DESIGN AND ANALYSIS OF VIDEO SUMMARIZATION APPROACHES USING ARTIFICIAL INTELLIGENCE TECHNIQUES

Thesis Submitted
In Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY
in
COMPUTER SCIENCE AND ENGINEERING

By

RUCHI GOEL
(2K18/PHDCO/18)

Under the Supervision of
DR. PRASHANT GIRIDHAR SHAMBHARKAR
Assistant Professor
Department of Computer Science and Engineering
Delhi Technological University

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

November, 2024

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Bawana Road-Delhi-42

CANDIDATE'S DECLARATION

I, **Ruchi Goel** hereby certify that the work which is being presented in the thesis entitled "**Design and Analysis of Video Summarization Approaches using Artificial Intelligence Techniques"** in partial fulfillment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the Department of **Computer Science and Engineering**, Delhi Technological University is an authentic record of my own work carried out during the period from **July 2018** to **November 2024** under the supervision of **Dr. Prashant Giridhar Shambharkar**, Assistant Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Bawana Road-Delhi-42

CERTIFICATE

Certified that **Ruchi Goel** (2K18/PHDCO/18) has carried out her search work presented in this thesis entitled "**Design and Analysis of Video Summarization Approaches using Artificial Intelligence Techniques"** for the award of for the award of the **Doctor of Philosophy** degree from Department of Computer Science and Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Dr. Prashant Giridhar Shambharkar

Assistant Professor (CSE)

Delhi Technological University

Place: Delhi, India

Date:

Design and Analysis of Video Summarization Approaches using Artificial Intelligence Techniques

ABSTRACT

The enormous increase of digital video data in today's fast-paced digital landscape has highlighted the importance of rapid and efficient data analysis and video retrieval approaches. Choosing a worthwhile video is not possible due to the vast amount of content. Time constraints also make it impossible to watch all videos through to the end. The process of condensing a video while maintaining its essential ideas and meaning using important keyframes, content and scenes is known as video summarization. The goal is to save time and effort by giving a succinct synopsis of the most crucial parts of the video rather than having to watch the whole video. Video summarization finds applications in various fields, such as surveillance, education, entertainment, and content browsing, enhancing accessibility and efficiency in video consumption.

The expansion in video content by the enhancement in multimedia technologies necessitates the development of novel browsing and understanding strategies. Design and Analysis of Video Summarization approaches using Artificial Intelligence Techniques is a thorough investigation that focuses on the creation and assessment of techniques for automatically producing succinct and informative video summaries. "Video summarization" is the process of distilling long video footage into shorter versions while keeping the most important details and pertinent information by removing redundant frames. "Artificial intelligence (AI) techniques" refer to a wide range of approaches and algorithms that make it possible for computers to carry out tasks that would typically require human intelligence. AI approaches are applied in the context of video summarization to automatically construct meaningful summaries, discover key aspects, and evaluate and comprehend the content of videos.

This research work contributes in providing video summary that makes video content easier to access by providing a more user-friendly entry point to complex video content. Multiple strategies are investigated to meet the various requirements of video summarization, such as key frame extraction and multimodal techniques. This helps a larger audience that has time or resources.
Firstly, efficiently discovering and navigating through a large number of videos is a big problem in smart cities, where security cameras add to the volume of video data and make efficient indexing and retrieval systems necessary. Video summarization appears as a critical solution that allows large-scale video collections to be stored, retrieved, and browsed while maintaining important aspects. With an emphasis on real-time applica-

tions, this research work offers a thorough overview of video summarizing approaches. After that analysis of real-time video is done using subtitles. Text summarizing techniques LSA and TextRank are used to assess the retrieved subtitles. The analyzed text is used to create a summary. This could be a brief text snippet emphasizing the important points, a list of key subjects presented, or even keywords that describe the video's content. The created summary is displayed alongside the video stream, allowing the user to follow along and understand the gist of the information as it unfolds.
Secondly, in a video, each frame represents a single point in time. However, several frames may include redundant information or slight differences. Keyframe extraction seeks to select a subset of these frames that best represent the visual content throughout the video. A method called TC-CLSTM Auto Encoder with mode-based learning is proposed for automatically selecting the keyframe, The autoencoder learns to recognize the most relevant elements in a video frame. These attributes can then be utilized to choose keyframes. Frames are rated based on extracted features and mode values, and the top-ranking frames are selected as keyframes. Thirdly, VEM a hybrid model is proposed for video summarization. this multimodal video summarization model presents a video summary using different multimedia modalities like text, audio and frames. To tackle text aspect subtitles are used. For audio component files are obtained in .wav format and from audio chunks From these audio chunks MFCC (Mel-frequency cepstral coefficients), Mel Spectrum, area Under the audio Curve, and audio peak after average cut-off were obtained and for the third aspect Mean Absolute Difference (MAD) is used to find important frame. Combining all aspects final summary is obtained. Another method TAVM using multimodal summarization is also presented. The BEiT vision transformer is used to identify items within the selected frame. For audio processing, speech-to-text converters are used to transcribe the audio content. Finally, in the final stage, the Summary Builder uses the GPT-3-based OpenAI API to build a summary of the information.

Lastly, Artificial intelligence (AI)-driven methods incorporating human presence detection and face identification lead to automatic summarization utilizing text and audio cues. The suggested framework intends to improve the efficacy and efficiency of video summarization by synthesizing various approaches, enabling quick understanding and retrieval of pertinent content amid the torrent of video data in the digital world.

Video summary provides timely insights to individuals in a variety of areas, including market research, surveillance, and media monitoring. This helps them to identify trends, anomalies, or crucial occurrences effectively. Additionally, by reducing the need to store, process, and transmit massive amounts of video data, these strategies aid in resource optimization. Video summarizing reduces costs and increases the effectiveness of systems that handle this type of data by condensing videos into brief representations. By utilizing different AI techniques, this research work seeks to produce different meth-

ods of video summarization. The study advances our knowledge of AI-driven methods for video summarization and offers suggestions for potential areas of research and useful applications to improve the administration and exploitation of video data.

ACKNOWLEDGEMENT

With great pleasure and a deep sense of gratitude, I would like to express my sincere thanks to my supervisor, **Dr. Prashant Giridhar Shambharkar**, Assistant Professor, CSE Department, Delhi Technological University. Without his constant motivation and encouragement, this research would not have been successful. He has always been my pillar of strength.

I would like to take this opportunity to thank **Prof. Vinod Kumar**, the HOD (Department of CSE), and **Prof. Rahul Katarya** Chairperson DRC (Department of CSE) for all their help, motivation, support, as well as for extending all the necessary processing and experimental facilities during my research work. I am grateful to **Prof. Rajni Jindal** Professor (Department of CSE) for her assistance, advice and guidance in making my aspirations come true. I would like to thank all faculty members and staff of the Computer Science and Engineering Department, Delhi Technological University for their encouragement and support.

I am extremely thankful to the Management of Maharaja Agrasen Institute of Technology, Rohini, Delhi, for permitting me to carry out my doctoral research work. I am also thankful to **Prof. Namita Gupta**, the HoD (Department of CSE) for all the support and encouragement.

I would like to extend my heartfelt gratitude to my dear friend and sister, **Dr. Pooja Gupta**, for her invaluable assistance, support, and encouragement. I would also like to express my sincere thanks to my friends **Ms Surbhi Upadhayay, Dr Garima Sharma, Dr Piu Jain, Mr. Nikhil Sharma** for their unwavering help and encouragement.

I would like to extend my profound sense of gratitude to my father **Sh. M.P Kuchhal**, mother **Smt. Santosh Kuchhal**, father-in-law **Sh. R.N Goel**, mother-in-law **Smt. Shanti Goel** and all family members for all the sacrifices they made during my research and also for providing me with moral support and encouragement whenever

required.

Last but not least, I would like to thank my husband, **Mr. Anuj Goel**, my son, **Aarav Goel**, for their constant encouragement and unwavering support, along with patience and understanding throughout my research.

I sincerely thank all who are not listed here but have been instrumental in making my journey a fulfilling experience.

Ms. Ruchi Goel
Department of Computer Science and Engineering
Delhi Technological University
Delhi-110042

LIST OF PUBLICATIONS

Publication in SCI/SCIE Journals:

1. Shambharkar, Prashant Giridhar, and Ruchi Goel. "Auto encoder with mode-based learning for keyframe extraction in video summarization." *Expert Systems* 40, no. 10 (2023): e13437.

2. Shambharkar, Prashant Giridhar, and Ruchi Goel. "VSEM: A Hybrid Model for Video Summarization." *Journal of Information Science and Engineering* 40, no. 6 (2024): 1253-1271.

Publication under review in SCI/SCIE Journals:

1. Shambharkar, Prashant Giridhar, and Ruchi Goel, "Keyframe Extraction via Peak Wave Analysis with Integrated Human Presence and Face Recognition, *Journal of Visual Communication and Image Representation, Elsevier*(2024).

Publication in ESCI Journal:

1. Shambharkar, Prashant Giridhar, and Ruchi Goel. "From video summarization to real time video summarization in smart cities and beyond: A survey." *Frontiers in big Data 5* (2023): 1106776

Publication in International Conferences:

1. Shambharkar, Prashant Giridhar, and Ruchi Goel. "Analysis of Real Time Video Summarization using Subtitles." *In 2021 International Conference on Industrial Electronics Research and Applications (ICIERA)*, (pp. 1-4). IEEE, 2021

2. Shambharkar, Prashant Giridhar, and Ruchi Goel "TAVM: A Novel Video Summarization Model Based on Text, Audio and Video Frames." *In 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp. 878-882. IEEE, 2023.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF TERMS AND ABBREVATIONS

NLP	Natural Language Processing
SVS	Single Video Summarization
MVS	Multiple Video Summarization
AVS	Automatic Video Summarization
DCT	Discrete Cosine Transform
AI	Artificial Intelligence
SVW	Sports Video in the Wild
OPF	Optimum Path Forest
VIRAT	Video Image Retrieval and Analysis Tool
AVOA	African Vulture Optimization Algorithm
QSAN	Query based Self Attentive Network
QAVOL	Query based deep African Vulture Leading
LSTM	Long Short Term Memory
Bi-LSTM	Bidirectional Long Short-Term Memory
GAT	Graph Attention Network
CFT	Contextual Features based Transformation
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
GAN	Generative Adversial Network
VAE	Variational AutoEncoder
IoT	Internet of Things
LSA	Latent Semantic Analysis
MFCC	Mel-Frequency Cepstral coefficients
FPS	Frames Per Second
MAD	Mean Absolute Difference
SVD	Singular Value Decomposition
TAVM	Text, Audio, and Video Mode
HPD	Human Presence Detection

FPD	Face Presence Detection
PWT	Peak Wave Time analysis
BEit	Bert Pretraining of Image Transformers
SOM	Self Organising Maps
VIRAT	Video Image Retrieval and Analysis Tool

CHAPTER 1

INTRODUCTION

1.1 Introduction

Video is the most challenging multimedia, as it incorporates all other media data (including text, pictures, graphics, and audio) into a single data stream. It is difficult to access video effectively due to its unstructured format and changing format length (1). Video information is a sequential data type that gives unlimited data through its moving content (2). Every minute, a large number of videos are uploaded to YouTube, IMDB, tourism sites, Flickr, and other video-sharing sites. Millennial video cameras are installed in public spaces, public transportation, banks, airfields, and other locations, resulting in a tremendous amount of data that is difficult to analyze in real-time. Social media stands as one of the most widely utilized applications globally, which allows users to communicate with each other and share media like pictures and movies. When a person is planning a vacation or wants to learn about a new subject, the first thing he does is look for available videos. There will be hundreds of suggestions for each search topic; navigating through these lengthy videos to find the essential video takes time, and also difficult to efficiently obtain this much data in such a short amount of time. To make effective use of video data, it must be accessible in a user-friendly manner. The challenge of analyzing video content to extract valuable or intriguing information is difficult and time-consuming. For this reason, it is critical to provide users with a brief video depiction of video material that allows them to get a sense of what the video is about without having to watch it in its entirety, allowing them to decide whether to watch the complete video. To address these concerns, work is underway to construct a video summary that summarizes the entire video in a short period of time. This is the purpose of a rapidly developing study field called video summarization (3). The generated video summaries may vary depending on the application, and the same video may have multiple summaries depending on the user's or application domain's requirements (4). The concept of video summarization is to make exploring a huge collection of video data faster and more efficient, as well as to achieve efficient access and representation of video content (5).

1.2 Need of Video Summarization

Besides time efficiency, video summaries are helpful in many ways:
• **Content promotion**: A video editor and marketer may use summaries as teasers or promotional material to attract viewers to watch the full video.
• **Learning aid**: In educational settings, video summaries can help students to review complex topics or lectures more efficiently.
• **Content accessibility**: Summaries make content more accessible to individuals with disabilities or those who prefer shorter, more concise versions.
• **Recap and memory aid**: After watching a lengthy video, a summary can serve as a recap to reinforce key takeaways and aid in the retention of information.
• **Content curation**: In video-sharing platforms and social media, users may curate video summaries to create playlists or compilations of related content.

1.3 Video Summarization

A video abstract, also known as video summarization, is a condensed version of the video's content. The five main foundations for accessing video content: multimodal analysis, video representation, summarization, browsing, and retrieval, have all advanced significantly in recent years (6). Video representation is concerned with the video structure. Summarising unscripted information (such as surveillance camera video) necessitates a "highlights" extraction structure that only catches the summary's most noteworthy events. whereas scripted information (such as news, and movie) uses a key-frame representation for each of the shots in a story.
Video representation is a crucial step in summarization for maintaining visual coherence, which in turn affects the overall quality of a summary. It consists of two main steps, namely, temporal segmentation, and feature representation.
A good video summary strikes a balance between the amount of information retained from the original video and the length of the resulting summary. A video summary must meet several criteria to be useful. Failure to follow these standards may have a significant impact on how a video event is understood (7). Video summarization helps us to quickly review lengthy videos by removing pointless and unnecessary frames. The goal of video summarization is to extract the most significant and instructive segments from the full-length video in order to create a comprehensive and succinct description (4). The produced summary is a collection of representative video frames (also known as video key-frames) or video fragments (also known as video key-fragments) which are stitched together chronologically to create a shorter video comprised of the generated summary

1.3.1 Steps of Video Summarization

To identify which parts of videos are to be removed, video summarization algorithms must rely on video content. There are three steps to video summarization (8) as indicated in Fig. 1.1.

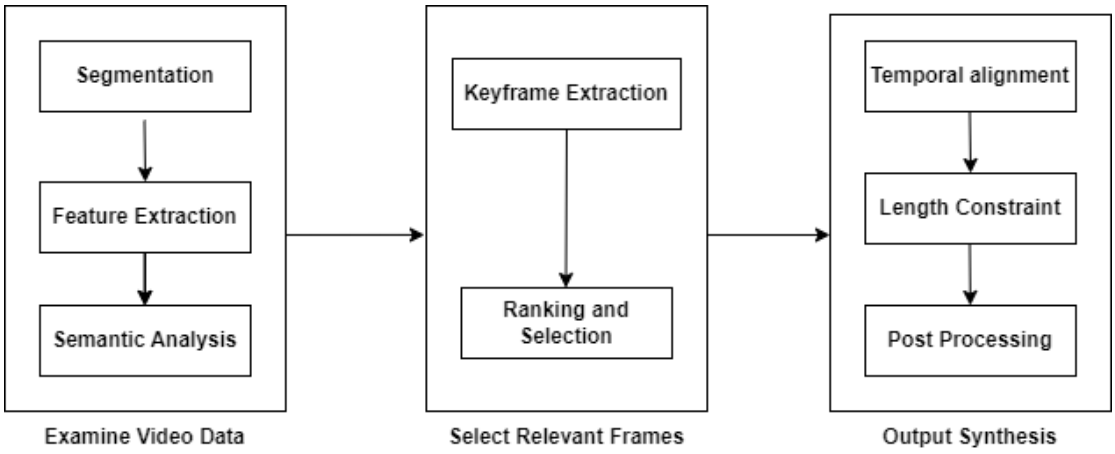


Figure 1.1: Steps to Video Summarization

1. **Examine Video Data:** The first step is to examine video data to determine the most important features, structures, or processes within the components visual, audio, and textual (audio and textual component if exists). It further consists of three parts.

   i Segmentation: The video is divided into its components called shots or segments. A shot is a series of uninterrupted frames constantly recorded by a single camera.

   ii Feature Extraction: Significant features are extracted from the video. Visual features like color, motion, and object recognition, as well as auditory aspects like sound analysis and speech recognition, can be included in this.

   iii Semantic Analysis: This stage involves analyzing each segment's content to determine its semantic meaning. It consists of two important parts: object and activity recognition and detection. The content of the video is comprehensively comprehended by object detection and recognition, which provide a snapshot of static features, and activity and action recognition, which capture dynamic aspects, resulting in an extensive overview of the video's main aspects.

2. **Select Relevant Frames:** This phase is to select relevant frames that represent the video's content.

i keyframe Extraction: This phase selects a representative frame that best expresses the essence of the segment and after that ranking and selection of keyframes are done. Appropriate frames are chosen by ranking the keyframes according to their significance, relevance, or aesthetic qualities.

3. **Output synthesis:** The last phase is output synthesis, which involves assembling the frames/shots into the original video.

   i Temporal Alignment: The coherent sequencing of significant frames or segments, the reflection of the chronological order and flow of events, and the preservation of temporal context and continuity are all dependent upon temporal alignment, which is an essential component of video summarization.

   ii Length Constraint: A major factor in deciding the summary's ultimate format and runtime is its length restriction. The length of the summary depends on the user's requirements. This limitation is put in place to make sure that the generated summary successfully communicates the key points of the original video while staying manageable and accurate.

   iii Post Processing: Post-processing is a crucial step in the output synthesis phase of video summarization which involves modifying and fine-tune the generated summary to improve its quality, coherence, and usability. This stage concentrates on refining the summary to make it more effective for its intended purpose and usually comes after the initial content selection and organization processes.

1.4 Video Summarization Techniques

The criteria for relevant frame selection (score prediction and keyframe selection) may differ for different users and application domains, a general framework for a video summarizing will not work for everyone. There are many approaches to video summarization static, dynamic, hierarchical, multi-view, image, and text summaries (9) as indicated in Table 1.

1.4.1 Summary Based

Summary based video summarization can be further classified into static, dynamic, hierarchical, multi-view, image, and text summaries (9).

1.4.1.1 Static Summary

Static summary is also known as keyframing or storyboard presentation. It's a montage of keyframes taken from the authentic video. A static summary is more suitable for

Approach Type	**SubTypes**	
Summary Based	1. Static	
	2. Dynamic	
	3. Hierarchical	
	4. Multi-view	
	5. Image	
	6. Text	
Preference based	1. Domain-specific	
	2. Query Based	2.1 Generic
		2.2 Query Focused
	3. Semantic Based	
	4. Event Based	
	5. Feature Based	
Domain Based	1. Pixel	
	2. Compressed	
Information Source Based	1. Internal	
	2. External	
	3. Hybrid	
Time Based	1. Real Time	
	2. Static	
	1. Supervised	
Training Strategy Based	2. Unsupervised	
	3. Semisupervised	

Table 1.1: Video Summarization Techniques

indexing, browsing, and retrieval (10). To assess video static summaries, Avila et al.(9) used a two-stream deep learning architecture that combined the k-means clustering algorithm with color information collected from video frames.

