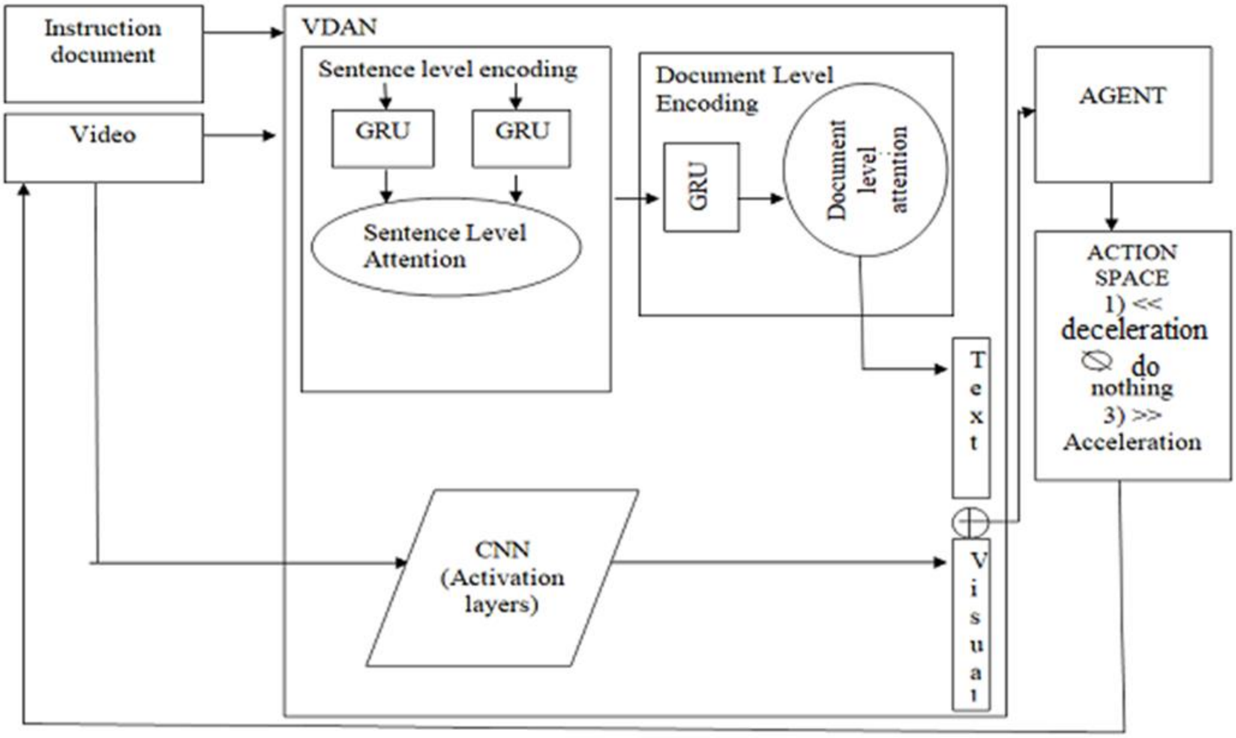
III. SUMMARIZATION

VIDEO SUMMARIZATION

The proposed system consists of two phases. They are video summarization and text summarization. The ultimate aim of the video summarization phase is to produce a fast-forward video. In this phase, initially the instruction document and the video are passed as input to the Visually-Guided Document Attention Network (VDAN). The main intent of VDAN is to extract the visual and textual features. For this purpose, an embedding space is created by the VDAN for encoding documents and images. The irrelevant frames in the video are dropped at the end. The visual features are extracted by the Convolutional Neural Network (CNN). Based on the extracted visual and textual features, the agent will decide an action from the three actions that are available. The three actions are acceleration, deceleration and doing nothing.

The agent is visually and textually guided with features produced by the VDAN to control the speed up rate. The action which is selected by the agent changes the skip rate and returns a reward based on the alignment of the textual and visual features in the embedding space. The visual features which are extracted by the VDAN are supposed to be close the visual stuff present in the embedding space only if they are semantically similar. Our network is trained without the use of any ground truth segment labels.

SSFF and FFNet are the two competitors of our method. Those two are based on ground-truth segment labels. Though our method does not use any ground-truth segment labels it is still aware of the instructions segments. This proves to be the biggest advantage of our method.

