International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 10 | October - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

An Overview of Video Summarization

Nimish Rajgure, Rohan Thoke, Pratiksha Yadav,Sahil Mandore
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering
Pune, India

Prof. Pallavi Bhaskare
Department of Computer Engineering Smt.
Kashibai Navale College of Engineering
Pune, India

***Abstract*— Video summarization is an important task in the field of computer vision, with applications in many areas, such as video surveillance, video indexing, and video recommendation. It aims to generate a concise and informative summary of a video by selecting the most important frames or segments.Web-based applications for video summarization offer a number of advantages over traditional desktop-based applications. First, they are more accessible, as they can be used from any device with a web browser. Second, they are more scalable, as they can be deployed on cloud servers to handle large volumes of video data. Third, they are more collaborative, as they can be used to share and discuss video summaries with others.**

Keywords: Text summarization, model, video.

I. INTRODUCTION

This study compared the performance of four different RNN models (feedforward, simple recurrent neural network (SRNN), gated recurrent unit (GRU), and long short-term memory (LSTM)) in predicting daily stock prices of selected listed companies of the Colombo Stock Exchange (CSE). The models were trained on closing, high, and low prices of the past two days. The feedforward models had the highest forecasting accuracy, with errors of approximately 99%. The SRNN and LSTM models generally produced lower errors than the feedforward models, but in some cases, their errors were higher. The GRU models had the highest forecasting errors. Extractive summarization extracts important sentences or phrases from the source documents and group them to generate summary without changing the source text. However, abstractive summarization consists of understandingthe source text by using the linguistic method to interpretand examine the text. The abstractive summarization aims to produce a generalized summary, conveying information in a concise way. The system works by first converting the video to a sequence of images. Then, it uses a combination of optical character recognition (OCR) and natural language processing (NLP) techniques to extract the text from the images. The OCR technique recognizes the individual characters in the images, and the NLP technique identifies the words and sentences in the text. The proposed system has several potential applications, it can be used to create searchable transcripts of educational and news videos, or to extract text from videos for other application.

This paper presents few models and methods commonly used for video-to-text summarization: Section II video to text summarization models. Section III methods for converting audio into text. Section IV describes future trends of summarization application. Section V concludes the paper.

II. VIDEO TO TEXT SUMMARIZATION MODELS

Video to text summarization is a challenging task that involves generating a concise textual summary of the content of a video. There are various approaches and models used for this purpose, often involving a combination of computer vision and natural language processing techniques, Here are a few models and methods commonly used for video to text summarization:

A. Recurrent Neural Networks

CARNN is a new type of neural network that uses chaotic behavior to improve its performance in noisy pattern recognition. CARNN has two main components: chaotic nodes and an attractor recurrent neural network (ARNN). The chaotic nodes create various variability around the formed attractors, while the ARNN supervises the evolution of these nodes to find the appropriate findings. The chaotic behavior of neurons enables CARNN to search effectively in attractor basins, resulting in better performance than ARNN and FNN in noisy pattern recognition[20].

Recurrent neural networks can approximate any non-linear function, given certain conditions, such as the number of nodes in the hidden layer and the approximation effectiveness..In other words, recurrent neural networks are very powerful and can be used to model complex relationships between data. However, it is important to carefully design the network architecture and training process in order to achieve good performance[23]. This study compared the performance of four different RNN models (feedforward, simple recurrent neural network (SRNN), gated recurrent unit (GRU), and long short-term memory (LSTM)) in predicting daily stock prices of selected listed companies of the Colombo Stock Exchange (CSE). The models were trained on closing, high, and low prices of the past two days. The feedforward models had the highest forecasting accuracy, with errors of approximately 99%. The SRNN and LSTM models generally produced lower errors than the feedforward models, but in some cases, their errors were higher.It provides better way to extract knowledge and generate meaningful information. The GRU models had the highest forecasting errors[21].

B. Convolutional Neural Networks

CNNs (Convolutional Neural Networks) are a type of deep learning algorithm that excels at visual data tasks like image recognition, object detection, and image classification. CNNs can learn and extract hierarchical features from images, making them ideal for computer vision. CNNs have had a major impact on computer vision and have achieved state-of- the-art results in many image processing tasks. They remain a core technology in deep learning and AI. CNN are a powerful tool for computer vision tasks. They have revolutionized the field and enabled new applications that were not possible before.