1.4.1.2 Dynamic Summary

It is also known as video skimming. Short dynamic sections (video skims) or audio and the most relevant video shots are chosen. The goal of video skimming techniques is to choose pictures or scenes from the full video and compile them into a relevant summary. Zhong et al. (11) offer a novel dynamic video summarization approach.
In static summaries, the motion component is lost. However, the technology makes video storage and retrieval easier, particularly in large video repositories. Storyboard layouts lack audio cues and may lack continuity, yet they are efficient in terms of calculation time and memory.

1.4.1.3 Hierarchical Summary

It represents a scalable and multi-level summary. It has several levels of abstraction, with the highest level having the fewest keyframes and the lowest level having the

most. V-unit (12) model for shot detection is used to structure videos following the hierarchical model to remove junk frames.

1.4.1.4 MVS (MultiView Summary)

MVS considers multiple points of view to create an informative summary. Smart IoT-based architecture with an embedded vision for detecting incredulous articles, and exchanging traffic volume statistics are proposed by the author (13).

1.4.1.5 Image summary

A single image or a collection of images is typically used for this type of summary. Images serve as synopsis rather than frames or shots. Authors (14) presented a paradigm that extends the single-image captioning transformer-based architecture to multi-image captioning.

1.4.1.6 Text Summary

These are summaries that consist solely of a paragraph-length textual summary of a video sequence. It is created utilizing Natural Language Processing (NLP) techniques and does not include any audio or visual descriptions. Text summaries are cost-effective in terms of storage and calculation, but they are unable to communicate all of the information since they lack the audio and visual components of the video sequence.

1.4.2 Preference Based

It is broadly divided into 5 categories listed video summarization is domain-specific, query-based, semantically based, event-based, and feature-based

1.4.2.1 Domain-Specific

Domain-specific video summarizing is a potential strategy for creating informative and targeted summaries that meet the specific needs of various areas. This technique has the potential to greatly improve information availability and comprehension across multiple disciplines by exploiting domain knowledge and personalizing the summarizing process. Kaushal et al. (15) provide a summary based on what is relevant to that domain, as well as other desirable domain features like representativeness, coverage, and diversity, for a given video of that domain.

1.4.2.2 Query-focused

It aims to create a diversified selection of video frames or segments that are both connected to the query and contain the original video data. While customizing video sum-

marizers appears to be a promising direction (16). It can be divided into two categories (17).

1. Generic
2. Query-focused

In the case of Generic video summarization, when a substantial scene transition occurs in a video, a broad summary is constructed by choosing keyframes. The keyframes are extracted when the cluster number for a frame changes. The visual components of the video are extracted using a pre-trained Convolutional Neural Network (CNN) which is then clustered using k-means, and a sequential keyframe-generating procedure follows. Query-focused summarization enables users to enter a search query or question about the video, and the summary is prepared to specifically answer that query.

1.4.2.3 Semantic-based

These are summaries that are generated based on the video's content and are mostly based on objects, actions, activities, and events, with a high level of interpretation based on domain expertise (18) (19).

1.4.2.4 Event-based

The objective is to develop and maintain succinct and coherent summaries that describe the current state of an event. Event-based video summarization is preferred over key frame-based summarization for surveillance video summarization. Many different applications use intelligent video surveillance systems to track events and conduct activity analysis (20).

1.4.2.5 Feature-Based

Features like motion, color, dynamic contents, gestures, audio-visual, voice transcripts, objects, and many others are used to classify feature-based video summarization techniques. Apostolidis et al. (21) used relevant literature on deep-learning-based video summarization and covered protocol aspects for summary evaluation.

1.4.3 Domain Based

It tailors the summary method to the video's domain or genre. It can be divided into pixels and compressed:

1.4.3.1 Pixel domain video summarization

It works by collecting information from the pixels of the frames in a video to summarize it. In most applications, a video is compressed, and decoding the video to summarize it takes a lot of time and space.

1.4.3.2 Compressed domain video summarization

It includes extracting features from compressed video by partially decoding. Fei et al.(22) devised a method for compressing a shot's most significant activities into a single keyframe in a compressed video stream in the compressed domain. It can provide a brief and colorful summary of video information. The original footage is represented by many keyframes created from one rich keyframe from each shot. Phadikar et al. (18) proposed a DCT (Discrete Cosine Transform) compressed domain image retrieval scheme. A feature set was created using edge histograms, color histograms, and moments. The best feature vector is then constructed using GA (Genetic analysis).

1.4.4 Information Source Based

Information source-based video summarizing is a technique that uses information sources other than the video to provide more thorough and insightful summaries. It is further classified as internal, external, or hybrid. At various phases of the video life cycle, a video summarizing algorithm evaluates a range of information sources to abstract the semantics associated with a video stream's content and then extract the various audio-visual cues. Based on the information sources they examine, the various methodologies reported in the literature can be divided into three groups (23):

1.4.4.1 Internal

Examine internal data extracted from the video stream generated during the video life cycle's production step. These methods extract semiotics from a video stream's image, audio, and text at low-level data for use in a video summary.

1.4.4.2 External

External information is any data source other than the video footage itself that can be used to supplement the comprehension and summary process. To look at data that isn't generated right away from the video stream, external summarization approaches are used. External information can be divided into two types: Contextual (not directly from a user's point of view) and User-based information (derived from a user's direct input). For contextual Hussein et al. (24) presented a video graph that is used to simulate

the long-term temporal structure of human activities. The semantic gap that internal summary approaches confront can be solved using external summarization techniques.

1.4.4.3 Hybrid

During any point of the video lifecycle, hybrid summarization algorithms examine both the movie's internal and external data. Hybrid summarization algorithms can leverage the semantics of the text to a greater extent, resulting in higher levels of semantic abstraction. This method is very well suited to summarizing domain-specific data. Kanehira et al. (25) devised a broad video summarizing approach that seeks to estimate the underlying perspective by taking video-level similarity into account, which is supposed to be obtained from the related viewpoint.

1.4.5 Time based Video Summarization

Time-based video summarizing selects and condenses video content based on temporal parameters, resulting in a succinct version that maintains the most important information or events. Depending on whether or not it is done on a live video, summarization can be classified as real-time or static, or on a video recording, respectively.

1.4.5.1 Real-time

In these circumstances, selecting crucial frames while the video is being captured depending on the context of the video, will be quite valuable. It's challenging, to sum up a video in real time because the output needs to be supplied quickly. In real-time systems, a late output is a bad output. This strategy is critical in situations requiring quick feedback or decision-making, such as live broadcasting, surveillance, and video conferencing.

1.4.5.2 Static based

A frame from the unified collection of frames collected from the source video is used to show the input video in a static summary (26). The most crucial elements of the original video are included in keyframes, which are a subset of frames. Static video summarization entails producing summaries after the complete video has been recorded. This method is utilized when immediate summarizing is not required.

1.4.6 Training strategy Based

Due to insufficient feature extraction and model selection, machine learning-based approaches can occasionally result in poor video summary quality. For example, a model with too few features may be inaccurate, whereas a model with too many features may

be overfitted (27). The following are some broad categories for a deep-learning-based video summarizing algorithms: Supervised approaches, Unsupervised approaches, and Semi-supervised approaches (21). The summary should keep keyframes from the original video. The same frames may be important for some at the same time and uninteresting for another viewer, thus, making a video summary a highly subjective word (28).

1.4.6.1 Supervised

These approaches are used to train a model using labelled data before generating video summaries. Deep neural networks have recently been used in video summarization. The temporal information is extracted using recurrent neural networks (29). For each video, these supervised approaches necessitate a huge number of frame- or shot-level labels. As a result, gathering many annotated films is expensive.

1.4.6.2 Unsupervised

There are no labeled data samples available in an unsupervised approach, so the frames are classified into several categories based on content similarity. Fajtl et al. (30) proposed a new soft attention-based simple network for sequence-to-sequence transformation, which is more efficient and less difficult than the current Bi-LSTM-based encoder-decoder network with soft attention. In an unsupervised manner, a deep summarizer network is used to reduce the distance between training films and the distribution of their summarizations. A summarizer like this can then be used to estimate the best synopsis for a new video (31).

1.4.6.3 Semi-supervised approach

This contains both labeled and unlabeled data. This mixture will often have a small bit of labelled data and a significant amount of unlabeled data.

Different video summarization techniques are available, so developers and consumers can choose the most suited one for their specific needs.

1.5 Video Summarization Framework

Selecting the right video summarization model depends on several factors (32):

i Who is the target audience?
Generic: Summarization caters to a broad audience without specific preferences.
Personalized: Summarization tailors to individual users based on their interests or viewing history.

ii What format is the desired output?

Text: Summarization generates a written summary of the video content.

Audio: Summarization creates an audio summary, potentially using spoken language or sound effects.

Video: Summarization produces a condensed video capturing key moments from the original.

iii What type of summary is needed?

Static: Summarization involves a set of representative images (key-frames) that showcase the video's content.

Dynamic: Summarization utilizes short video segments (key-segments) stitched together to represent the video's essence.

iv How many videos are involved?

Single Video Summarization (SVS): Summarization focuses on a single video.

Multiple Video Summarization (MVS): Summarization analyzes and summarizes multiple videos.

v How are the factors determining the summary derived?

Internal: Summarization relies solely on the content of the video itself.

External: Summarization incorporates additional information like captions, titles, or user preferences.

1.6 Applications of Video Summarization

Video summarization has become a game-changer across numerous fields, offering a powerful tool to condense and extract valuable information from vast amounts of video data. Here's a glimpse into its diverse applications:

- **Enhancing Security and Surveillance:**

  Security personnel can quickly scan through hours of footage, identifying suspicious activities or events with ease thanks to concise video summaries.

- **Revolutionizing Content Exploration:**

  Video platforms and databases leverage video summarization, allowing users to efficiently browse extensive video collections and preview their content without committing to full viewing.

- **Shaping the Way News and Media Consumed:**

  News outlets utilize video summarization to create bite-sized highlights of key events, keeping viewers informed on the go. Similarly, broadcasters can summarize sporting events or live broadcasts for viewers who missed the original airing.

- **Transforming Education and Learning:**

  Students and educators can leverage video summarization to navigate lengthy educational videos, gaining quick access to key concepts, lectures, or demonstrations.

- **Streamlining Legal and Forensic Processes:**

  Lawyers and investigators can efficiently review video evidence, such as security footage or body camera recordings, by utilizing video summaries to pinpoint critical moments or relevant details.

- **Empowering Medical Professionals:**

  In the medical field, video summarization assists in analyzing medical imaging sequences or surgical videos, allowing healthcare professionals to review procedures and monitor patient progress more efficiently.

- **Automating Video Editing:**

  Content creators and filmmakers can leverage video summarization to automate the editing process, identifying highlights and key moments for trailers, promotional videos, or social media content.

- **Preserving Personal Memories:**

  Individuals can use video summarization tools to create condensed summaries of their personal videos, capturing significant events or experiences from wearable cameras or other sources.

- **Real-Time Event Detection and Monitoring:**

  Video summarization algorithms can be deployed for real-time event detection and monitoring in various applications, including traffic surveillance, crowd management, and environmental monitoring.

- **Enhancing Human-Computer Interaction:**

  In human-computer interaction research, video summarization helps analyze user behaviour by summarizing interactions with user interfaces or digital environments, facilitating usability studies and interface design improvements.

In conclusion, video summarization plays a crucial role in managing and extracting actionable insights from the ever-growing volume of video data. This technology enhances efficiency, empowers informed decision-making, and revolutionizes how we interact with and utilize video content across various domains.

1.7 Motivation

Video summarization is the process of condensing and compacting a video, or creating an abstract or summary of the video. The main motive for doing video summarization research is time conservation. Watching a complete video of more than 1 minute would take a significant amount of time and effort, which is unacceptable. Additionally, summarized videos demand less bandwidth and storage, which makes them ideal for situations where network bandwidth or storage are limited. All things considered, video summarization is an essential tool for effective content consumption. As the amount of video content increases, it becomes more difficult to browse and retrieve it. As a result, effective methods for video summarization have to be developed to automatically reduce long videos into shorter ones .

Artificial intelligence techniques provide significant advantages in video summarization by adeptly managing the inherent uncertainty, imprecision, and complexity present in video data. Moreover, these techniques also integrate multiple modalities of information like visual, audio, and textual data, thereby enabling more comprehensive video summarization. Artificial intelligence techniques generate summaries that capture the richness in several aspects of the underlying video footage by efficiently combining information from many sources. Video Summarization embrace life long learning as it opens doors to diverse learning experiences, enabling us to explore a wider range of topics and perspectives without feeling overwhelmed by time.

1.8 Problem Statement

With the rapid growth of user-generated videos, being able to browse them effectively is becoming increasingly vital. Video summarization is seen to be a potential method for effectively realizing video content by identifying and selecting descriptive frames from the video. An automatic video summary would be advantageous for everyone who wants to save time and learn more in less time as video content continues to grow at a rapid rate. Currently, most techniques primarily rely on visual features. However, incorporating audio analysis (e.g., speech recognition, sentiment analysis) and text analysis (e.g., captions, subtitles) can provide a richer understanding of the video's narrative and context.

Secondly, Understanding the temporal evolution of information within a video is crucial for capturing important events and their progression.

Real-time applications require efficient processing to ensure summaries are generated without significant delays. This involves designing lightweight algorithms that minimize computational resources without sacrificing accuracy.

Based on this, the problem statement is defined as follows:

"Can the summaries generated by video classification approaches be automated using artificial intelligence techniques to generate time and memory-efficient summaries?" While artificial intelligence techniques have demonstrated encouraging outcomes in tackling the issues in video summarization, further efforts are required to enhance and improve their functionality for broader use.

1.9 Research Objectives

Developing efficient algorithms and techniques for reducing large video content into shorter, more digestible summaries while maintaining the most pertinent and significant information is the main goal of video summarization.

The research objectives of this thesis are as follows:

- **RO1**: To investigate the video summarization techniques and how these techniques are used in different applications.

- **RO2**: To propose and implement a video summarization model for use in real-time scenarios using multimedia components (Audio and Text).

- **RO3**: To Propose a model for summarizing a video that establishes a spatiotemporal relationship that generates a time and memory-efficient summary.

- **RO4**: To evaluate and compare the performance of the proposed method with existing state-of-the-art methods.

This research focuses on the significance of assessing video summarizing methods, investigating their applicability in many fields, developing a real-time processing algorithm that is effective, and emphasizing memory and time efficiency. It also emphasizes the necessity of using multimedia elements, having a thorough comprehension of video content, and striking a balance between efficiency and accuracy.

1.10 Thesis organization

This thesis is organized into eight chapters, each focused on specific aspects of the research and contributing to the understanding and evaluation of the proposed nature-inspired optimization models. Finally, the configuration of the thesis is as follows:

Table 1.2: Mapping of ROs to Publications

Research Objectives(ROs)	Publications
RO1	• Shambharkar, Prashant Giridhar, and **Ruchi Goel**. ”From video summarization to real time video summarization in smart cities and beyond: A survey.” Frontiers in big Data 5 (2023). Vol. 5, 1106776 DOI: 10.3389/fdata.2022.1106776
RO2	• Shambharkar, Prashant Giridhar, and **Ruchi Goel**. ”Analysis of Real Time Video Summarization using Subtitles.” In 2021 International Conference on Industrial Electronics Research and Applications (ICIERA), pp. 1-4. IEEE, 2021. DOI: 10.1109/ICIERA53202.2021.9726769
RO3	• Shambharkar, Prashant Giridhar, and **Ruchi Goel** “VSEM: A Hybrid Model for Video Summarization” Journal of Information Science and Engineering. • Shambharkar, Prashant Giridhar, and **Ruchi Goel** ”TAVM: A Novel Video Summarization Model Based on Text, Audio and Video Frames.” In 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS),pp. 878-882. IEEE, 2023. DOI: 10.1109/ICTACS59847.2023.10389978
RO4	• Shambharkar, Prashant Giridhar, and **Ruchi Goel** ”Auto encoder with mode-based learning for keyframe extraction in video summarization.” Expert Systems 40, no. 10 (2023): e13437. DoI: https://doi.org/10.1111/exsy.13437

- **Chapter 1: Introduction**
  This chapter offers a comprehensive examination of video summarization, including a basic overview, importance, and key components. It explores a range of methods, models, and applications related to video summarization, elucidating the motivations driving research in this field. The chapter also clearly defines the problem statement, research questions, and research objectives. Finally, the layout of the thesis is also discussed.

- **Chapter 2: Literature Review**
  Chapter 2 establishes the foundation for the research. A thorough summary of pertinent papers and current work in the field of video summarzation is provided in this chapter to help put the study in perspective. It also provides an extensive survey of various datasets commonly employed in video summarization research, as well as an exploration of evaluation metrics used to assess summarization performance.

- **Chapter 3: Analysis of Real Time Video Summarization using Subtitles**
  Chapter 3 presents that essential components of a video can be summarized by the summation of subtitles in a video using text summarization and video mapping algorithms. The audio-generated version of the summary subtitle file will be played with the summarized video. Real-time image and video processing involve producing output while simultaneously processing input. To improve the summary process, text retrieved from subtitles is integrated with the ongoing video processing in real-time video summarization analysis. Specifically, when subtitle files are used, the Latent Semantic Analysis (LSA) technique performs better than the Text Rank method. On average, using the Text rank method saves 65-70% and the LSA method saves the user 75-80% of their time. This technique not only saves the viewer's time but also provides relevant content in a short amount of time.

- **Chapter 4: Auto Encoder with Mode-based Learning for Keyframe Extraction in Video Summarization**
  Chapter 4 illustrates a hybrid algorithm to handle constrained optimization problems. It uses a self-adaptation strategy and a parameter-free penalty function, with substantial experiments and comparisons with other techniques. A novel supervised learning method, "TC-CLSTM Auto Encoder with Mode-based Learning," is proposed for automatically choosing keyframes or important sub-shots from videos, where TC stands for Time Distributed convolutions and CLSTM stands for convolution-long short-term memory. Mode-based learning is the use of annotation mode for deciding key frame or frame importance. The key is to find the

importance of frames based on the importance score provided in the annotation file. Those frames are combined to form a video summary. An encoder is used to reduce the number of frames. The key frames chosen should uniformly represent the entire video, minimizing redundant or missing data. Convolutional LSTM is used to classify frames accurately and finally decoder is used to reconstruct the frame. F- Measure is employed as the quantitative metric for assessing the outcomes of the experiments performed on the TVSum dataset.

- **Chapter 5: VSEM: A Hybrid Model for Video Summarization**
  Multimedia features (text, audio, and image) play an essential role in a video summary; combining these can be effective. In this chapter a hybrid model (VSEM) for the video summarization is proposed. Audio and image features are combined with text, and a hybrid model is proposed. The first part concerns the subtitles in a video. Hybrid text summarization is proposed using text and frequency summarization and is compared with the state-of-the-art methods. To tackle the audio aspect of the multimedia, sound chunks are taken and from these audio chunks MFCC (Mel-frequency cepstral coefficients), Mel Spectrum, area under the audio curve, and audio peak after average cut-off were obtained. The third aspect of images or video frames is traversed by considering the changes in each frame using Mean Absolute Difference. The experimental results on TVSum show that the multimedia components—text, audio, and image—can offer the summary task more information and accuracy than a single visual feature.