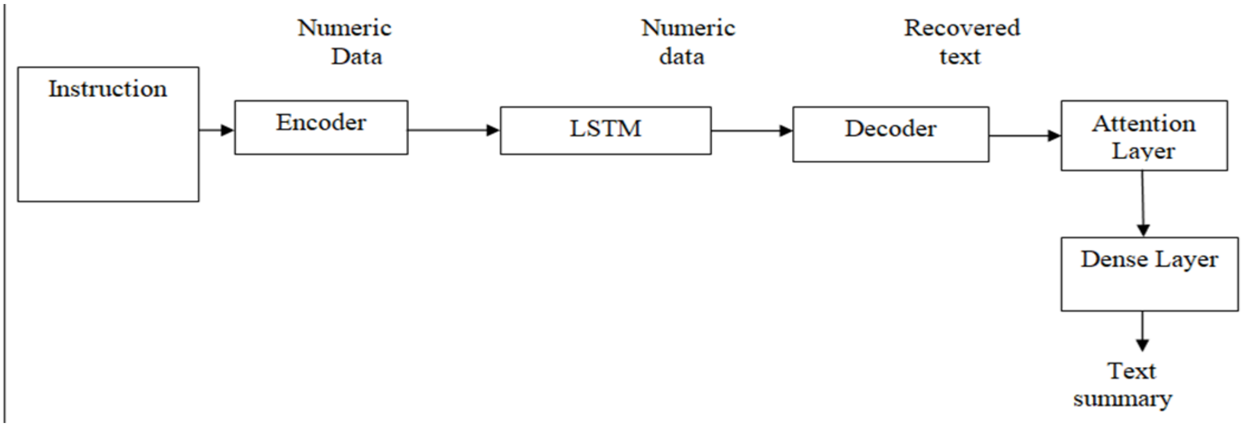


TEXT SUMMARIZATION

The next phase is the text summarization phase. In our project we have used the encoder decoder Recurrent Neural Network (RNN). The main reason to choose RNN instead of Natural Language Processing (NLP) is the short and inaccurate summaries that are generated by NLP technique. In case of RNN, the current output always depends on the previous output. This helps it in creating summaries that are more human like. We also have Long Short Term Memory

(LSTM) to provide some extended functionality for RNN. For implementing our neural network we have used Keras library of Python. After the data is loaded, contraction mapping is done to understand the contracted words of that particular language. Next is the data pre-processing step. In this step, data is cleaned by means of stopwords removal. The sentences are tokenized for further processing. The encoder in this phase is used for the conversion of words into numeric data. The decoder does the reverse process of encoder.

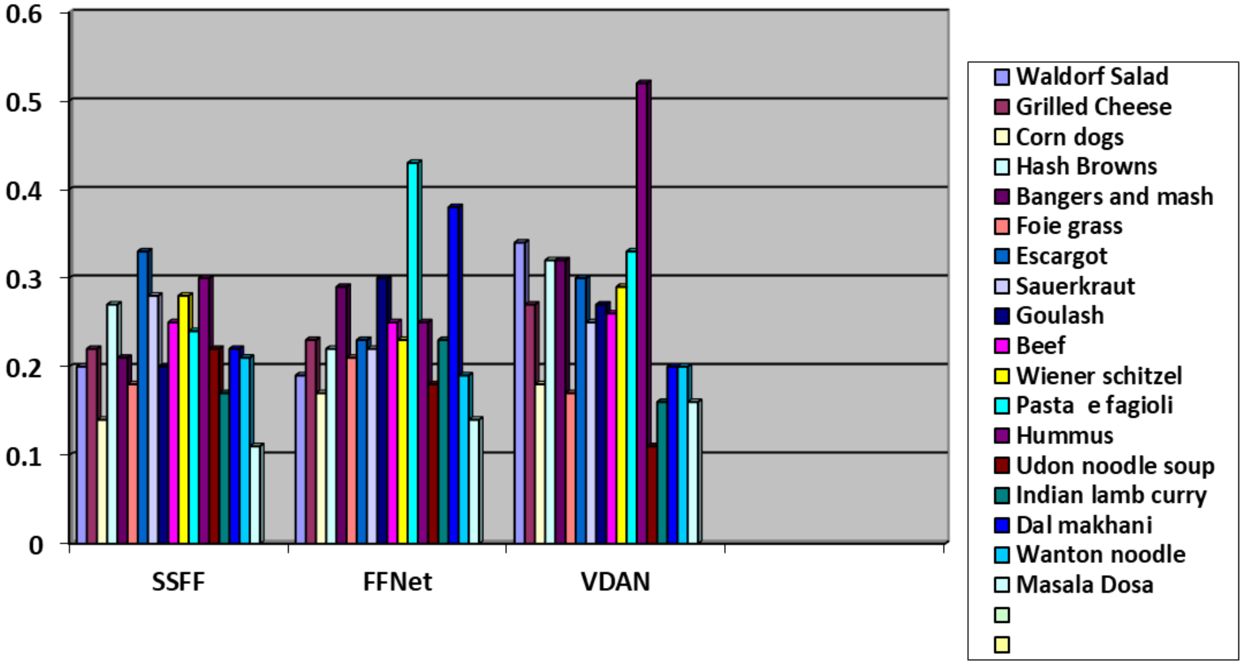
The attention layer is used for choosing only the useful information and discards the rest by means of cognitive mapping. The main use of dense layer is to represent the matrix vector multiplication that takes place in the neurons.