This study proposes a hybrid approach using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks to improve the accuracy of Intrusion Detection Systems (IDS). The proposed approach is able to accurately classify malicious traffic according to attack types. The results are validated on NSL-KDD and CICIDS2017 datasets, achieving multiple classification accuracy of 98.1 and 96.7, respectively.The authors propose a new method for detecting and preventing network-based cyber attacks. They use a hybrid approach that combines two types of deep learning algorithms: CNNs and LSTMs. CNNs are good at extracting spatial features from data, while LSTMs are good at extracting temporal features from data. The authors' proposed approach is able to extract both spatial and temporal features from network traffic data, which allows it to more accurately detect and classify cyber attacks.The authors evaluated their proposed approach on two different datasets of network traffic data: NSL-KDD and CICIDS2017. The results show that their proposed approach achieves high accuracy in classifying both normal and malicious traffic. Additionally, their approach is able to accurately classify malicious traffic according to attacktypes[24].

C. Transformer Models

Transformer models have revolutionized the field of natural language processing (NLP) and have been successfully applied to a wide range of tasks beyond NLP. Some approaches use reinforcement learning to learn how to select and summarize the most important segments of a video.Models are trained to maximize a reward. They were introduced in the paper "Attention Is All You Need" by Vaswani et al. in 2017. Transformer models are known for their parallelization, scalability, and ability to capture long- range dependencies in data[15].

The authors verify the accuracy and validity of the STM using a 2.5 MVA and 6300/420 V three-phase transformer as a test object. They also discuss the use of the STM to analyze some important faults in transformers.The STM can be used to design and optimize transformers, to investigate the effect of faults on transformers, and to develop new diagnostic methods for transformers. The results of this paper can be used to develop new in-situ SERS sensors for the detection of furfural concentration in transformer oil. This could help to improve the reliability and efficiency of transformer maintenance and inspection which enables easy usage in text summarization[9].

This paper uses density functional theory to simulate the adsorption mode of furfural on a metal surface and its SERS characteristics. The simulation model of the furfural molecule and silver cluster is built, and the SERS signals from different vibration modes are calculated and identified. By comparing the theoretical calculation results with measured furfural spectra, the mode of furfural molecule adsorbed on metal surface is determined. These results can provide a theoretical basis for the in-situ SERS detection of furfural dissolved in transformer oil.

D. 3D CNNs

3D Convolutional Neural Networks (3D CNNs) are a type of neural network architecture designed for processing three-dimensional data, typically used in tasks involving video analysis, medical imaging, and other spatiotemporal data. While traditional 2D CNNs are used for image analysis, 3D CNNs are specialized for capturing spatial and temporal features in videos and volumetric data.

This paper proposes a novel 3D-CNN for video super- resolution that does not require motion alignment as preprocessing. The proposed 3DSRnet maintains the temporal depth of spatio-temporal feature maps to maximally capture the temporally nonlinear characteristics between low and high resolution frames, and adopts residual learning in conjunction with the sub-pixel outputs.3DSRnet first deals with the performance drop due to scene change, which is important in practice but has not been previously considered. Experimental results on the Vidset4 benchmark show that 3DSRnet outperforms the state-of-the-art method with average 0.45 dB and 0.36 dB higher in PSNR, for scale 3 and 4, respectively[11]. The proposed architecture includes optimizations such as locality exploration and block alignment with 3D blocking, which further improve performance and accuracy. The authors also develop an analytical model and tool to predict the optimized parameters for hardware customization based on user constraints[12].Experiments show that a 15-bit mantissa design using single-precision accumulation on a Xilinx ZC706 device can be 8.2 times faster than an Intel i7-950 processor at 3.07 GHz with only 0.4% accuracy loss. The results show that the transfer learning approach using 3D CNNs achieves impressive results on all metrics, with relatively short learning time. This suggests that the proposed approach is a promising approach for aggressive action recognition in video content[13].

E. Reinforcement Learning

Reinforcement Learning (RL) is a subfield of machine learning that deals with how agents should take actions in an environment in order to maximize a cumulative reward. It is inspired by behavioral psychology and is concerned with how software agents ought to take actions in an environment in order to maximize a reward Reinforcement learning is widely used for solving problems where the optimal action can be learned through trial and error. It has been successfully applied to a wide range of tasks, from training agents to play video games to controlling autonomous robots.

The findings of this paper could help to improve the performance and efficiency of DRL agents in a variety of applications. For example, curriculum learning could be used to train self-driving cars more quickly and safely. The authors also suggest that it may be possible to develop systems that allow machines to teach each other, which could lead to new and innovative ways to train AI models[15].Reinforcement learning (RL) is a type of machine learning that allows agents to learn how to behave in an environment by trial and error. RL agents are rewarded for taking actions that lead to desired outcomes and penalized for taking actions that lead to undesired outcomes. Over time, the agent learns to take the actions that maximize its expected reward[16].