- **Chapter 6: TAVM: A Novel Video Summarization Model based on Text, Audio and Video Frames**
  In this chapter a novel method TAVM (text, audio, and video mode) is proposed that will provide the video summary using different multimedia elements of text, audio, and frames. The proposed method is separated into three parts. The process begins with video processing, where the BEiT vision transformer is employed to recognize objects within the chosen frame. Following that, Audio Processing comes into play, which uses speech-to-text converters to transcribe the audio content. Finally, in the last step, the summary builder utilizes the GPT-3-based OpenAI API to generate a summary of the content. The experimental analysis on the benchmark dataset SumMe demonstrates the effectiveness of the proposed approach.

- **Chapter 7: Keyframe Extraction via Peak Wave Analysis with Integrated Human Presence and Face Recognition** In this chapter an innovative artificial intelligence (AI)-driven approach to address the challenges of identifying keyframes and summarizing video content efficiently is proposed. The method-

ology involves the integration of three models: Model 1 for human presence detection (HPD), Model 2 for face presence detection (FPD) using Peak Wave Time analysis (PWT) and Model 3, an advanced YOLO V face recognition model. The synergistic integration of these approaches is intended to meet the issues faced by the growing volume of video data. The proposed approach on TVSum dataset aims to contribute to the field of video summarization by enhancing accuracy and reducing time requirements through a multi-model approach combining audio and video.

- **Chapter 8: Conclusion and Future Scope**
  The last chapter covers the results of the suggested study, provides a summary, and suggests potential directions for further research.

1.11 Chapter Summary

Video summarization addresses the growing demand for efficient information access, content discovery, and personalized learning experiences in information-rich world. This chapter provides an overview of the research work done for thesis. It describes the fundamental components of the research statement, the goals of the study, and the effective techniques for video summarization. A succinct explanation of the important terminology used throughout the thesis is also provided. Lastly, It provides a synopsis of the thesis organization and structure.

CHAPTER 2

Literature Review

The massive data uploads on social sites have raised the interest and need for video summarizaton and also spurred the rapid growth of Automatic Video Summarization (AVS).(33). AI-powered software can analyze videos and extract vital information, resulting in succinct summaries that highlight the most relevant topics. Video summarizing, or the act of reducing video footage into a succinct and instructive style, is becoming more important in today's information-rich culture. Researchers have investigated many approaches to attain this goal, each with its own strengths and weaknesses. Video summarization is broadly divided into static and dynamic summary (34). Dynamic summaries, also known as video skims, are created by analyzing the audio and visual information of a video. A static summary or keyframe-based summary is a grouping of the necessary keyframes that are required to construct the desired summary and are chosen in a sequential order (3).
Video summary approaches can be classified into the following groups based on the methodology used for summarization.

2.1 Feature based Summarization:

This method focuses on extracting significant elements from video, such as:

- **Visual features:** Color, texture, action, and recognizing objects are used to identify key frames or segments (13).
- **Audio features:** Analyzing voice, music, and other aural cues to better understand the video's narrative flow.
- **Text features:** Using captions, subtitles, and other textual information about the video.

These features are then used to select representative frames or segments for the summary.
To identify keyframes in a video, zhong et al. (35) has extracted higher-level visual features in their proposed approach. A technique called Graph Attention Network (GAT) is

used to analyze the visual features of each frame and transform them into more meaningful, high-level features. This is achieved through a mechanism called Contextual Features-based Transformation (CFT). Within GAT, a new ”Salient area size based” spatial attention model is introduced. This model is based on the observation that humans naturally pay more attention to larger and moving objects in a scene. By focusing on these salient areas, the model extracts more relevant visual features from each frame. The high-level visual features extracted by GAT are then combined with semantic information processed by a Bidirectional Long Short-Term Memory (Bi-LSTM) network. Bi-LSTM analyzes the sequence of frames in the video, capturing the overall context and meaning. By combining these two sources of information (visual features and semantic context), the system can more accurately determine the probability of each frame being a keyframe. This helps identify the most important and informative frames that best represent the video content.

The authors (26) proposed approach that first extracts features from unique video frames (after redundancy removal) using a two-stage process. In first stage leverages pretrained Deep Convolutional Neural Networks (CNNs) proven effective in image/video analysis. Four such models (AlexNet (36), GoogLeNet (37), VGG-16, and Inception ResNet v2) are employed in a ”Multi-CNN” approach. This method combines the strengths of each CNN by concatenating their extracted features into a richer, unified representation. The specific models were chosen based on a previous study that analyzed individual CNN performance for this task. This approach is efficient and scalable but might miss semantic relationships and context within the video. Rani et al. (38) proposed a static video summary method combining four distinct visual features—correlation, mutual information, color histogram, and color moments. Through the use of several visual elements, the approach seeks to minimize information loss by capturing all the details required for evaluating changes in the visual content of frames. After the frames are fused, they are grouped using Self-Organizing Maps (SOM), and the euclidean distance between the frames is used to determine which frames are the most representative within each cluster.

2.2 Search-based Summarization:

This approach treats video summarization as a retrieval task:

Query-based summarization: A major hurdle in video summarization research is user subjectivity. People have diverse tastes and priorities when it comes to what they find important in a video. This subjectivity makes it difficult to create a single ”one-size-fits-all” summary that will satisfy everyone. What one viewer might find crucial, another might skim over entirely. This emphasizes the challenge researchers face in creating summaries that cater to individual preferences. This allows users to specify a

specific query or topic related to the video, and the system retrieves relevant segments that address the query.
Authors (16) developed a system that utilizes a ”memory network” to focus on different parts of the video (frames and shots) based on the user's query. This allows the summary to be tailored to the user's specific interests within the video. To train and evaluate their system, they created a new dataset called UT Egocentric (UTE). This dataset includes detailed annotations that label concepts present in each video shot. Authors proposed a new method for evaluating video summaries based on the concept annotations in their dataset. This method allows for a more precise assessment of how well the summary captures the relevant information based on the user's query.
Using a three-layer generative network, Zhang et al.(39) presented a novel method of video summarization in 2018 . This technique is based on the Generative Adversarial Network (GAN) model known as TPAN. Generative adversarial networks are a type of deep learning system where two neural networks compete against each other. One network (generator) tries to create something realistic (like a video summary), while the other (discriminator) tries to identify if it's real or artificially generated. Through this competition, both networks improve their abilities. In 2020, Xiao (40) proposed a video summarization method named CHAN, which consists of two parts: a feature encoding network and a query-relevance computing module. CHAN employs a convolutional network with local self-attention mechanism and query aware global attention mechanism to learn the visual information of each shot. To solve video description issues, authors (41) develops a Query-biased Self-Attentive Network (QSAN). The network creates a generic summary and a query-focused summary based on semantic information. It computes query-relevant scores for each shot using a hierarchical model, a query-aware scoring module, and a library of video captions to produce the summary that is query-focused.
Authors (42) have proposed Query-based Deep African Vulture Learning (QDAVOL) tackles challenges in MultiVideo Summarization (MVS) and aims by understanding user queries and aims to create user-specific, informative, and well-organized MVS summaries. In order to ascertain user intent, choose keyframes, and guarantee cogent flow, the summary makes use of query and online image analysis, event-based object detection, African Vulture Optimization Algorithm (AVOA), and similarity-based frame closeness.

2.3 Similarity-based summarization:

identifies segments in the video that are similar to pre-defined summaries or reference videos, providing summaries based on existing knowledge or user preferences. This approach offers flexibility and personalization but requires well-defined queries or ref-

erence summaries and can be computationally expensive. The relevant studies that we mainly review are divided into four areas.

- **Clustering-based approaches**
  In 1998, Zhuang et al. (43) first proposed a clustering approach for video summarization. Video is divided into smaller units based on scene changes. After that K-means clustering was used within shots to group frames within each shot based on their color similarities (using color histograms). The center frame of each cluster is used to represent the visual content of that shot in the summary. Avila et al. (9) in 2011 built upon existing clustering methods to improve video summarization. The authors first identified potentially informative frames by removing irrelevant ones. This step helped focus on the most valuable content. The remaining candidate frames were then grouped based on their visual similarities using the k-means clustering algorithm. To ensure a concise and informative summary, they filtered out redundant frames within each cluster. The remaining frames were considered the final video summary. Authors (44) determined the optimal number of clusters based on the variations in visual content between adjacent frames. This approach allowed for a more flexible and adaptable summary creation process.
  Clustering-based video summarization encompasses various types, including partitioned, spectral, K-means, and similar methods (3).

- **Shot segmentation-based techniques:**
  A technique known as "shot segmentation-based video summarization" breaks a video into shots or segments according on shifts in the audio or visual content. After that, these parts are dissected and chosen to provide a synopsis that includes the main ideas of the video. A key phase in the summarizing process is shot segmentation, which aids in locating important scenes or events in the video. Shot segmentation can be done using a variety of techniques, such as those based on color histograms, motion analysis, or audio features. The resultant synopsis attempts to offer a streamlined rendition of the original video while maintaining its main ideas and organization. Authors (45) have used zero-shot action recognition for considering the correlation of action-action, label-label and action-label at the same time.

- **Sparse subset selection-based methods:**
  Sparse subset selection is a technique used in video summarization that focuses on identifying a small, representative subset of elements from the original video that effectively captures its essence. These elements can be individual video frames (keyframes), short video segments (subshots), or even features extracted from

the video content. Authors (46) have used block-sparsity based sparse dictionary selection method for video summarization and designed greedy algorithms called SBOMP. Video frame are treated individually and the relationship among frames plays an important role in keyframe extraction. The authors proposed a method called Multi Scale Deep Feature Fusion Based Sparse Dictionary Selection (MSDFF-SDS). This method allows for adjusting the contribution of each scale of features through a balance parameter. Moreover, it leverages row-sparsity consistency of the simultaneous reconstruction coefficient to select a minimal number of keyframes.

- **Graph Based Video Summarization** Graphs help the model understand the general organization and flow of the content by simply capturing the relationships between the various segments of the video. A potential method that makes use of graph structures to represent and analyze video footage in order to produce summaries is called "graph-based video summarization. A frame, shot, or section of the video is represented by each node. These nodes are connected by edges, which are frequently created by temporal proximity, semantic content, or visual feature similarity. The authors (47) developed a new graph-based structural difference analysis model and a graph-based metric to assess how two frames are different from one another. Median graphs are obtained as the corresponding keyframes, and this graph's structural difference can reflect any potential disparities between continuous frames. The authors (48) have proposed SumGraph a recursive graph modelling network for video summarizing that depicts a relation graph with frames as nodes linked by semantic relationships.

2.4 Deep Learning based Summarization:

This approach utilizes deep learning architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to automatically learn features and representations from the video. The main methodologies used are:
End-to-end learning: directly maps raw video frames to summary representations, bypassing the need for explicit feature extraction.
Attention mechanisms: focus on specific parts of the video deemed crucial for understanding, enabling the model to prioritize important information.
Generative models:create new video summaries like text descriptions or highlight reels, allowing for more creative and flexible summarization formats. Mahasseni et al. (49) introduced a method for video summarization that combined three key components: LSTM-based key-frame selector is used to identify the most important frames in the video based on their content and sequence. Variational AutoEncoder (VAE) compressed the selected keyframes into a lower-dimensional representation, capturing the

essence of the video and trainable discriminator acts as a judge, trying to distinguish between the original video and the reconstructed video generated from the summarized keyframess. The key innovation lies in an adversarial training process. The key-frame selector and VAE try to ”fool” the discriminator by creating a summary so good that it appears indistinguishable from the original video. Meanwhile, the discriminator continuously learns to better differentiate between the real and summarized versions. Through this competition, the system progressively improves its ability to generate informative and accurate video summaries. Building upon this idea, Apostolidis et al. (50) further refined the training process by introducing a step-by-step, label-based approach. This optimization led to even better summarization performance in their experiments. Deep learning approaches offer promising results in capturing complex relationships and generating diverse summaries but require large datasets for training and can be computationally expensive. Yuan et al. (51) employed a deep-learning approach to automatically summarize videos. They used 3D-CNN to extract both low-level and high-level features from the video frames. These features were then combined using a specific technique to create a comprehensive representation of the video content. Next, they employed a recurrent neural network called a convolutional LSTM to capture the relationships between different parts of the video across time. This allowed the model to understand the flow and progression of events within the video. Finally, they devised a novel loss function, called the Sobolev loss, to guide the training process. Important scores assigned are matched to each video frame by the model with the ground truth scores provided by human experts. Secondly the temporal structure of the video is exploited by considering the importance of frames in relation to their neighbors in the sequence. Deep learning techniques include reinforcement learning, unsupervised, supervised, and weakly supervised methods (34). It is anticipated that deep learning models will advance in intelligence and produce improved video summarizing methods as research on the subject progresses. Models have proven a considerable improvement in recognizing crucial moments and producing relevant summaries.

2.5 Multimodal Summarization:

This approach combines information from multiple modalities (visual, audio, and text) to create a more comprehensive understanding of the video content. The creation of a multi-modal summary by humans necessitates the utilization of their pre-existing understanding and external knowledge to generate the content (52).

- Jointly analyzing visual features and speech recognition: provides a richer understanding of the video’s narrative and context.
- Integrating captions and subtitles: leverages pre-existing textual information to

enhance summary accuracy and informativeness.

Authors (53) has done text summarization of subtitle file using LSA then video mapping is accomplished by selecting a video from the subtitle file that corresponds to the sentence. The authors (54), suggested a model for summarising subtitles based on LDA. LDA-generated keywords list was used to summarise the subtitles of instructive videos. The authors (55) treated each sentence as a document and created a sentence summary with a threshold equal to the total of all sentences TF-IDF values. Extractive text summarization was found to lower the original content by 60%. Moreover, removing stop-words has no bearing on the final report. Authors (56) proposed a architecture in which a movie including subtitles is taken as input, and divided into various scenes. Each scene is described in a single sentence, and the descriptions and subtitles are merged to provide a final summary. Three of the video's four main scenes/storylines are described in the summary. For scene description generation, the S2VT (Sequence to Sequence—Video to Text) algorithm is used, and for extractive text summarization, MUSEEC (MUltilingual SEntence Extraction and Compression is used (56).

Multimodal summarization can improve summary quality but requires sophisticated techniques to effectively handle and integrate information from different sources.

2.6 Video Summarization Evaluation

Measuring the effectiveness of video summarization frameworks can be tricky. Unlike other fields, there's no single gold standard for quantitative evaluation. By combining quantitative and qualitative approaches, researchers can gain a more comprehensive understanding of a video summarization framework's effectiveness. Here's a breakdown of some common metrics used:

2.6.1 Information Retrieval Metrics:

- **Precision:** This measures how much of the information in the summary is actually relevant to the original video. A high precision means the summary avoids including irrelevant details. A formula to calculate precision is given in Eq. (2.1) as given below.

$$Precision = \frac{Total\ Number\ of\ Relevant\ frames \bigcap Retrieved\ frames}{Retrieved\ frames} \tag{2.1}$$

- **Recall:** This measures how much of the key information from the original video is captured in the summary. A high recall indicates the summary doesn't miss

important points. Recall is calculated as given in Eq. (2.2) below.

$$Recall = \frac{Total\ Number\ of\ Relevant\ frames \bigcap Retrieved\ frames}{Relevant\ frames} \quad (2.2)$$

- **F-Score:** This combines precision and recall into a single metric, providing a balanced view of the summary's performance. F-measure is as given in Eq. (2.3).

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \quad (2.3)$$

2.6.2 Additional Metrics:

- **Accuracy:** This is a more general measure of how accurately the synopsis represents the original video's content. It can be difficult to specify accurately for a video summary.
- **Computational Time:** This metric assesses the efficiency of the summarization process, measuring how long it takes to generate a summary. Faster processing is desirable in real-time applications.
- **Number of Keyframes:** This statistic measures the efficiency of the summarization process by determining how long it takes to produce a summary. Real-time applications benefit from faster processing.
- **Compression Ratio (CR):** This compares the original video's size to the size of the summary, demonstrating how much data has been compressed. A higher CR indicates a more effective summary. A formula to calculate compression ratio is given in Eq. (2.4)

$$Compressionratio = \frac{Number\ of\ keyframes}{Total\ number\ of\ frames} \quad (2.4)$$

- **Area Under Curve (AUC):** This metric is used in some studies to assess the overall performance of a summarization model based on its ability to distinguish relevant and irrelevant information.

2.6.3 Metrics for Neural Network-based Summarization:

- **Classification Accuracy:** In some neural network approaches, summarization is treated as a classification task. This metric measures the model's ability to correctly classify frames or segments as important or not for the summary.
- **Training Time:** This measures how long it takes to train the neural network model used for summarization. Faster training times are generally preferred.

- **Error Rate:** This metric assesses the number of mistakes made by the summarization model, such as including irrelevant information or missing important parts.

While these metrics provide valuable insights, it's crucial to remember that quantitative evaluation alone may not tell the whole story. Human judgment plays a vital role in assessing the quality of a video summary. Factors like coherence, informativeness, and engagement are often evaluated through qualitative methods like user studies.

2.7 Datasets

Video summarizing datasets are essential for training, assessing, and comparing various methodologies, as well as for promoting cooperation and reproducibility. Below mentioned datasets as mentioned in Fig. 2.1 are crucial for advancing innovation and advancement in the area.

A **generic video summarization** dataset consists of a varied selection of movies with

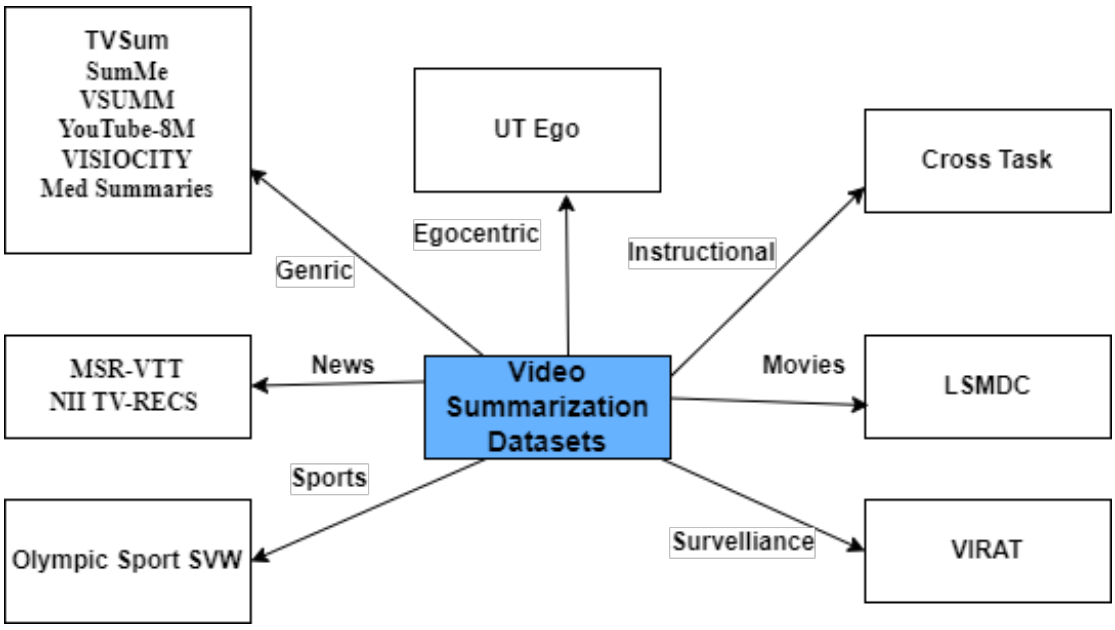


Figure 2.1: Video Summarization Datasets

accompanying summaries or annotations that cover a range of subjects, styles, and durations.