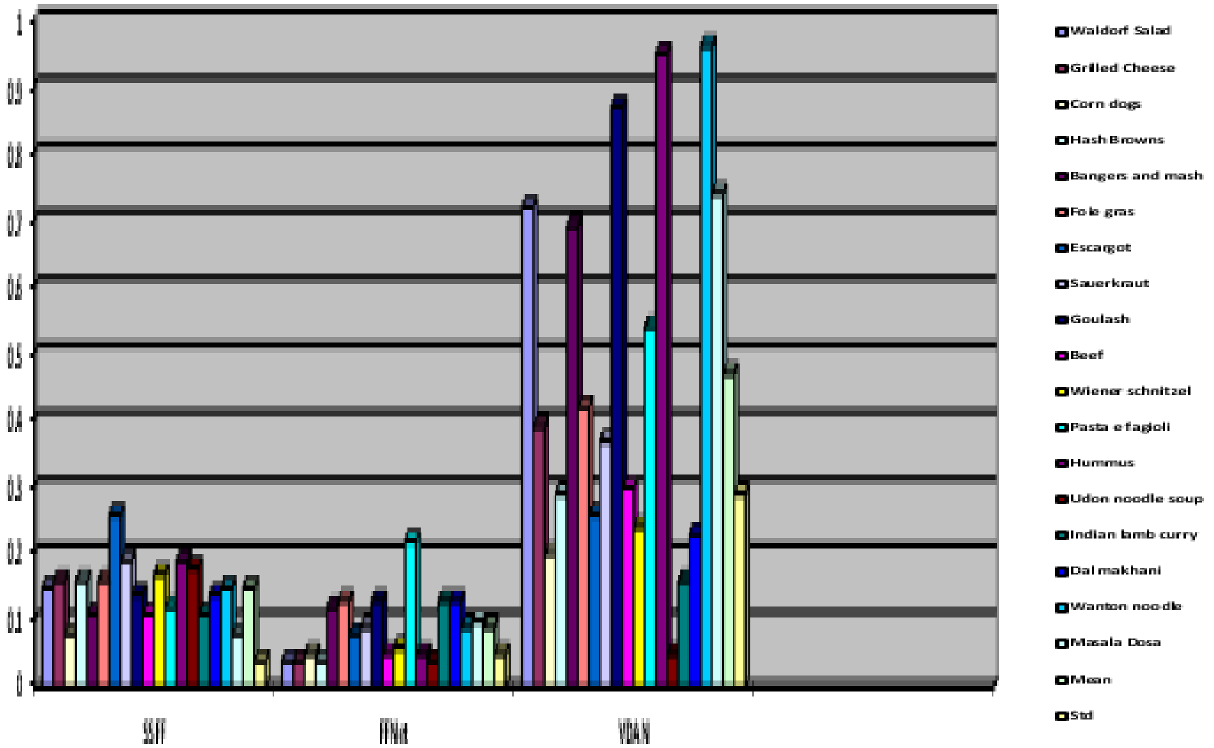
IV. RESULT

The following are the results of the video summarization on different datasets. The experiment was tested on all those datasets and results were noted.