F. Temporal Segment Networks

Temporal Segment Networks (TSN) is a deep learning architecture designed for action recognition in videos. It was introduced in the paper "Temporal Segment Networks for Action Recognition in Videos" by Limin Wang et al., which was presented at the European Conference on Computer Vision (ECCV) in 2016. TSN addresses the challenge of recognizing actions in videos by effectively modeling temporal information while maintaining the benefits of spatial features. Temporal Segment Networks have demonstrated significant improvements in action recognition tasks, especially when compared to traditional approaches that analyze individual frames or use long video clips.

The proposed spatio-temporal pattern recognition model is rooted in competitive network models from Amari (1978) and Arbib (1989). However, the author's model introduces distinct network structure and parameters, resulting in the temporally-continuous transformation of spatially-distributed patterns into temporally-segmented spatially-local representations. Notably, the model features an adjustable hysteresis, ensuring that the largest input consistently prevails, setting it apart from previous works in the field[4]. This paper proposes a non-parametric distribution based approach for event detection in sensor network data. The approach uses multiple sliding windows at different scales to obtain the distribution of the data. The temporal data stream is then segmented and potential event-bearing candidates are identified by comparing the present and past statistical behavior of the data. The paper also discusses the effect of optimum bandwidth selection on accuracy, the range of allowable window sizes, and computational speed[2].

G. Hybrid Model

Hybrid models refer to the combination of different machine learning or deep learning techniques, architectures, or components to address specific tasks or challenges. These models integrate multiple approaches to leverage the strengths of each component, aiming to improve overall performance or efficiency. Hybrid models are commonly used in various fields to tackle complex problems. Hybrid models offer the flexibility to address complex problems that may require diverse approaches. The choice of components and their integration depends on the specific problem, data, and desired outcomes.

He The established formalism for hybrid systems is the hybrid automaton. A hybrid automaton is a finite state machine that controls the switching between continuous modes. Each mode is described by a continuous state equation and a set of discrete events that can trigger a switch to another mode.The paper discusses the advantages and disadvantages of each description and the established formalism, using the bouncing ball example. It concludes that the established formalism, the hybrid automaton, is the most general and flexible way to model hybrid systems. However, it is also the most complex and difficult to use. This letter proposes a hybrid physical model-driven and data-driven approach for linearizing power flow models. The proposed approach combines the existing physical-equation-based linear power flow model with a linearized error model obtained using a partial least squares regression data-driven approach.The proposed hybrid linear model retains the useful inherent information from the physical model and utilizes the ability of data analysis to extract the inexplicit linear relationships. Simulations on four test systems have shown that the proposed hybrid linear model exhibits a much better performance on branch power flow calculation than other linear power flow models[27].

III. METHODS FOR CONVERTING AUDIO INTO TEXT

The best method for converting audio into text will depend on your needs and budget. If you need a quick and affordable transcript, an online ASR service or ASR software program may be a good option. If you need a highly accurate transcript, a human transcription service may be a better option.

A. Automatic Speech Recognition

Cochlear implant (CI) users have difficulty understanding speech in reverberant environments, such as noisy rooms. Automatic speech recognition (ASR) systems can recognize speech in reverberant environments better than CI users can. This study developed a hybrid recognition-synthesis CI strategy that uses an ASR system to recognize speech in reverberation and then synthesizes the speech waveforms so that they can be presented to CI listeners. The results showed that this hybrid strategy can significantly improve speech intelligibility for CI users in reverberant environments.[32].

This paper proposes a new method for automatic punctuation generation in speech-to-text transcriptions. The method uses a combination of acoustic and language models to predict the punctuation marks that should be inserted into a raw word sequence. The acoustic model uses prosodic features, such as pause duration, to identify potential punctuation points. The language model uses three components: a trigger-word model, a forward trigram punctuation predictor, and a backward trigram punctuation predictor. The trigger-word model identifies words that are likely to be followed by a punctuation mark. The forward and backward trigram punctuation predictors predict the next punctuation mark based on the previous two punctuation marks.The separation of the acoustic and language models allows them to be trained on different corpora. The results show that the method achieves an F- measure of 81.0% and 56.5% for voicemails and podcasts, respectively, on reference transcripts[33]. Automatic speech recognition is a better way to extract information from text and video which converts audio to text and text to summary.