- **SUMME:** The SumMe dataset comprises 25 videos on various subjects like cuisine, travel, sports, and so on. The majority of these are unedited videos. For each video, 15- 18 users are employed to choose key shots and write video descriptions with a total of 41 subjects participating. The annotation is binary, indicating whether or not a frame should be included in the summary. Each video is

between 1 to 6 minutes in length.

- **TVSum:** 50 different kinds of videos, including news, documentaries, and vlogs, along with 1,000 annotations of shot-level relevance scores that were gathered through crowdsourcing (20 of each video) at 30 frames per second are included in the TVSum dataset (57). Without requiring (expensive) user study, the video and annotation data allow an automatic evaluation of different video summarizing approaches.

- **YouTube-8M:** The YouTube-8M dataset (58) comprises millions of YouTube videos, each tagged with one or more categories, however it is not specifically created for summarization. Video summarization can be achieved through its application in large-scale unsupervised or weakly supervised learning methods.
- **VISICOSITY:** is a vast compilation of 67 lengthy videos that cover six distinct categories: sports (soccer), TV shows (Friends), education (tech-talks), birthday celebrations, and weddings (15).
- **MED SUMMARIES:** A new dataset for the assessment of dynamic video summaries is called "MED Summaries". It has 160 videos with annotations in total: 60 videos for validation and 100 videos for testing.Within the test set, ten event categories are present.

Video Surveillance systems produce vast amounts of video data, it is not possible to maintain and check whole data manually. The domain of video summarization techniques are vital because it facilitates the effective assessment and identification of prospective incidents by automatically extracting the essence of surveillance videos. In Video Surveillance domain VIRAT dataset is used.

- Compared to previous action recognition datasets, the **VIRAT** (Video Image Retrieval and Analysis Tool) video dataset (59) is intended to be more realistic, natural, and difficult for video surveillance domains in terms of quality, background clutter, scene diversity, and human activity/event categories.

For **Instructional videos** Cross Task dataset is used .

- **Cross Task** dataset(60) of 4.7K movies and 83 jobs about house repairs, cooking, handicraft, and auto maintenance. These exercises and their methods are taken from wiki How, a website that offers solutions to a variety of problems, and the videos are from YouTube.

Sports Videos in the Wild (SVW) is a Video Dataset for Sports Analysis.

- A collection of videos called Sports Videos in the Wild (SVW) (61) that were taken while participating in sports or watching games by users of Coach's Eye, the top sports training app on mobile devices. The dataset comprises 4100 videos that were chosen after examining about 85,000 recordings. It includes 30 sports categories and 44 actions.

2.8 Research Gaps

With the rapid growth of user-generated videos, being able to browse them effectively is becoming increasingly vital. Video summarization is seen to be a potential method for effectively realizing video content by identifying and selecting descriptive frames from the video. An automatic video summary would be advantageous for everyone who wants to save time and learn more in less time as video content continues to grow at a rapid rate. The following research gaps have been identified through a comprehensive literature review:

- Most video summarization techniques primarily rely on visual features. However, incorporating audio analysis (e.g., speech recognition, sentiment analysis) and text analysis (e.g., captions, subtitles) can provide a richer understanding of the video's narrative and context.

- Secondly, understanding the temporal evolution of information within a video is crucial for capturing important events and their progression. Real-time applications require efficient processing to ensure summaries are generated without significant delays. This involves designing lightweight algorithms that minimize computational resources without sacrificing accuracy.

- A set of characteristics is required for video analysis in order to describe visual information (Features are usually taken from the pixel values of video frames).

- To give a meaningful summary in real-time video summarization, both the spatial and temporal relationships among data must be captured.

- Currently it is difficult to create summaries that are customized to the preferences of the user, such as varying length and content focus, due to the methods' lack of adaptability and user control.

- AI models may have difficulty summarizing videos with abstract concepts or narratives. They may struggle to understand the underlying theme or story arc, resulting in summaries that lack important details.

- Efficient algorithms are required for summarizing videos while they are being received or transmitted. Due of their high computing costs, complex approaches may not be viable for real-time applications.

2.9 Overview of Relevant Studies and Research Work done

Summarizing multimedia content has not received as much attention from researchers as text summarization has over the years (62). In (63), authors provide an extractive summary using two text summarization algorithms and video mapping algorithms. The information in the video can be effectively condensed by using multiple keyframes or key-shots (26). In contrast to the discipline of computer vision, there has been a significant advancement in the evaluation of text summaries in the NLP community. First, NLP approaches were developed to assess the caliber of text that had been machine-translated from one language to another. Authors (64) employed an existing text summarising evaluation and map a video summary into text. This has the benefit of allowing semantic comparisons to be made between outlines. However, it also means that the judgement does not include visual elements like shaky cameras, as long as a specific piece of content is portrayed. By measuring the number of sub-shots that overlap between a given video summary and a ground-truth video summary, authors (65) develop VERT. This system assesses video summaries compared to a provided video summary. The drawback of pixel-based distance is another drawback of this technology. Additionally, individuals frequently struggle to create a video synopsis that accurately reflects the video instead of writing, which they find easier to produce. Asynchronous text, image, audio, and video-based summary of the video was suggested by Haoran Li et al.(66). After analysing each asynchronous component separately and using several optimization techniques on the summary, a more accurate final textual summary is generated. Salience matching is also carried out to improve the relevance of the summary. A temporal and spatially driven method was put out in (67) in which the number of keyframes were automatically determined and extracted using Optimum-Path Forest (OPF) clustering before being utilized to create the final summary. Finding important frames in video summarization is an important and tedious task. A deep learning-based approach for learning video representation was proposed by Michele Merler et al. (68). Deep learning models like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), etc., are used to process the audio and visual content to learn the representation. A video that serves as the final summary is produced. To understand the spatiotemporal representations from video, the authors (69) recommended using a ranking-based method to summarise the video in many stages. Ali Javed et al. (70) suggested a technique for enumerating the cricket video's audio-visual components. The final summary for the cricket match videos is prepared by identifying the critical

frames for the audio and visual content. To increase the effectiveness of summarization, authors used an Audio-Visual Recurrent Network (AVRN) to include audio and visual information in video summary tasks (71). Utilizing the latent consistency between audio and visual data is possible with the audio-visual fusion LSTM. The self-attention video encoder can detect global dependencies throughout the entire video stream. In (72), abstractive summaries of narrated instructional are generated on several subjects, including sports, cooking, and gardening, using transfer learning. A pre-trained BERT encoder and a transformer decoder with random initialization are used in the transformer design. Auto-generated instructive video scripts using the BertSum abstractive summarization model have a quality level comparable to descriptions chosen randomly from YouTube user submissions.

Videos include a wealth of information in the form of vision, text, and audio, and this information learning has been investigated in video summarization. Video summarization effectively and efficiently speeds up video processing, storage, administration, and retrieval, making comprehending and analyzing certain circumstances or occurrences in long recordings easier(67).

Zhang et al. (73) define video summarising as a technique of selecting a subset of video shots and used Long Short Term Memory (LSTM) units to generate video summaries. To extract video summaries, Zhao et al. (74) used a layered structured adaptive network to enhance the LSTM network. To combine audio and visual information into the video summarising job, authors in (75) propose an AudioVisual Recurrent Network. The suggested model makes use of two-stream LSTM for sequential audio and visual feature encoding, dynamic multimodal information fusion, and a self-attention module for global video data gathering. Because recurrent neural networks are unsuitable for analyzing the complicated structure of lengthy films. The authors (76) proposed a video summarization technique based on the merging of three modalities(text, audio, and image) and the recognition of significant video portions. For each frame, the audio, visual, and text saliency ratings are linearly integrated to get a multimodal saliency score. A method is proposed for automatically generating soccer highlights based on segmented images and audio-visual characteristics (77). The strong audio information improves the performance of the summarization system. The sports scoring mechanism combines descriptions and provides bespoke highlights based on end-user preferences.

Deep learning algorithms are being used to automatically create an internet trailer for any movie based solely on its subtitle (78). The proposed method looks for important textual components in the movie subtitle file that can be used to classify the movie into the right genre. Authors (63) have used Text Rank and LSA algorithm to produce a video summary using subtitles, audio, and frames. The extractive summary is produced using the proposed approach. The LSA approach outperforms the text rank method when used to a subtitle file. The authors (56) provide a structure for segment-

ing a movie with subtitles into scenes, giving one-sentence summaries of each, and then putting them all together into a summary. While MUSEEC ((Multilingual Sentence Extraction and Compression) is used for extractive text summarization, the S2VT (Sequence to Sequence—Video to Text) algorithm provides scene descriptions. Jangra et al. (79) have proposed a video summary generation method using Joint Integer Linear Programming. By expanding and manually annotating the multi-modal summarization dataset, the author has generated their own text-image-video dataset. The suggested approach extracted the most relevant video with 44% accuracy.
Overview of research work done is as below:

1. Video Summarization using Subtitles:
   This method focuses on summarizing videos by evaluating the subtitles or closed captions. Text summarizing techniques are used on the extracted text to highlight key points and provide a short summary. This method is efficient for videos with accurate subtitles, but may not work well for videos without subtitles or with incorrect ones.

2. Video Summarization using Autoencoders:
   This method uses autoencoders, a form of deep learning model, to extract key features from video frames. One study proposes an autoencoder that uses mode-based learning (TC-CLSTM Auto Encoder). This model is intended to learn useful representations of video data, possibly focusing on the most informative parts for keyframe selection. This technique has the advantage of being able to capture complex visual aspects while using minimum training data.

3. Hybrid Summarization:
   This study suggests a method for combining standard text summary and frequency-based summarization. It generates a summary by analyzing textual material from the video, most likely subtitles or transcripts, and identifying keywords or common phrases. According to the study, this hybrid strategy beats other methods in terms of evaluation measures.

4. TAVM: A Novel Video Summarization Model (Text, Audio, and Video Frames):
   This approach delves into multimodal summarization, which incorporates data from several sources. It uses text (captions, subtitles), audio (speech recognition), and video frames (visual features) to produce a more comprehensive overview. The suggested TAVM model employs various strategies for processing each data type before combining the extracted information to produce a summary. This method provides a more thorough knowledge of the video information than approaches based on a single modality.

5. Keyframe Extraction via Peak Wave Analysis with Integrated Human Presence and Face Recognition:
   This approach describes an innovative AI-driven technique for addressing the hurdles of effectively recognizing keyframes and summarizing video information. This methodology combines three different models. This integrated technique combines human presence detection (HPD), face presence detection (FPD) using Peak Wave Time analysis (PWT), and an enhanced YOLO V face identification algorithm. It is intended to address the challenges caused by the growing volume of video data. By combining these approaches, the proposed model aims to improve accuracy and reduce time requirements in video summarizing. Additionally, it incorporates both auditory and video cues to enhance its performance.

2.10 Chapter Summary

By enabling users to more effectively and perceptively explore the immense ocean of video content, video summarizing has the potential to completely transform how we interact with and comprehend the ever-expanding video landscape. This literature survey chapter explores the range of approaches used by scholars studying video summarization. It describes the various approaches investigated as well as the plethora of strategies, datasets, and assessment standards applied in order to accomplish this goal.

Publication

The Literature survey work is published in:
Shambharkar, Prashant Giridhar, and Ruchi Goel. "From video summarization to real time video summarization in smart cities and beyond: A survey." *Frontiers in big Data* 5 (2023): 1106776.

CHAPTER 3

Analysis of Real Time Video Summarization using Subtitles

The ever-increasing volume of user-generated material need more intelligent video navigation and discovery methods. Video summarization is seen to be a potential method for effectively realizing video content by identifying and selecting descriptive frames from the video. An automatic video summary would be advantageous for everyone who wants to save time and learn more in less time as video content continues to grow at a rapid rate. The essential components of a video can be summarized by the summation of subtitles in a video using text summarization and video mapping algorithms. To improve the summary process, text retrieved from subtitles is integrated with the ongoing video processing in real-time video summarization analysis.

3.1 Introduction

Whether a live stream on a personal blog or a security camera in a manufacturing facility, video data is a common asset used daily. Smart cities face various complicated challenges from managing transportation networks to securing people to enhancing emergency response times. Smart camera video data provides a rich, time-based record of urban surroundings, but its sheer volume and complexity make it challenging to analyse and use. It is necessary to provide smart cities with fast and accurate information to increase efficiency and quality of life. The volume of digital video data has expanded dramatically in recent years due to the growing use of multimedia applications in domains such as education, entertainment, commerce, and medicine (80). Smart video can collect rich data in almost real-time, these datasets can also be very large, expensive to transmit and store, and labor- and time-intensive to analyze. Secondly, This task is far more complex than analyzing text documents because of the video's multimodal character, which sends a wide range of semantics in various formats, including sound, music, static images, moving images, and text (81). The enormous video data must be managed correctly and efficiently to maximize the usability of these huge recordings. As a result, video summarization is an important and rapidly expanding study field. Users may manage and explore large videos more effectively and efficiently with the help of a video summary (82). This work aims to identify and establish the video summarising

approaches that have been discovered in the literature, with a particular emphasis on real-time video summarization. The phrase ”real time” refers to the amount of elapsed time to summarise a video that is smaller than the original video's duration. Real time video summarization aids in the indexing and retrieval of videos from a library. It also aids the consumer in deciding whether or not to view the complete video (83)
Real-time image and video processing involve producing output while simultaneously processing input. The typical frame rate is connected to real-time image and video processing. The current capturing standard is typically 30 frames per second. To process all the frames in real time, they would have to be processed as soon as they were captured. So, if the capture rate is 30 frames per second, 30 frames must be processed in one second.
Existing methods for a video summarizing generally take either an offline (27) or an online (29) approach. To generate a summary, offline techniques require knowledge of and access to the complete video stream. Such solutions, on the other hand, necessitate storing the entire video at the same time, which is resource-costly and/or unfeasible (for example, for unboundedly long video streams).
Alternatives to the aforementioned include online or streaming video summarizing tools. An online summarization method takes a video stream and generates a summary on the go and at any time as the data stream elements come, without relying on any future data. Because they simply maintain a small piece of the previous video stream (or information related to it) in memory, such approaches can be made to use less memory. This situation is particularly interesting because online methods are computationally less expensive than their batch counterparts when batch processing a video is too resource-intensive on a device, then an application needs access to the historical summary, or for unboundedly long video streams.

3.2 Application of Real Time video Summarization

Applications for real-time video summarization are diverse and span many different fields where it is essential to have immediate access to summarized video data. A real-time video summarization system can process videos online, eliminating the possibility of backlogs. Fig. 3.1 shows the application of real time video summarization in different fields.

Real-time feedback and processed images from sensors are required for a variety of real-time applications as shown below.

- **Video Surveillance**: Real-time video summarization allows surveillance systems to quickly identify and highlight significant events such as breaches, unapproved access, or dubious activity. These occurrences can be summarized in real time,

Figure 3.1: Real Time Video Applications

allowing for timely notifications to law enforcement or security staff.

- **Traffic Monitoring**: Allows for better mobility and efficiency on roads by helping authorities to efficiently manage traffic operations, increase road safety, and optimize traffic flow. Real-time traffic incident, accident, or road danger detection and highlighting is possible with summarization algorithms.

- **Medical Videos**: Real-time video summarization from medical recordings improves patient care, medical education, research, and quality control in healthcare environments by giving users prompt access to vital data and insights from treatments and procedures. Decision-making and necessary modifications are made easier for surgeons and healthcare teams by summaries of live video feeds from operating rooms or medical operations, which give them instant insights into the procedure's progress.

- **News and Media**: By using real-time video summarization, news organizations and broadcasters can swiftly take the most important points from recorded or live news pieces. Then, in real-time, viewers can get succinct summaries of significant occurrences or breaking news, enabling quicker information transmission.

- **Sports and Internet Videos**: Broadcasters may instantly deliver highlights and significant moments from live sporting events, including basketball games, tennis tournaments, or soccer matches. Short recaps of games can be shared by fans on social media, which sparks conversation and interest among followers.

- **Militray Applications, Drones and Robots**: The successful and safe execution of military operations and missions are done. It also helps to improve situational

awareness, decision-making, and operational efficacy in a variety of mission-critical scenarios.