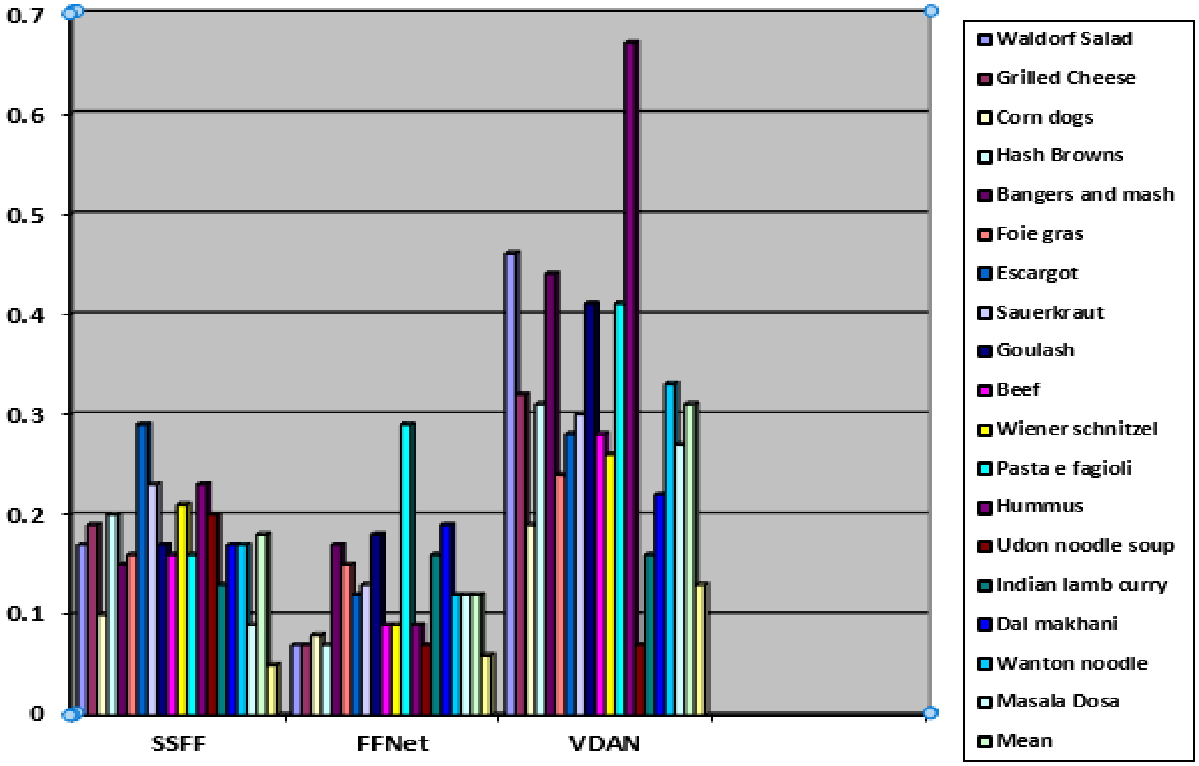
The results were tested on the basis of three parameters. They are precision, recall and F1 score. The first graph depicts the results on the basis of the precision parameter. Precision is a parameter that indicates the accuracy of the built model.



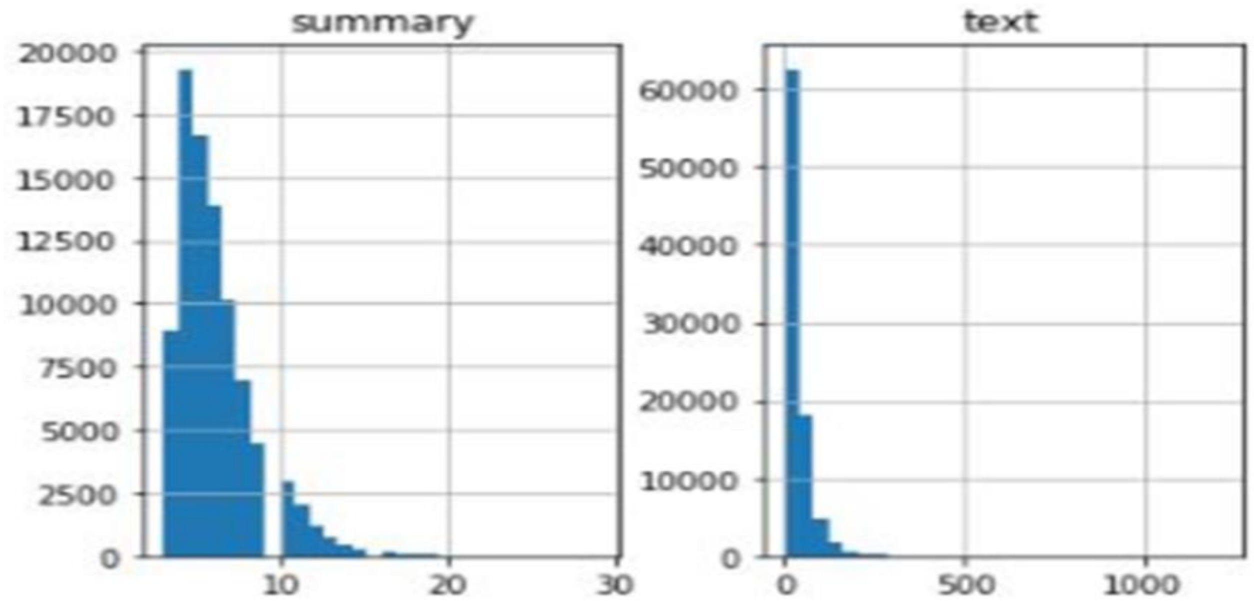
The next graph shows the results on the basis of the recall parameter. Recall is the number of actual positives that are captured by our model.