This paper compares two techniques for training automatic speech recognition (ASR) systems on noisy speech with a technique for training ASR systems on clean speech. The techniques were compared using a speech recognition accuracy measure and 14 types of noise, including noise from household appliances and computers, street and transport, teaching rooms, and lobbies.The results showed that training on noisy speech allows ASR systems to achieve higher recognition accuracy in noisy environments. For example, an ASR system trained on noisy speech can achieve a 95% recognition accuracy with a minimum signal-to-noise ratio (SNR) of 10 dB, while an ASR system trained on clean speech can only achieve the same recognition accuracy with a minimum SNR of 20 dB.In other words, training on noisy speech makes ASR systems more robust to noise. This is an important finding because it means that ASR systems can be used in a wider range of environments, even in environments with a lot of noise[34].

B. Human transcription

Transcribing handwritten text can be sped up by using off- line Handwritten Text Recognition (HTR) techniques, which produce an initial draft transcription of a handwritten text image. However, this draft transcription usually contains errors that must be corrected by the transcriber by providing feedback.The usual approach is post-editing, where each error is corrected without modifying the rest of the current transcription. A more sophisticated approach can employ the current modification to provide a new whole transcription, hopefully with less errors.Apart from that, feedback can be provided in different modalities: keyboard input, online handwritten text, or speech. Each of these modalities has different advantages and disadvantages in terms of ambiguity, derived errors, and final transcription time.This study evaluates the transcription productivity of different feedback modalities in the assisted transcription of a historical handwritten text document in Spanish[35].

This paper describes a new system for transcribing audio and video from TV broadcast news. The system builds on an existing audio-only transcription system by adding the ability to process visual information, such as video recordings.The new system uses a combination of deep neural networks (DNNs) and hidden Markov models (HMMs) to recognize speech in the audio signal. It also uses convolutional neural networks (CNNs) to classify the visual signal.In addition, the system includes modules for detecting and identifying human faces, TV logos, and company logos in the video frames. It also has a module for optical character recognition (OCR) to extract text from the video frames.The system was tested on a large database of 817 hours of TV broadcast news, which was completely transcribed. The system also includes the ability to intelligently search the transcribed data based on both audio and visual signals[36].

Human transcriptionists listen to audio recordings and type them out manually. Human transcription is generally more accurate than ASR, but it is also more expensive and time- consuming. Human transcriptionists give you complete control over the transcription process. You can provide specific instructions on how you want the transcript to be formatted and what types of information you want to be included.

Digital archives are the best way to capture human experiences. But before we can use them effectively, we need to describe their contents. The MALACH project is funded by the National Science Foundation and aims to improve access to large spoken archives by advancing the state-of-the-art in automatic speech recognition (ASR), information retrieval (IR), and related technologies for multiple languages.

This paper describes the ASR research for the English speech in the MALACH corpus. The corpus contains unconstrained, natural speech with disfluencies, heavy accents, age-related coarticulations, uncued speaker and language switching, and emotional speech collected from over52,000 speakers in 32 languages.

The paper describes this new testbed for developing speech recognition algorithms and reports on the performance of well-known techniques for building better acoustic models for the speaking styles seen in this corpus. The best English ASR system to date has a word error rate of 43.8% on this corpus[37].

IV. FUTURE TRENDS OF SUMMARIZATION APPLICATION

Video summarization applications are still in their early stages of development, but they have the potential to revolutionize the way we interact with video content. Here are some of the future trends in video summarization applications:

- More accurate and informative summaries: As video summarization algorithms continue to improve, we can expect to see summaries that are more accurate and informative. This will make it easier for us to quickly and easily understand the content of a video.
- Personalized summaries: Video summarization applications will be able to learn our individual preferences and generate summaries that are tailored to our interests. This will make it easier for us to discover new and relevant video content.
- Interactive summaries: Video summarization applications will become more interactive, allowing us to drill down into specific aspects of a video or to skip ahead to the parts that are most interesting to us. This will make it easier for us to consume video content in a way that is efficient and effective.
- New applications: Video summarization applications will be used in new and innovative ways. For example, video summarization could be used to generate educational content, to create personalized video feeds, and to develop new forms of entertainment.

Video summarization applications have the potential to revolutionize the way we interact with video content. As video summarization algorithms continue to improve andnew applications are developed, we can expect to see

V. CONCLUSION

Video summarization application have the potential to revolutionize the way we interact with video content. As algorithms continue to improve and new applications are developed, we can expect to see video summarization application become an essential tool for students, researchers, business professionals, and consumers alike. One of the most

promising approaches to video summarization is to use deep learning. Deep learning models can be trained to learn the complex relationships between different aspects of video content, such as audio, video, and text. This allows them to generate more accurate and comprehensive summaries than traditional methods.