3.3 Stategies Employed in the Architecture

Automatic text summarization methods are desperately needed to deal with the ever-increasing amount of text data available online, to improve both the discovery and consumption of relevant information. One of the main reasons for implementing text summarization into our system is that it improves the efficiency of the text summarizing process. Text summarization can be accomplished in two ways: extractive and abstractive. In extractive summarization, only the key sentences or phrases are taken from the original content whereas new sentences are created from the source material in abstractive summarization. Depending on whether it is done on a live video or a recorded video, video summarization can be characterized as real time or static. Video frames are frequently generated at such a quick rate in real-world circumstances that longer segments make sense (84). Subtitle summarization is analogous to document summarization in that it focuses on significant information to include in the summary, such as a sentence or not. For text summarization, LSA and TextRank algorithms are used. TextRank is a graph-based text processing ranking model that may be used to determine the most relevant phrases in a text as well as keywords (85). LSA is a text summarising method that uses statistical approaches to assess a text's semantic structure. This technique prioritizes keywords in a phrase over linguistic features and word order . The proposed architecture includes the following major steps as shown in Fig. 3.2

3.3.1 Conversion and Summarization of Text File

The challenge with video summarization is determining which video parts are "essential" and extracting them (86). The system accepts any URL video as input. This technique is intended for videos that only include subtitles. The system starts by downloading the video and then timestamp-based caption is checked from the extracted file and subtitle files from the user's provided link. To apply summarization techniques, the subtitle file is converted to a text file (.txt). Cleaning of the subtitle file is done in the preprocessing step in which stop words, lemmatization, and stemming are done. In the next step summarization of the converted text file is done using Latent Semantic Analysis (LSA) and TextRank approach. TextRank is a technique for extracting information from documents. It is founded on the premise that words that appear more frequently have greater significance. As a result, sentences with a high frequency of words are crucial. The system then assigns ratings to each sentence in the text based on this. Unlike Latent Semantic Analysis (LSA), which is an unsupervised method for extracting

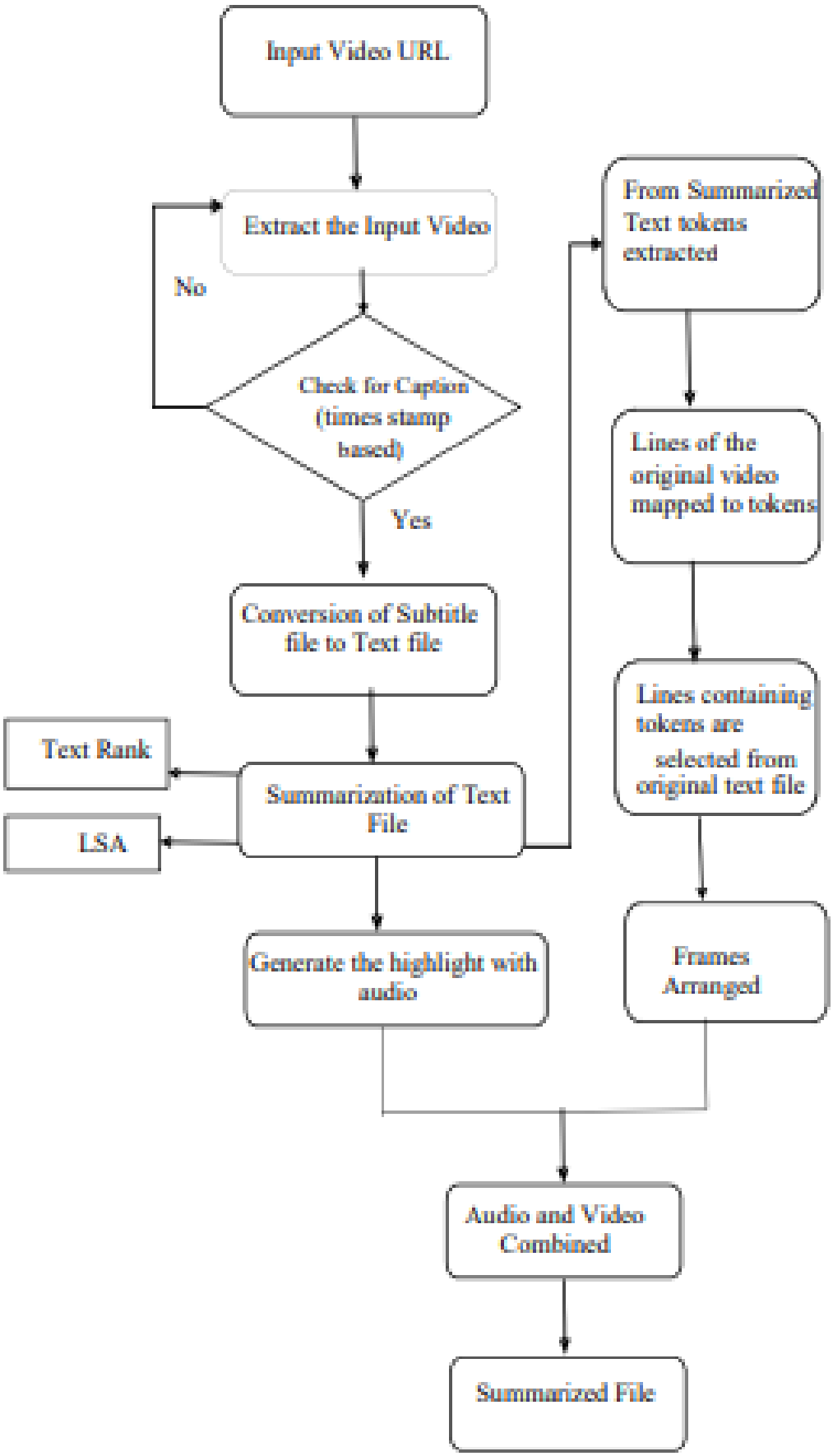


Figure 3.2: Proposed Architecture

a representation of text semantics from observed words. LSA is based on the idea that words with similar meanings are more likely to appear in comparable situations (87). The model then guesses which words should appear as incomparable documents (i.e., contexts) and hence be close to each other in the semantic space based on a study of the words that do and do not co-occur in the corpus. The LSA method begins by constructing a term-sentence matrix, in which each row represents a single word from the input (n-words) and each column represents a phrase (m sentences). The entry aij in the matrix represents the weight of the word i in sentence j. The weights of the words are calculated using the TFIDF technique, and if a phrase does not contain a word, the weight of that word is 0. The matrix is then transformed into three matrices by using Singular Value Decomposition (SVD). The summary includes the highest-scoring sentences.

3.3.2 Video Generation

The time range of the individual summarised sentences is estimated once the subtitle file's summary has been prepared. The number of seconds each sentence in the video takes is referred to as the time range. After then, for each sentence, the start and finish segments of the time range are determined separately. The start and end segments of a sentence refer to the start and finish times in seconds. The system then collects the highlights that correspond to the time range using these start and end segment times, resulting in several short pieces of video in line with the summarised subtitle text file. These crucial frames were matched with the summarised subtitle text sentences, and the final video is created by adding them to the video. The system then extracts tokens from the summary text that can be mapped to the frames in the videos Various essential frames are extracted and when combined, summarized video is obtained.

3.3.3 Audio Generation

The audio, when combined with the video, allows for a thorough understanding of the information presented in the video. Because the frames have been concatenated, the audio in the summary video will be inconsistent. Speech recognition is a key component in a variety of applications, including home automation, artificial intelligence, and so on. In Python, there are numerous APIs. For converting text to speech., used for converting text to speech. A simple pyttsx3 library of python programme is used in this architecture that translates entered text into audio that can be saved as an mp3 file. There is no need for an internet connection or any kind of delay when using this program. The audio, in addition to the video, aids in a thorough grasp of the information offered in the film. As a result, the system's final result is absolutely clear.

3.3.4 Creating final Summarized File

A video summary is created by sampling video clips and combining them with descriptions. Titles, subtitles, descriptions, inquiries, comments, and other forms of data are associated with an original video (88). After the audio and video highlight were generated next step is to combine video with audio. The semantic video record is created by deconstructing any video's caption document without sacrificing the content and video quality. Following the voice and text preprocessing, the video and generated audio are combined and synchronized to produce a nice and clean video and audio output.

3.4 Experiment Result and Analysis

Nowadays, many people use online video streaming services. YouTube is today's most popular and frequently utilized online video platform. Various videos of differing lengths were selected from the popular video sharing network YouTube and adapted for use on the suggested system. TextRank and LSA are used to calculate the experimental outcomes of the original video and the summary video. This is also noticed that the runtime of the original video and the summary video differ significantly. On average, using the Text rank method saves 65-70% and the LSA method saves the user 75-80% of their time. LSA technique will save more time as compared to TextRank. These techniques not only save the viewer's time but also provides relevant content in a short amount of time. Different educational videos for kids are taken for testing. First Input video and their corresponding subtitle file are downloaded. Initially, the video consists of sentences 62 sentences which is summarized into 20 sentences in total. In the next step to apply LSA and TextRank algorithms, extractive summarization techniques are used. The subtitle file is converted to the text file and then preprocessing is done. Video is obtained after getting highlights from the summarized text. We obtained various highlights from the video as shown in Fig. 3.3.

Then summarized video is given voice using python pyttsx3 and at last, the final summarized file is obtained. It is observed that the original video was of 156 seconds and the summarized video is of 108 seconds. Total time saved is 65- 70% using TextRank algorithm.

3.5 Summary

The growing prevalence of video content on the internet necessitates a more efficient method of expressing or managing it. The summary of videos can be used as a means of accomplishing this. Video summaries can be generated automatically with subtitle files by utilizing NLP-based algorithms. Instead of training algorithms with large

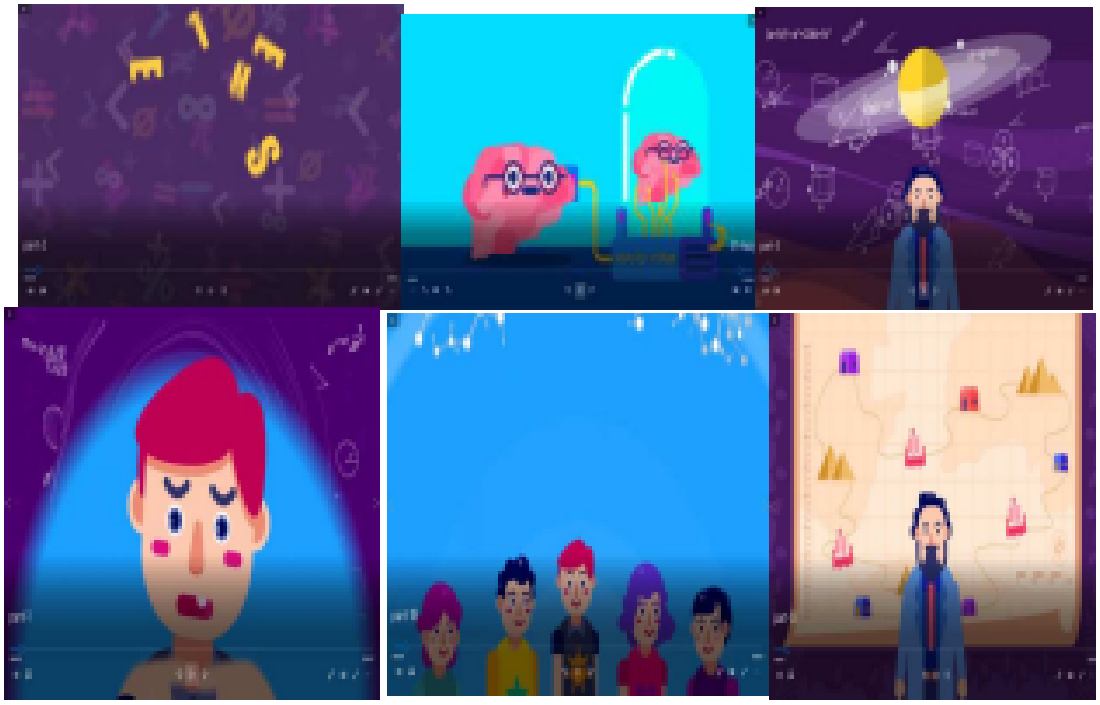


Figure 3.3: Various Highlights of Summarized Video

video datasets, this strategy makes use of text processing, which is quicker and easier to understand

The suggested method's main contribution is:

1. To improve the summary process, text retrieved from subtitles is integrated with the ongoing video processing in real-time video summarization analysis.

2. Specifically, when subtitle files are used, the Latent Semantic Analysis (LSA) technique performs better than the TextRank method.

In general, real-time video summarization has advantages in terms of accessibility and immediacy, but it also has drawbacks in terms of processing effectiveness, content complexity, and summarization accuracy. Algorithms for real-time summarization could find it difficult to adjust to dynamically changing content, including scenes that change quickly or unexpected events. As a result, summaries may become out of date or irrelevant to the situation at hand. This can be solved with our next suggested approach, detailed in the next chapter.

Publication

The work discussed in this chapter is published in:
Shambharkar, Prashant Giridhar, and Ruchi Goel. "Analysis of Real Time Video Summarization using Subtitles." *In 2021 International Conference on Industrial Electronics Research and Applications (ICIERA)*, (pp. 1-4). IEEE, 2021.

CHAPTER 4

Auto Encoder with Mode-based Learning for Keyframe Extraction in Video Summarization

The exponential increase in video consumption has created new difficulties for browsing and navigating through video more effectively and efficiently. The proliferation of advanced photographic devices and the improvement of internet connectivity have brought exponential growth in technology. Computers use their magic to create these video summaries. The original video is automatically analyzed, and crucial components—like important frames or brief video snippets are extracted. The most important elements of these selected segments are then combined to create a brief video synopsis. Selecting video frames that best convey the main ideas of the information is known as key frame extraction. Usually, selection criteria like diversity, significance, or visual originality are used to determine these frames. Key frame-based summarization techniques choose a subset of frames to produce a condensed version of the video. A novel supervised learning method, "TC-CLSTM Auto Encoder with Mode-based Learning," is proposed for automatically choosing keyframes or important sub-shots from videos, where TC stands for time-distributed convolutions and CLSTM stands for Convolution Long-term Memory. Mode-based learning is the use of annotation mode for deciding key frame or frame importance. Without the need for text like subtitles, autoencoders are able to learn rich representations directly from the visual content of videos. This makes it possible to comprehend the video information more thoroughly, taking into account aspects like motion patterns, spatial relationships, and visual characteristics that might not be adequately conveyed by subtitles alone. The key is to find the importance of frames based on the importance score provided in the annotation file. Those frames are combined to form a video summary. Encoder is used to reduce the number of frames. The key frames chosen should uniformly represent the entire video, hence minimizing redundant or missing data. Convolutional LSTM is used to classify frames accurately and finally decoder is used to reconstruct the frame. Convolutional layers can recognize the same pattern regardless of where it appears in the input since they are translation-invariant. The same set of filters are used over the whole input area to accomplish this characteristic.

4.1 Introduction

Video summaries condense lengthy videos into shorter clips packed with the most valuable information. This saves viewers time by letting them grasp the essence of the video quickly. Video is the most widely used type of visual information, which has grown in popularity swiftly. The task is to analyze the video's multi modal data in search of relevant cues (such redundant information) that can guide a conclusion (89). High computational power and advanced vision techniques have increased the scope of video summarization techniques. Video summarization helps us to quickly review lengthy videos by removing pointless and unnecessary frames.

Choosing a frame plays an important role (90) so the main idea is to summarise the important sections of the frames. Consider the video "Teacher leaves home in the morning for college and returns home after taking lectures and leaves again for taking lectures and returns to home in the evening" as an example. Even if the frames associated with the "in college" and "at home" sequences may look similar aesthetically, the semantic flow of the film dictates that none of the frames should be removed and that they are all crucial. The decision of whether or not to label a frame as summary-worthy depends on nearby frames, which makes video summarising by nature a sequential task. If a specific frame has been given a high value by the summarizer system, the nearby frames should also be given higher importance. Using an excessive number of nearby frames can lengthen the summary and reduce its usefulness. Additionally, the process of creating summaries might get more laborious as a video gets longer.

This chapter formulates video summarising as a key frame selection technique from the video. Every time a significant change in the scene occurs in the video, the intention is to automatically give a video summary in the form of key frames. The proposed model differs in terms ofthe Encoder and Decoder that have been used to lavish the information in the best method possible. It helps in better and faster learning compared to state-of-the-art models. It is shallow and hence requires lower resources.

The video provides both spatial and temporal information in the form of the appearance of each frame and the motion between frames (91). The collection of all the frames that make up the video constitutes a video summary.

$$V = \{F_1, F_2, F_3 - - - - - - F_s\} \tag{4.1}$$

Eq. (4.1) represents sequential video frames: where 'V' is the video, F_i shows i^{th} frame and the number of frames in the entire video is s. The goal of the video summary by keyframe is to choose a portion of V that is shorter in length and covers almost all of

V's crucial frames. The task of summarising videos requires the use of spatiotemporal features (92).

4.2 Application of Keyframe Extraction in Video Summarization

Keyframe extraction plays an important role in video summarization because it identifies representative frames that accurately convey the essence of the video information. Keyframes act as visual foundations, allowing users to move through videos more effectively. By using keyframes as reference points, users can easily navigate to specific areas of the video that are of interest to them without having to watch the entire video. Keyframe extraction methods improve the usefulness and accessibility of video footage in a variety of applications. Keyframe extraction is used in surveillance systems to detect significant events or actions caught in video footage. By picking keyframes that reflect noteworthy activities or anomalies, security professionals can quickly evaluate footage and respond to potential threats in real time.
Keyframe extraction can also be used in video recommendation systems to evaluate video content and suggest similar or relevant videos to consumers based on their interests. By comparing keyframes retrieved from multiple videos, recommendation systems can find commonalities and offer videos that are relevant to the user's interests.

4.3 Stategies Employed in the Architecture

Temporal characterization uses correlations between images to monitor changes. When we examine one image, spatial characterization is relevant. It includes but is not limited to the coordinates, intensity, gradient, and resolution. An approach to accomplish this goal is to give each frame a score based on importance, then the frames with the highest scores should be chosen. The development of the summary and the prediction of the importance score appear to be the two main components of this activity.

$$Max_x = \Sigma_{i=1}^{5} x \tag{4.2}$$

x is the score that is given to frames and based on that final frames are selected. Score x is decided based on the annotation file that was used in the data.

$$V_s = \{S_1, S_2 - - - - - - S_n\} and \quad V_s \subset V \tag{4.3}$$

where s represents the video summary and n is the number of selected frames from the summary based on x. frames are taken as input and x is estimating significance ratings of frames at the frame level. The significantt frame scores assess the relationship between the matching frames and the original video, with "important" referring to the

relationship between the video content and the high-level semantics. The frame with the highest frame-level relevance score is more representative.

For each frame of the video in regular intervals of time, The following steps are followed

Step 1: Frames were extracted from the videos using the open cv library.

step 2: The mode of the importance of each frame was taken from the annotation file.

step 3: A new dataset consisting of frame path and frame importance was created.

step 4: Frames are fed to CLSTMAE (Convolutional Long Short Time Memory Auto Encoder model)(93).

Step 5: Using the mode of importance the target model is trained and monitored for a loss (sparse cross entropy) and accuracy.

4.3.1 Model Architecture

The proposed architecture is shown in Figure 4.1. The model consists of three sections: An encoder, a ConvLSTM section, and a decoder. The frames from videos were first extracted using the OpenCV module. These frames were then fed into the encoder. The encoder consists of two convolution blocks each consisting of three layers namely the convolution layer, batch normalization layer, and activation layer (ReLU activation is used). The extracted tensor is passed to the ConvLSTM layer which captures temporal and spatial features. The input is reconstructed at output using the decoding segment consisting of two de-conv blocks. Each de-conv block consists of a Conv-Transpose Layer, Batch Normalization, and Activation Layer (ReLU activated). The final tensor is flattened and fed to Dense, which returns the probability or importance of the frame score. Any frame having an importance score greater than 3 is considered important, the rest frames are discarded.

4.3.2 Feature Extraction

Understanding and determining the most representative video frames depend on feature extraction. The suggested approach uses Conv layers for feature extraction. It is 21 21-layer deep residual network. To extract image information from each frame of the input video, residual neural networks are used(94). Convolutional networks often consist of multiple convolution layers stacked on top of each other. As the network deepens, higher-level features are extracted, which capture more abstract and complex patterns. Lower layers detect simple features, while deeper layers detect more global and high-level structures. Convolutional layers capture local patterns such as edges, corners, and textures. These low-level features are obtained by detecting variations in intensity or color gradients within small receptive fields. The proposed approach is implemented using the TensorFlow library.