The next graph shows the results on the basis of F1 score.



The following graph depicts the result of text summarization.



IV. CONCLUSION

Thus a time saver has been introduced in the form of a summarizer. With the help of summarizing a quick idea could be derived without spending hours. This will not just save time but help in different aspects of life. Video summarizer with the text summarizer can prove to be boon in the forthcoming stages of life. People who hate reading can also get to know content by reading the summarized version of long paragraphs and videos instantly.

V. FUTURE ENHANCEMENT

Video summarization opens a wide branch in compact representation of the multimedia data. Based on the application different type of summary could be derived from the videos. This could be enhanced further by getting voice as input and performing the summarization. This would further ease the way of getting things known. This would particularly help in long business meeting where people are supposed to listen to every single line. This would prove to be a time saver

REFERENCES

[1] Amit bora, Shanu Sharma [2018],” A Review on Video Summarization Approcahes: Recent Advances and Directions”, International Conference on Advances in computing, communication control and networking (ICACCCN).

[2] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. [2017].” Diverse sequential subset selection for supervised video summarization”. In Advances in Neural Information Processing Systems.

[3] Bor-Chun Chen, Yan-Ying Chen, Francine Chen (2017), Joint Video Summarization and Captioning with Recurrent NeuralNetworks.

[4] Chandra Khatri, Gyanit Singh, Nish Parikh [2018],” Abstractive and Extractive Text Summarization using Document context vector and recurrent neural networks, IEEE paper.

[5] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I.Metsai, VasileiosMezaris, IoannisPatras[2021],” Video Summarization Using Deep Neural Networks “,IEEE paper.

[6] J.N.Madhuri, R.Ganeshkumar [2019] ,”Extractive text summarization using sentence ranking”, International conference on Data Science and Communication.

[7] Kaiyang Zhou, Yu Qiao, Tao Xiang [2018], Deep Reinforcement for Unsupervised Video Summarization with Diversity-Representativeness Reward.

[8] Madhu S.Nair , Jesna Mohan [2019],”Video Summarization using Convolutional Neural Network and Random Forest Classifier”, IEEE region 10 conference.

[9] Moratanch, chitrakalagopalan[2017],”A survey on Extractive text summarization”, IEEE

International Conference on Computer, Communication, and Signal Processing. Rameshnallapti, bowenzhou, Cicero dos santo, bingxiang, caglargulcehre[2017],"Abstractive text summarization using Sequence-to-Sequence RNNs", IEEE paper.

[10] Narendra andhale, L.A.Bewoor[2016], "An Overview of Text Summarization Techniques", International conference on computing communication control and automaion (ICCUBEA).

[11] N.Moratanch, S.Chithrakala [2017]," A Survey on Extractive Text Summarization", International conference on computer, Communication and signal processing.

[12] Pradeep choudhary, Sowmya P.Munukulta, K.S.Rajesh, Alok S.Shukla [2017], "Real Time Video Summarization on Mobile Platform",IEEE International Conference on Multimedia and Expo.

[13] Rockson Agyeman, RafiqMuhammad and GyuSang Choi (2019), Soccer Video Summarization using Deep Learning, IEEE Conference on Multimedia Information Processing and Retrieval,2019(MIPR).

[14] Shagan sha, sourabh kulhare,Allison gray, subashini venugopalan, Raymond Ptucha [2017],"SemanticText Summarization on long videos",IEEE winter conference on Applications of Computer Vision

[15] Tejero-de-Pablos, Y.Nakashima, T. Sato, N. Yokoya, M. Linna and E. Rahtu, "Summarization of User-Generated Sports Video by Using Deep Action Recognition Features," IEEE Transactions on Multimedia.

[16] VarunLuthra, JayantaBasak, Prof. Santanu Chaudhury and K.A.N.Jyothi(2018), A Machine Learning Approach to Video Summarization.

[17] Vinay Rajpoot, Sheetal Girase [2018], "A Study on Application Scenario of Video Summarization", Second International conference on Electronics, Communication and Aerospace Technology(ICECA).

[18] Washington Ramos, Michel Silva, Edson Araujo, Leandro Soriano Marcolino, Erickson Nascimento, [2020], Straight to the Point: Fast-forwarding Videos via Reinforcement Learning Using Textual Data, IEEE Conference.

[19] Yudong jiang, Kaixu Cui, Bo peng, changliang zu [2019]," Comprehensive Video Understanding: Video Summarization with Content-Based Video Recommender Design", International Conference on Computer vision workshop.