REFERENCES

[1] Wang, Y., Du, S., & Zhan, Y. (2008). Adaptive and Optimal Classification of Speech Emotion Recognition. 2008 Fourth International Conference on Natural Computation. doi:10.1109/icnc.2008.713
[2] Beigi, M., Chang, S.-F., Ebadollahi, S., & Verma, D. (2009). Muti-scale temporal segmentation and outlier detection in sensor networks. 2009 IEEE International Conference on Multimedia and Expo. doi:10.1109/icme.2009.5202496
[3] Fukuzawa, K., Komori, Y., Sawai, H., & Sugiyama, M. (1992). A segment-based speaker adaptation neural network applied to continuous speech recognition. [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. doi:10.1109/icassp.1992.225879
[4] Tanaka, T. (n.d.). Spatio-temporal pattern recognition by competitive networks. Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan). doi:10.1109/ijcnn.1993.716809
[5] Wang, S., Chen, Z., & Ding, Z. (2019). The Unified Object Detection Framework with Arbitrary Angle. 2019 5th International Conference on Big Data and Information Analytics (BigDIA).
[6] Dulani Meedeniya, Isuru Ariyarathne, Meelan Bandara, Roshinie Jayasundara, Charith Perera, "A Survey on Deep Learning Based Forest Environment Sound Classification at the Edge", ACM Computing Surveys, vol.56, no.3, pp.1, 2024.
[7] Pencea Maria Larisa;Ruxandra Tapu 2022 International Symposium on Electronics and Telecommunications (ISETC)
[8] Tatevik Ter-Hovhannisyan;Karen Avetisyan 2022 Ivannikov Memorial Workshop (IVMEM)
[9] Zhaoliang Gu;Mengzhao Zhu;Wenbing Zhu;Ran Xu;Jiabin Zhou;Jian Wang;Qingdong Zhu
[10] M -S. Chaouche;H. Houassine;S. Moulahoum;S. Bensaid;D. Trichet2019 19th International Symposium on Electromagnetic Fields in Mechatronics, Electrical and Electronic Engineering (ISEF)
[11] Soo Ye Kim;Jeongyeon Lim;Taeyoung Na;Munchurl Kim 2019 IEEE International Conference on Image Processing (ICIP)
[12] Hongxiang Fan;Ho-Cheung Ng;Shuanglong Liu;Zhiqiang Que;Xinyu Niu;Wayne Luk 2018 28th International Conference on Field Programmable Logic and Applications (FPL)
[13] Anton Saveliev;Mikhail Uzdiaev;Malov Dmitrii 2019 12th International Conference on Developments in eSystems Engineering (DeSE)
[14] Kartik Hegde;Rohit Agrawal;Yulun Yao;Christopher W Fletcher 2018 51st Annual IEEE/ACM International Symposium on Microarchitecture(MICRO)
[15] Martin Kaloev;Georgi Krastev 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)
[16] Le Lyu;Yang Shen;Sicheng Zhang 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)
[17] Weihua He;Yongyun Wu;Xiaohua Li 2021 IEEE 5th Information Technology,Networking,Electronic and Automation Control Conference(ITNEC)
[18] Chunyu Chen;Xinsheng Wu;An Chen 2020 International Symposium onAutonomous Systems (ISAS)
[19] Lu Dongdong;Tie Wenjie;Lei Songlin;Qiu Xiaolan 2021 CIE International Conference on Radar (Radar)
[20] Azarpour, M., Seyyedsalehi, S. A., & Taherkhani, A. (2010). Robust pattern recognition using chaotic dynamics in Attractor Recurrent Neural Network. The 2010 International Joint Conference on Neural Networks (IJCNN). doi:10.1109/ijcnn.2010.5596375
[21] Samarawickrama, A. J. P., & Fernando, T. G. I. (2017). A recurrent neural network approach in predicting daily stock prices an application to the Sri Lankan stock market. 2017 IEEE International Conference on Industrial and Information Systems (ICIIS). doi:10.1109/iciinfs.2017.8300345
[22] Chernigovskiy, A. S., Tsarev, R. Y., & Knyazkov, A. N. (2015). Hu's algorithm application for task scheduling in N-version software for satellite communications control systems. 2015 International Siberian Conference on Control and Communications (SIBCON).
[23] Cong, S., Yu, M., & Dai, Y. (2010). Approximation performance