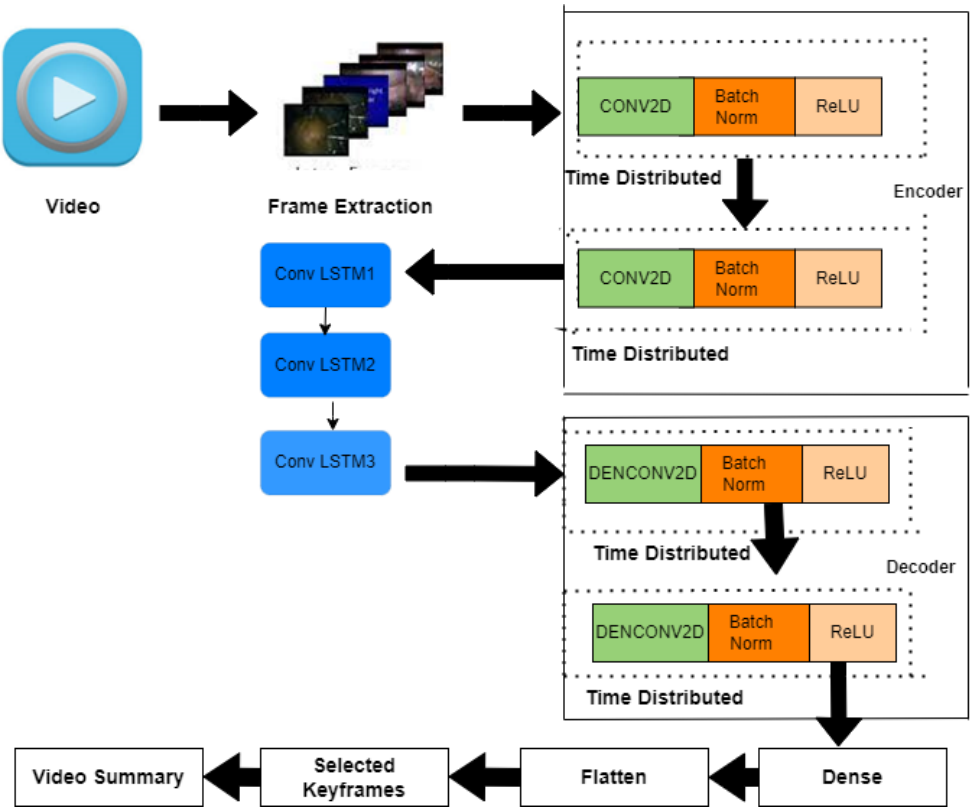


Figure 4.1: Proposed Architecture

4.3.3 Frame Extraction

Keyframes are digital pictures that give users of video retrieval systems the most comprehensible information summary. The summary should, as much as feasible, convey the shot's primary message so that the viewer is aware of the video's intended message from the outset. The videos were mp4 and extracted frames were in png format. From the given set of videos, frames were extracted using an open cv. The size of the frames was kept constant at 244 X 244. The frame rate is 25 fps which matches the original frame rate. This is done to match the annotations provided. As the annotations given for the data are based on a 2-second sequence, a frame rate differing from the pre-defined rate will result in anomalous and incorrect results. Spatial features take about the features that are collected and learned from the current frame that is being processed by the model. As video is a sequence of frames, the frames present before and after the current frames play a crucial role, thus, temporal features help in learning these features across the time dimension, i.e., the effect of different frames across time. These features in general are the weights that determine the excitation or inhibition level of neurons. sample image of frames is shown in Figure 4.2.

Figure 4.2: Frame Extraction

4.3.4 Data Creation

Provided annotation files consist of 20 judges data which contain the importance of each frame at a scale of 1-5 depending upon the individual's opinion like
1: Not Important;
2: Less Important;
3: Neither Important nor Redundant;
4: Important;
5: Very Important
The mode of this data was taken, resulting in one importance value for each frame. The arithmetic mode gives an insight into which frames are important in terms of multiple viewer experiences rather than forming the hypothesis based on a single user. This above process was carried out using Python and SciPy modules.

4.3.5 Model Training

The frames were fed to the proposed architecture. It underwent three different steps:

1. Encoding The size of frames was reduced in this step using time-distributed convolution layers.

2. Conv LSTM The memory of previous frames and current frames were combined to gain a better transformation of experience for the machine and thus classify frames more correctly.

3. Decoding The tensors were then used for reconstructing the frames, which were finally used for obtaining the importance of each frame on a scale of 1-5.

ConvLSTM is fundamentally similar to LSTM, but the difference is that it can also extract spatial information, much like CNN, in addition to learning the temporal correlation of input. ConvLSTM can therefore be used to extract spatiotemporal characteristics. The usage of this feature is required for video data modelling. The following are the pertinent formulas:

$$f_t = \sigma(W_f * [h_{t-1}, x_t, C_{t-1}] + s_f) \tag{4.4}$$

$$b_t = \sigma(W_b * [h_{t-1}, x_t, C_{t-1}] + s_b) \tag{4.5}$$

$$c_t = tanh(W_c * [h_{t-1}, b_t] + b_c) \tag{4.6}$$

$$c_t = f_t \oplus c_{t-1} + b_t \oplus c_t \tag{4.7}$$

$$o_t = \sigma(W_i * [h_{t-1}, x_t, c_{t-1}] + b_o) \tag{4.8}$$

$$H_t = o_t \oplus tan_h(c_t) \tag{4.9}$$

where * denotes the convolution operation, t represents the time step, h is the forget layer, w weights, $\tan_h$ represents the activation function, $\oplus$ is Hadamard product and b_t represents the input gate layer.

The extracted frames are first loaded into the memory using an open CV. The frames are loaded in the form of a 3-D array with dimensions as H X W X Channels. These frames are then appended to a list which is later converted to a numpy array. This array is then fed to the model during training. The first layer of the model, as evident in model architecture is the input layer, which accepts tensors (in our case image array) as input. This layer consists of a perceptron with dimensions H X W X Channel. Accuracy is used as the training performance metric in a cross-entropy loss model. Learning (determining) good values for all of the weights and the bias from labeled samples is what training a model is all about. A loss is the result of an incorrect prediction. Loss is greater if the model's forecast is imperfect; else nil. The Model training aims to identify weights and biases with low loss. For every model, a higher loss is worse (poor prediction).

4.3.6 Frame Analysis

Our model predicts a frame's importance equal to the mode of its importance as scored by judges, it is considered as right prediction otherwise it s considered as wrong prediction. The 0 represents less important frames and 1 represents important frames. Dynamic programming maximizes scores of highlight segments using the knapsack problem. Figure 4.3 shows some video frame analysis of the video we have taken as a reference from the dataset. The reference shared was based on the fact that the summary generated by them is not similar to the one present in the annotations. The one we have

created is based on the similarity between the mode of annotations and our generated summary. The variables at the current state are fed to the model to get the accuracy and loss at each training epoch.

	frame	target
0	frame\video\-esJrBWj2d8\0.png	1.0
1	frame\video\-esJrBWj2d8\1.png	1.0
2	frame\video\-esJrBWj2d8\2.png	1.0
3	frame\video\-esJrBWj2d8\3.png	1.0
4	frame\video\-esJrBWj2d8\4.png	1.0

Figure 4.3: Video Frame Analysis

4.3.7 Removing Redundant frames

The superfluous frames make the generated video summary longer than necessary and less effective as a result. The mode operation is used to eliminate duplicate frames; Frames with a score less than 3 are discarded and higher than that are considered. As a result, there are fewer frames, which results in a more effective summary.

4.3.8 Summary Generation

Despite the fact that the significance of the final nominated skims is sufficient it is possible that some visually comparable sequences will be chosen as summarised skims. The production of the final summary is difficult because of this issue, which is overcome in the suggested technique by only choosing the final summary frames whose probability is highest for the informativeness class. In order to create a diversified and representative summary, the frames with the highest probability are considered in the post-processing phase and are combined to create a single video. The final summaries are displayed with probability ratings in 3, and they include the frames from a sequence with the highest probabilities.

Algorithm 1 Algorithm for Generating a Summary

Process of Video Summarization V_s
Input: Video V,
Mode O
Output:
Video summary SV
Start Process:
Extract(V,O)
X←No. of frames(V)
For i = 0 to X1, do
Observe the current frame F[i]
Pred Mode Train (f[i], Mode(O))
Discard the frame
end for V
saved SV

4.4 Experimental Result and Analysis

4.4.1 Dataset Used

TVSum (95) contains 50 videos in a variety of genres. (such as news, how-to videos, and user-generated material), with crowd-sourced annotations of their shot-level relevance scores. Each shot is scored by 20 annotators, who use binary labels and 5-level significance rankings to describe their annotations. Most of the videos last between one and five minutes. The main advantage of using TvSum is its capacity to provide summaries of any length. TVSum is a great tool for segmentation and importance score (96).

4.4.2 Evaluation Metrics

F- Measure will be employed as the quantitative metrics for assessing the outcomes of the experiments performed on the datasets.
for one video on TvSum Number of frames = 1740
Number of frames selected as important 240
Video shorten = (1 - (240/1740)) * 100 = 86%
Average Shortened - 84.37%
The number of frames in one video of the dataset was 1740 and our approach will shorten the original video by 86% in length.
Table 4.1 shows five videos of different genres from the TVSum dataset and the time saved during each video.

Given a generated summary G_s and a ground-truth summary GT_s, the precision P

Table 4.1: Comparison of Different Genre Videos

Sr. No	Original video duration (in min)	Summarized video duration (in min)	Time saved (in min)
1	2:28	01:05	01:23
2	8:30	03:01	05:29
3	03:09	00:42	02:27
4	03:53	01:34	02:19
5	04:36	00:53	03:43

Table 4.2: Comparison with Different State of Art Techniques

Author	Model Name	F_Score(%)
Liu et al. (97)	3DST	58.1
Lan et al. (98)	Adv-Ptr-Der-SUM	58.3
Apostolidis et al. (21)	AC-SUM-GAN	60.6
Fajtl et al. (30)	VASNET	62.37
Sreeja et al. (99)	Bi-convolutional LSTM GAN	69
Proposed Method	**TCCLSTM**	**84.35**

and recall R are computed as

$$P = \frac{overlapped\ duration\ of\ G_s\ and\ GT_s}{Duration\ of\ G_s} \times 100\% \tag{4.10}$$

$$R = \frac{overlapped\ duration\ of\ G_s\ and\ GT_s}{Duration\ of\ GT_s} \times 100\% \tag{4.11}$$

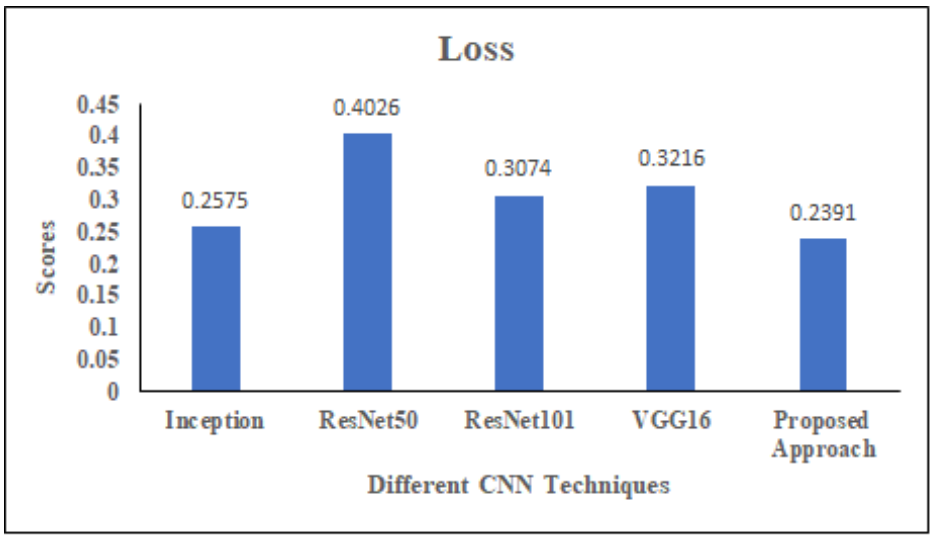
F-measure is

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \tag{4.12}$$

Table 4.2 shows a comparison with different state of art methods. Our approach achieves an F-score of 84.35 which is mentioned in bold in the table.

To determine how well the model is doing for these two sets, the loss is computed for both training and validation sets. A loss, unlike accuracy, is not expressed as a percentage. Each sample's training and validation set errors are added together to create this value (100).

For computer vision applications, CNN is the most widely used artificial neural network and is also used in various applications in real-time video applications. convolutional neural networks (CNN) prioritizes particular visual features to distinguish between images. The effectiveness of the suggested strategy is tested using a few of the most well-



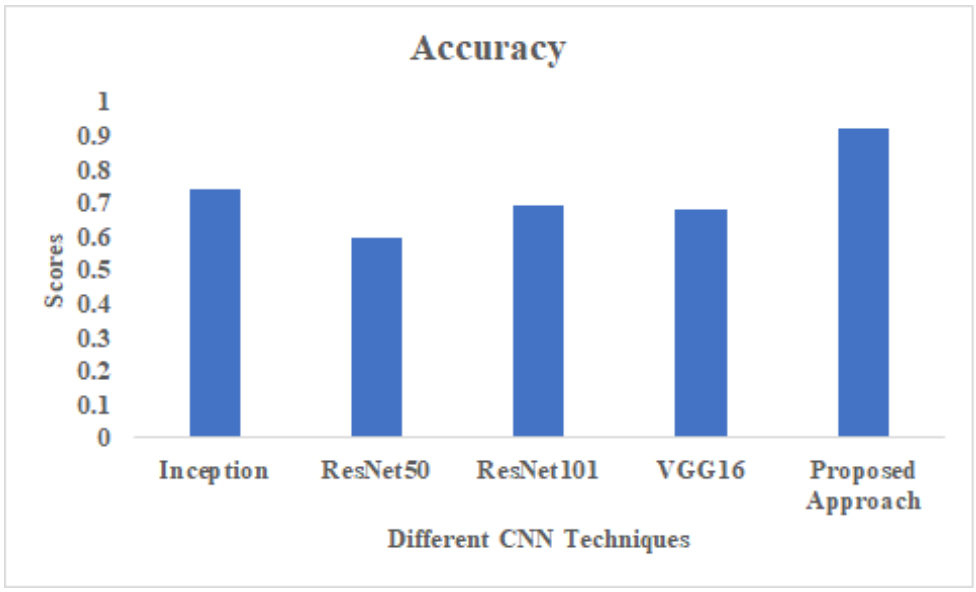


Figure 4.4: Comparison With Different CNN Techniques. (a) Loss, (b) Accuracy

liked CNN architectures as shown in Table 4.3. Their loss and accuracy were evaluated and debated and also shown in Figure 4.4.

Table 4.3: Comparison With Popular CNN Architecture

Different CNN Techniques	Accuracy Score	Loss	F1
ResNet50	0.5973	0.4026	0.59
VGG16	0.6783	0.3216	0.65
ResNet101	0.6925	0.3074	0.69
Inception	0.7442	0.2575	0.73
Proposed Method	**0.92**	**0.2391**	**0.84**

4.5 Summary

To summarize, automated video summary techniques shorten the time required to browse and view the entire video. The whole video should be represented in a succinct and clear video summary, and it is also important that the video's semantic information should be preserved. The suggested method's main contributions are:

1. The created summary using TCCLSTM gives an overview of all the main actions in a little amount of time, making the concise representation of large video streams extremely crucial.

2. By shortening the time needed for overall navigation and video retrieval, it will also enable administrators and other stakeholders to make decisions about the selection of desired material more quickly and effectively. According to the autoencoder architecture (TC-CLSTM), the model may be able to capture temporal information between frames because it uses a Long Short-Term Memory (LSTM) network. This might help locate keyframes in a video sequence that indicate important changes or transitions.

The Auto Encoder with Mode-based Learning method appears to be a solution that holds potential for keyframe extraction in video summary. Its capacities to operate with minimal data, learn useful representations, and maybe utilize temporal information are all advantageous features.
Although the Auto Encoder with Mode-based Learning for Keyframe Extraction has benefits like ease of use and computational efficiency, its primary reliance on visual features can limit its ability to fully capture the complexity and richness of video content, which could result in potential flaws in the quality of the summaries that are produced.This can be solved with our next suggested approach, detailed in the next chapter.

Publication

The work discussed in this chapter is published in:

Shambharkar, Prashant Giridhar, and Ruchi Goel. ”Auto encoder with modebased learning for keyframe extraction in video summarization.” **Expert Systems** 40, no. 10 (2023): e1343

CHAPTER 5

VSEM: A Hybrid Model for Video Summarization

Examining several media modalities, including text, audio, and visual information, is necessary for video representation (66). The video format encodes audio or text information that describes the order, structure, and content of each frame in a moving video image. A modality in the multi-modal space depends on how particular media and associated elements are organized inside a conceptual architecture. These modalities involve specific techniques or methods to encode heterogeneous information harmoniously and may include textual, visual, and aural modalities. Multi-modal learning, especially audiovisual learning, has recently garnered a lot of attention and has the potential to make many computer vision tasks (101). However, current video summarization techniques only consider the visual data and ignore the text and audio data. Text and audio modality can help the visual modality comprehend the structure and content of the video more effectively, which will also help the summary process. Assumption is that models may generate a better and more comprehensive knowledge of the underlying data, reveal new insights, and allow a wide range of applications by combining information from varied sources such as text, pictures, voice, and video.

5.1 Introduction

Technology development has caused a quick increase in multimedia data on the internet, making it difficult for consumers to access crucial information quickly (102). Video is the most challenging multimedia (including text, pictures, graphics, and audio), as it incorporates all other media data into a single data stream and is difficult to access effectively due to its unstructured format and changing format length(1). Video information is a sequential data type that gives unlimited data through its moving content (2). Think about using YouTube to look for educational or tourist videos. Most people wouldn't want to watch or listen to these lengthy recordings, but a video clip may offer a streamlined and palatable recap. It is inefficient to browse through the millions of returned results. It would be much simpler to view a brief description of each result. Secondly, because of the limited storage space, it is also necessary to summarise videos without losing much information. These issues can be solved by summarizing

the essential information from the vast amount of available content. Video summarization methods pique the viewer's interest by choosing exciting scenes from the original video (8). By highlighting significant portions of the original video, video summarising techniques can grab the audience's attention (103). The viewer can comprehend the information without watching the clip. To extract specific critical frames from a video, video summarization creates a representative summary with a smaller file size. Both the identification of the various activity sequences across time and the accurate summary of each series with the next are necessary for adequate video description approaches (104).

Additionally, eliminating redundant and useless video content may have uses in video retrieval, storage, and indexing (105). It will also increase the effectiveness of associated video analysis tasks, including action recognition and video captioning (106). Manually summarising and editing videos requires a lot of time and work. An automatic summarizing strategy is required to identify relevant incidences in the original video content. Video summary is said to be good if it possesses high recall, high precision, and low redundancy rate (4). Creating a good video summary requires thoroughly comprehending the video's structure and semantic content.

One of the significant challenges involved with the video summarizing problem is the decreasing processing costs associated with producing consistent video summaries from large amounts of data. Another challenging task is the effective fusion of multimedia resources, such as audio, text, image, and video (107). The significant occurrences can be automatically identified by evaluating the text, audio, and visual elements. Retrieving information from audio or visual content is still tricky because high-level semantic information needs to be recovered from low-level audio or visual data.

Video-based applications are used in various fields, such as security and surveillance, personal entertainment, medicine, sports, news videos, educational programs, movies, etc. The video is made up of a succession of images and some pertinent information. Textual information comprises the information's linguistic shape, whereas audio consists of speech, music, and numerous different noises. Rich media includes video, frequently combining other media forms, including text and audio (108).

5.2 Application of using Hybrid Model in Video Summarization

Hybrid architectures represent a major improvement in video summarization. Hybrid models develop more thorough and informative summaries by combining information from numerous sources, capturing the depth of video content.

A hybrid model, combines voice recognition and visual analysis, can identify essential news parts, speaker changes, and important graphics for a short summary. Secondly

analyzing both spoken and visual content can aid in identifying key concepts, examples, and explanations for a more comprehensive video overview in education.

5.3 Application of using Multimodal Architecture in Video Summarization

Multimodal architecture focuses on using information from many modalities (text, audio, and video) to provide summaries that are more useful and accurate than previous methods. Multimodal integration can be used in numerous applications like movie summarization, news summarization, lecture notes summarization in education field. Multimodal analysis may extract essential scenes, plot lines, and even emotional clues from language and music, resulting in summaries that reflect the full movie experience. Multimodal summaries can help users with vision impairments (audio descriptions) or hearing impairments (text summaries created by voice recognition).

5.4 Strategies Employed in the Architecture

It is observed that multimedia features (text, audio, and image) play an essential role in a video summary, and combining these can be effective. Video summaries on YouTube are currently based on the relevance of the frames in each video. To accurately summarise a video, a three-pronged approach is followed, as shown in Fig. 5.1. Thus, the problem statement is divided into three parts. The first part concerns the subtitles in a video and for this a text summarization tool is employed to convert the subtitles into a shorter version that includes complete sentences. Each line of the subtitles of importance is considered. Thereby, the whole list of sentences in the subtitle file acts as a corpus. Each line in the corpus is already mapped to the timestamp relative to the video. The summarization results are then divided and mapped into a list of sentences based on the timestamp.

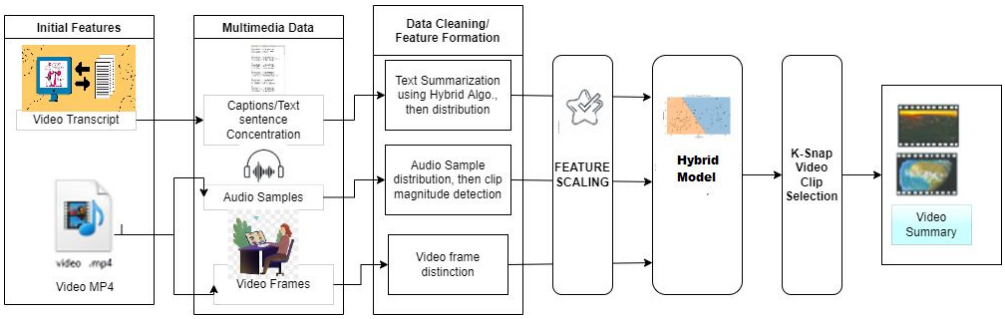


Figure 5.1: Video Frame Analysis

To tackle the audio aspect of the multimedia, files are obtained in .wav format,

and for 0.1 seconds (Persistence of Hearing) of the video, an array of audio samples are taken. These samples can be considered separate sound waves with their troughs and crests. From these audio chunks MFCC (Mel-frequency cepstral coefficients), Mel Spectrum, area under the audio curve, and audio peak after average cut-off was obtained. The multimedia, files in the .wav format are used, and a variety of audio samples are taken for each frame of the video. These samples can be considered separate sound waves with their troughs and crests. The amplitude of each such change is compared to obtain the magnitude of that wave. This list of magnitudes is linked to timestamps in the video and is used to determine the frames where the video is silent and where it is lively.

The third aspect of images or video frames is traversed by considering the changes in each frame. In a video, the fps (frame rate per second) used to be 24 (generally)(109). Still, with the advancing technology, it is often 120, 240, or even 300. For this high frame rate, the change in a video frame is minute and insignificant upon regular inspection. It is necessary to consider the picture array (pixel arrangement within the image) to determine the precise changes in the next frames. Thus images are treated as an n-dimensional array rather than an image. Mean Absolute Difference (MAD) is used to track changes over a few n-dimensional arrays with the same dimensions. This allows us to find the crucial frames in the video and spot changes.

The accuracy of our summarized video is compared to the gold standard of the video summarization dataset. The redundancy could affect how accurate the video is.

$$C_{vs} = [\{T_1 + T_2 + T_3 - - + T_n\} \cup \{A_1 + A_2 + - - + A_n\} \cup \{F_1 + F_2 + - - + Fn\}] \quad (5.1)$$

n=Total no

T1.....Tn Text in each frame of the video

A1.....An Audio of each frame of the video

F1.....Fn Number of frames in the video

$$C_{vs} = T_s \cup A_s \cup F_s \quad (5.2)$$

Where Cvs = Total Combined video Summary, which is the combination of Ts (Text Summary), As (Audio Summary), and Fs (Image summary). Final summary Cvs keeps the length of the summary to a minimum while omitting none of the crucial information from the original data. If we have original video V, then the length of Combined summary LCvs is less than the length of original video summary Vs.

$$|LC_{vs}| < |LV_s| \quad (5.3)$$

5.5 Steps Followed

Video summarization has developed to meet the demands of massive video data. The main goal of video summarization is to identify pertinent and significant video (110).
For each video, the following steps are followed:
1. Download the video transcript and then video
2. Apply the MAD on the video frames and find changes in frames and keyframes.
3. Apply the text summarization algorithm to the video transcript.
4. Extract the .wav file from the .mp4 file to extract the audio features.
5. Combine all the 3 types of Scores to get a unified score.
(The clips with the value above cut-off from that score is the summary)
The design of the proposed VSEM system consists of stages as shown in Fig. 5.2.

5.5.1 Preparation of Text File

In video summarization, main challenge is the key frame selection approach that takes into account the keyframes relevance as well as the temporal relationships between the frames in the video. The next difficult task is to create a system to evaluate the accuracy and completeness of the chosen keyframes.
Text summarization systems extract brief information from a document. By the summary generated by a system, the user can determine whether a document is relevant to his or her needs without reading the entire document.
There are two methods of producing automatic text summaries extractive and abstractive (38). The extractive approaches evaluate each sentence's relevance before choosing the best-scoring ones with the least amount of redundancy. The methods for abstractive text summarization take the original text's location and extract its most important details. Abstractive techniques are more accustomed to the human summary, which is more precise, logical, and expressive. Since the captions dialogues cannot be altered in a video, extractive summarization is used. For the extractive summarization approach, a hybrid of text rank summarization and frequency summarization is implemented.
An unsupervised graph-based content extraction technique called text rank employs the Bag of Words via Word2Vec to give words a numerical value and then uses a cosine similarity matrix, a page rank implementation, and a sentence graph to assess the value of sentences.
The drawback of text rank is that it omits keywords that, despite being relevant in context, have a lesser chance of appearing. The obvious approaches to increase text rank are to use terms that are semantically related to one another and to avoid selecting incorrect vital keywords (111).
The premise of frequency summarising is straightforward: sentences with high-frequency words in the paragraph, excluding stop words from the nltk toolkit, are rated highest. It

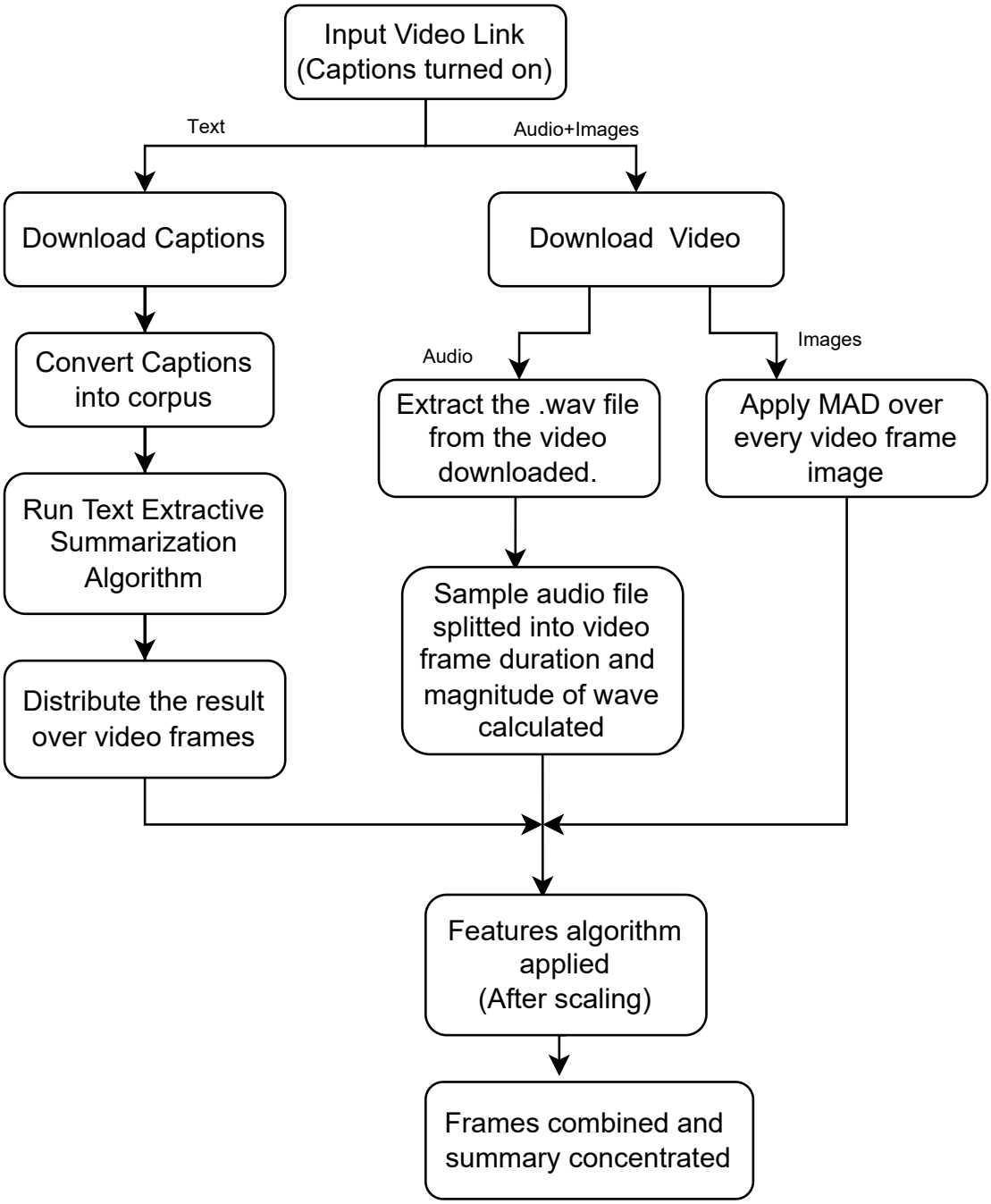


Figure 5.2: Flow Chart of Proposed VSEM Model

is possibly the most straightforward and most often used summarization technique.
In the frequency summarization algorithm, a dictionary with the frequency of occurrence of that word is taken, ignoring the stop words from nltk. In the text rank summarization algorithm, a graph is initialized with the weights corresponding to similarity matrix values. For the hybrid, alongside the similarity matrix values, the occurrence frequency of the words is also considered, as shown in Fig. 5.3. In Hybrid summarization, frequency and text rank summarization are considered. Input is a transcript of the video, and then scores of each sentence are taken using text algorithms. After

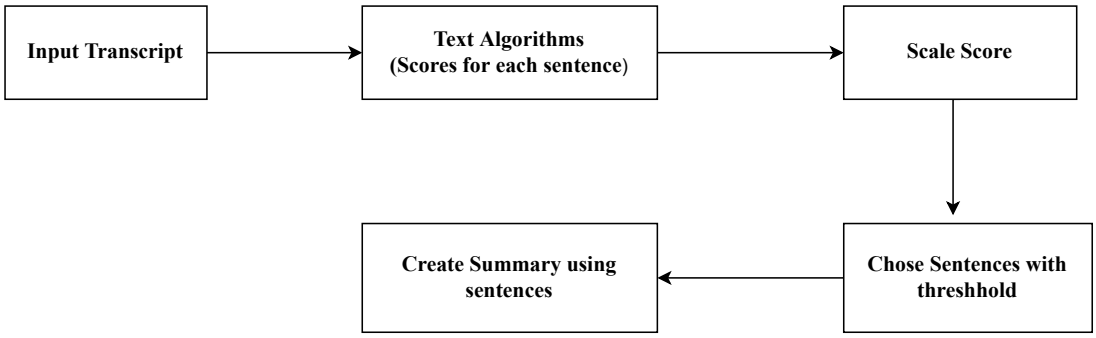


Figure 5.3: Hybrid Summarization

that, scaling is done as shown in an Algorithm 2. Minimum and maximum are taken in scaling; after that, iterations are saved. Sentences are chosen with a threshold.

To verify the effectiveness of the above summarization method, it was tested against

Algorithm 2 Scaling Algorithm

```
Scale_3(lst:list): Input Variable Name - lst , type - list
        Result: Scaling values
min_val = min(1st):                          //The current minimum value in the list
 if (min_val < 0) then
    for i in range(len(lst)): do
       lst[i] = lst[i] - min_val
    end
end
min_val = min(lst)
 max_val = max(lst)
 if (max_val == min_val) then
    for i in range(len(lst)): do
       lst[i] = 0.5
    end
    return 1st
end
diff_val = max_val - min_val                 //Gap between max val and min val
 for i in range(len(lst)): do
    lst[i] = (lst[i]-min_val)/diff_val
end
return 1st
```

Bert-extractive-summarizer from CNN-daily mail news text summarization.

Rouge	Frequency	Text Rank	Hybrid	Bert extractive Text Summarizer
Rouge-1	0.2854	0.2114	0.3207	0.2876
Rouge-2	0.1012	0.0563	0.1272	0.0998
Rouge-l	0.2577	0.1924	0.2955	0.2652

Table 5.1: Rouge Score

Table 5.1 depicts the average rouge score for the algorithms with Bert-extractive-summarizer

for the first thousand instances in cnn-dailymail test data set, where the reference summary is highlighted. Here, it can be seen that the hybrid summary score (rounded up to the 4th decimal place) has the best value here as compared to its components of text rank summarization and frequency summarization, as well as BERT Extractive Text summarization with the threshold of 0.80. Fig. 5.4 illustrates how different algorithms will produce different results for a sentence, with some being more focused on one aspect of the paragraph than others.

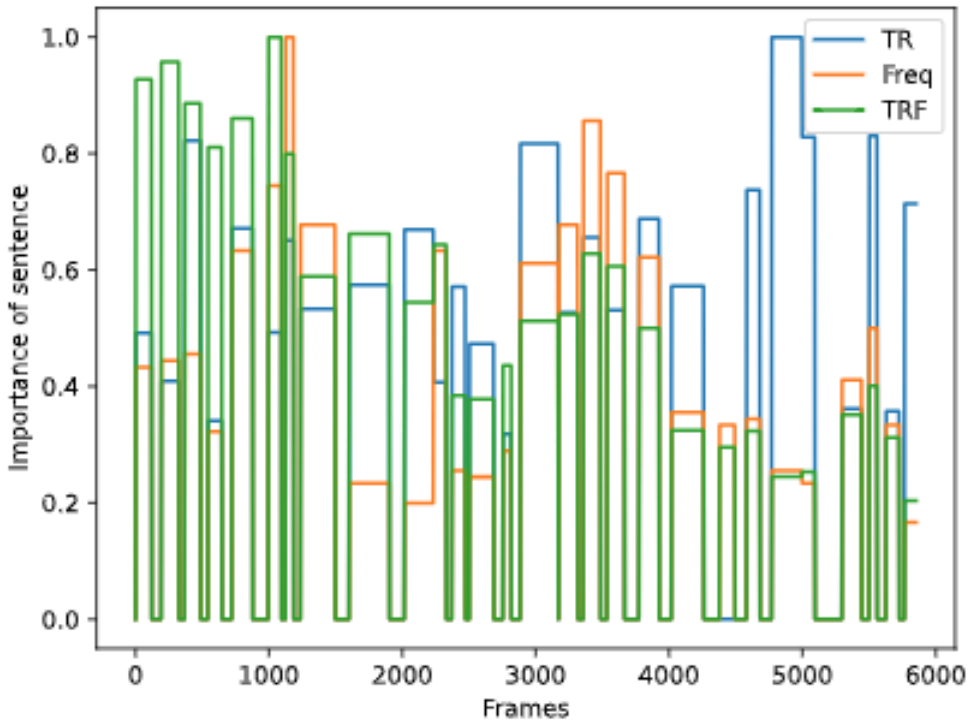


Figure 5.4: Text Score by Hybrid along with its origin algorithms

5.5.2 Audio Separation

Another crucial aspect of a video is its audio, which can be combined with its visual elements to create a powerful summary (112). According to Coutrot et al., (113), the sound will affect viewers visual attention while watching videos, and the strength of this influence varies over time. Subjects will glance in different directions with and without the audio, and the eye fixations gathered during the audio-visual test condition are more concentrated. In comparison to just visual features, audio-visual elements can produce greater results.

To isolate the audio data from the video, the MoviePy toolkit is employed. As each presentation's timeline indicates the alignment of the video, audio and slides, the entire

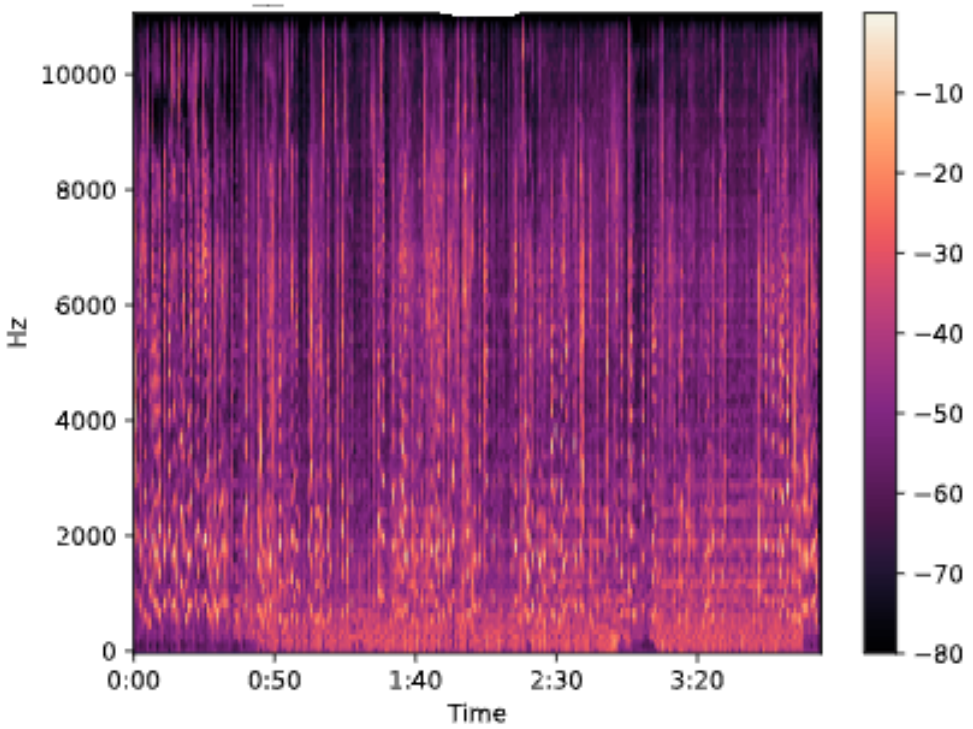


Figure 5.5: Mel Power Spectrogram

audio file is divided into a series of audio segments following the timing of the slide switch, ensuring that each audio clip is correctly aligned to a slides page, originally MP4 speech. After obtaining the wave file Mel Spectrogram, Mel Frequency Cepstral Coefficients, area of audio displayed, and audio amplitude are extracted from an audio chuck of length 0.10 sec (Persistence of Hearing is 0.10 sec.), then the value is distributed over frames. The time vs frequency graph over decibel for the MEL spectrogram of video is shown in Fig. 5.5.

In Mel Spectrogram, the Hertz value is converted to Mel Scale, i.e. instead of taking a regular interval of frequency, the regular interval of pitch is taken, which is better suited when working with human perception models. this is the only reason that audio chunks of 0.10 seconds were picked, as to co-inside with the persistence of hearing. The time versus frequency graph varies over MFCC as shown in Fig. 5.6.
MFCC, as the name sounds, are the coefficients that makeup MFC. They are representations of an audio chunk's cepstral. The nonlinearity of the human hearing system about various frequencies is also taken into account by MFCC. An audio sequence can be divided into separate segments based on the temporal shift of the MFCC, each of which contains music in the same genre or speech from the same individual.
The Mel-frequency cepstrum (MFC) mimics the response of the human auditory system

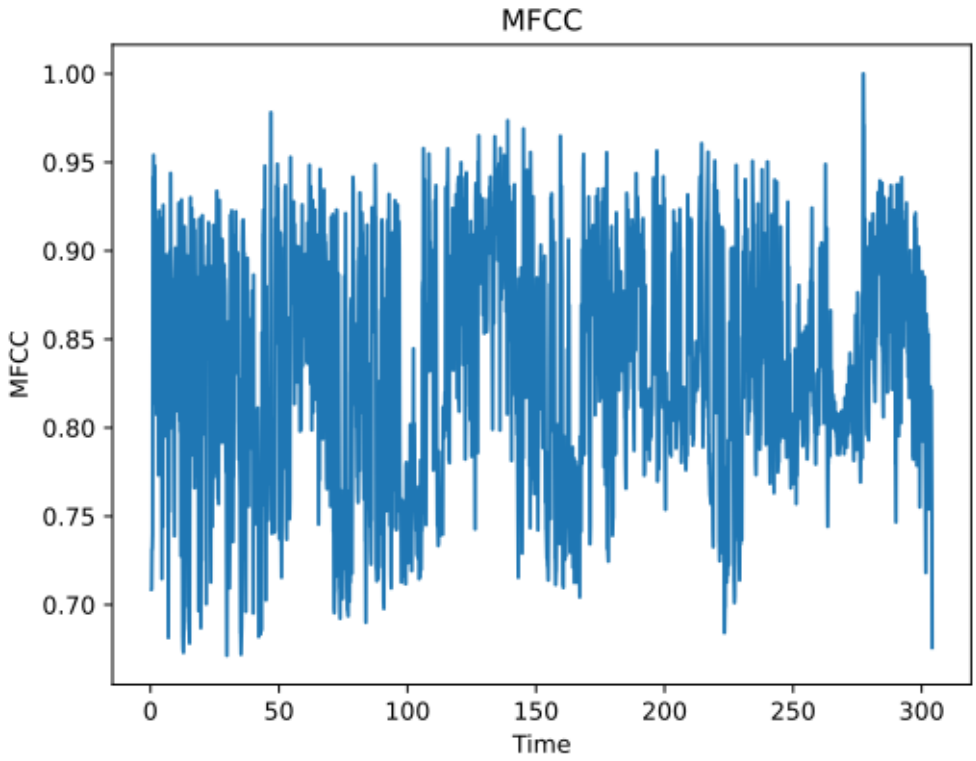


Figure 5.6: MFCC Per 0.1 sec

more accurately than the linearly spaced frequency bands included in the conventional spectrum because the frequency bands of the Mel scale are evenly spaced. To represent the basic signal two properties of the signal were taken, namely Area Under Curve (AUC) and audio peak with audio average as threshold are taken as shown in Fig 5.7.

The AUC graph gives a rather scaled version of the actual audio signal, as the actual signal often contains very high peaks near either extremity due to the abrupt ends, causing noise to take over the signal. If the initial value of the signal is low, then the noise error can be very significant.

The average value of the signal is 0.49968383 after scaling it; due to noise variations in the signal, the end part got an unwanted boost, dwarfing the rest of the signals and making the signal seem flat-line as shown in Fig. 5.8

Though these noise signals (when faced with a low value) can be significant, they are more like an impulse than an actual signal; thus, taking AUC makes them a minor error. Working on this, a max value to represent all the signal chunks (sampled at 0.10 sec.) can be taken to represent the peaks of the signal after removing an average signal value component. An audio peak graph is useful for detecting silence in the given audio file. The audio peak represents the crests, thus the amplitude of the audio signal, and also gives an idea of a zoomed-in top of a sound wave as shown in Fig. 5.9.

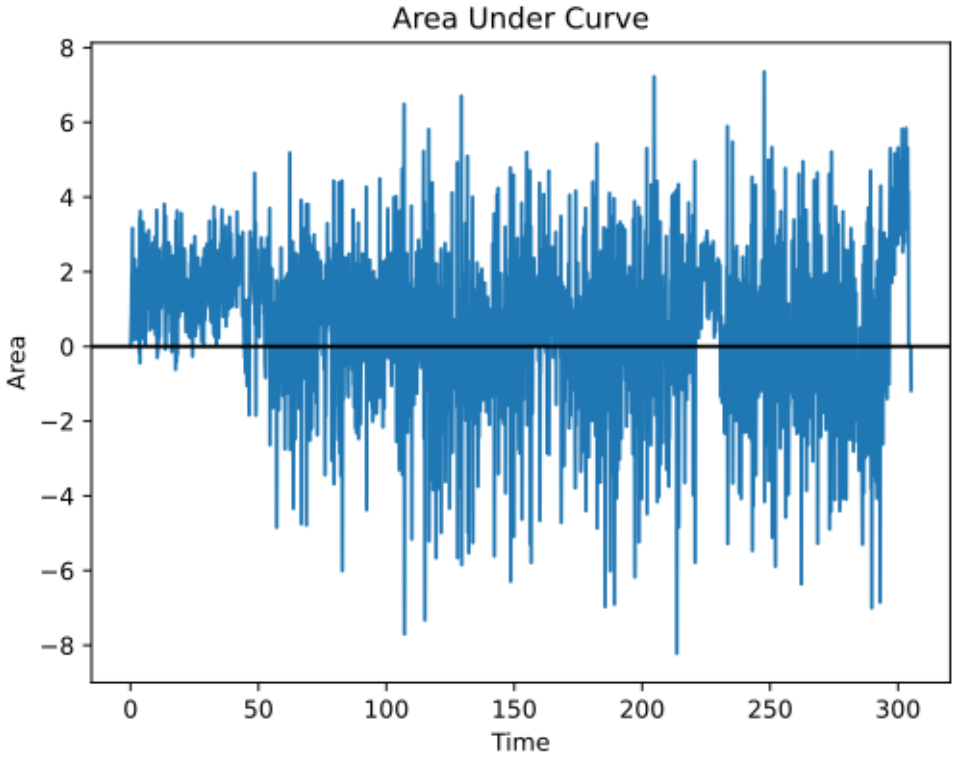


Figure 5.7: Area under Curve

5.5.3 Frame Analysis

A video is essentially moving images with audio; the part moving images is traversed by the term Frames Per Second (FPS). It is independent of image resolution, and audio sample rate (66). It is related to the speed of the video. Thus, a clip with a high rate of change in the frame is more interesting than a clip with a low rate of change in the frame. It should be noted that when recording a video, if the video has a real frame rate of 30 (i.e., 30 frames per second), there is frequently a delay (i.e., 1 percent), making the video's actual frame rate 29.97 (approx.). This frame rate should be rounded back to 30 when processing the video.

For the images (Video Frames), it is to be noted that they can be classified into two types

1) 3D-like colored images which have pixel arrays of form 'm x n x 3'

2) 2D-like grayscale or black-and-white images which only consist of either black pixels or white pixels and have an array of form m x n

In a video, since all the video frames are of the same size, it can be assured that all the video frames are of the same dimensions. Due to this postulate, instead of treating the video frames as an image, they can be treated as an n-dimension array. Finding alterations between two n-dimensional arrays is easier and more precise; For this, the Mean

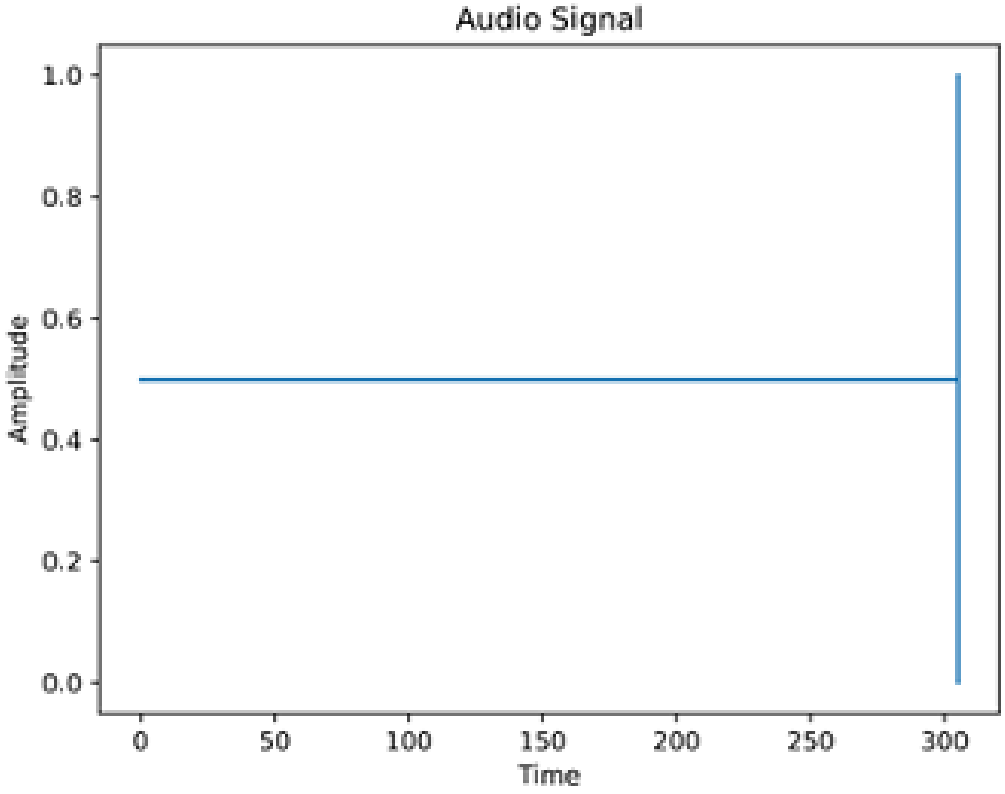


Figure 5.8: Audio Signal

Absolute Difference (MAD) algorithm is used.
Due to the high frame rate, two adjacent frames are often practically a copy of the other, but on closer inspection, they are not, and every frame is distinct, some more than the other, and to find those distinctions, MAD is applied as shown in Fig.5.10.

This MAD is then again used to identify key frames in the video, keyframes are those where there is a rather abrupt change in the image, the object in the spotlight changes, and a new object is introduced. This significant change can be identified by setting a threshold to see when the change spikes. This method is similar to Jenks natural breaks algorithm (114), which in itself is like a variation of the K-mean algorithm. Here, the threshold is the sum of mean and standard deviation over MAD, and it is set as a classifier to detect if there is a keyframe. A higher frequency of keyframes denotes a higher degree of change and movement in the video, as shown in Fig. 5.11.

5.6 Experiment Result and Analysis

5.6.1 Dataset

The experiment is carried out on the TV SUM dataset (95), which is a benchmark data set of video summarization. Fifty structured videos on ten distinct topics were

acquired from YouTube for the TVSum dataset. Videos are professionally edited about news, cookery, education, and others. All videos are of the length of fewer than 10 minutes. The shots are produced by evenly dividing the video into 2-second chunks, and 20 annotations of shot-level relevance scores are included. Fig. 5.12 dataset shows a partial image of the data set.

5.6.2 Evaluation Measure and Results

The TVSum dataset, as previously mentioned, is used to assess the performance of our approach. ROUGE scores are used to evaluate the textual summary. The dataset TVSum is split in a ratio of 20:7 for training and testing, the first 20 for training and the latter for testing. In TVSum, for each video, there are 20 individual human evaluations given. Only the first human evaluation is picked up for all the summaries to avoid ambiguity in the model. The video compression threshold taken is 0.80.
The outcomes of VSEM were compared to those of the following video summarising techniques, which likewise use the TVSum data set. Calculating the F1 measure between the predicted and reference summaries is the most used evaluation strategy. Indicating which frames from the original movie are chosen for the summary, let y_i signify a label with the values 0 or 1 (y_i = 1 if the i-th frame is selected, otherwise 0).

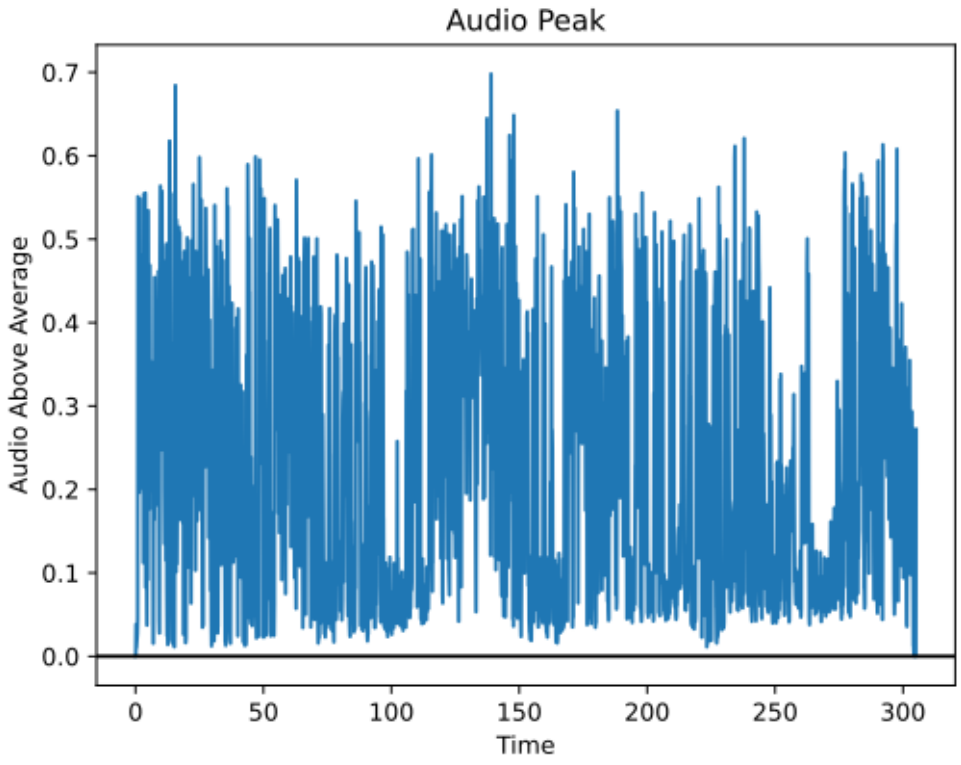


Figure 5.9: Audio Peak

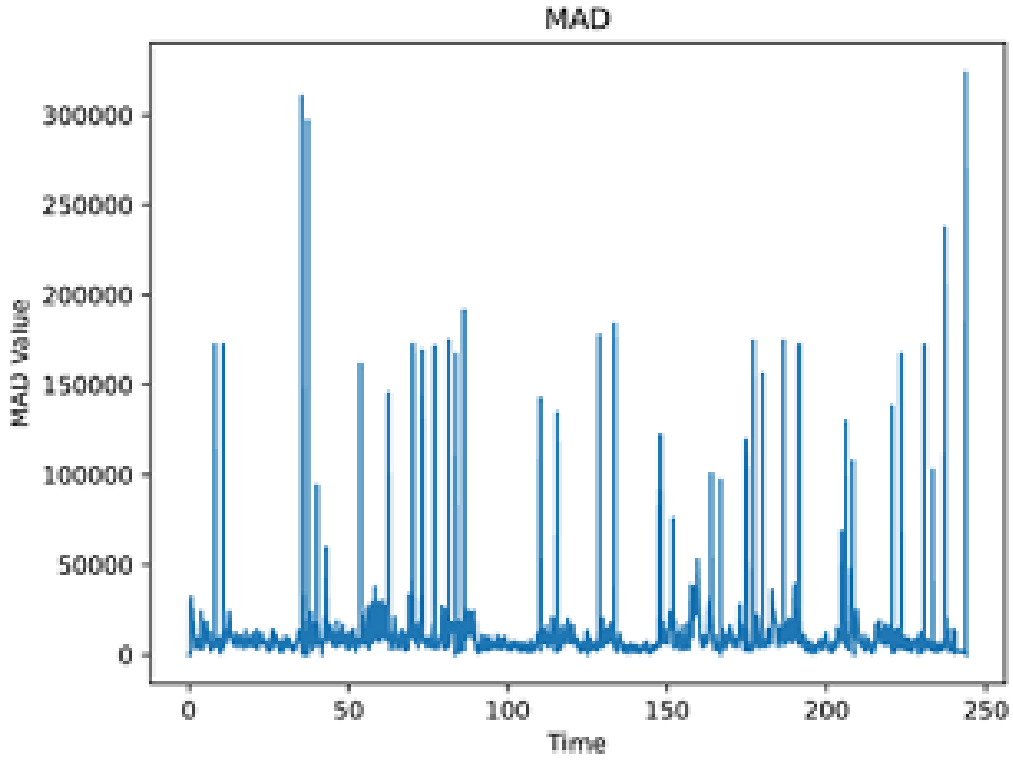


Figure 5.10: MAD for Keyframe Identification

To evaluate the summary's quality, the F-score (F1) is computing as given in Equation (5.4).

$$F1 = (TP + TN)/(TP + TN + FP + FN) \tag{5.4}$$

Where
F1 = Accuracy
TP → True Positive, Frames selected by both predicted and human summary
TN → True Negative, Frames rejected by both predicted and human summary
FP → False Positive, Frames rejected by human summary but accepted by predicted summary.
FN → False Negative, Frames rejected by predicted summary but accepted by human summary.

Mean is the F1 score on TVSum dataset.

The F1 score using different models is shown in Table 5.2. F1 score is found on dataset video and from tresults, it is found that Stochastic Gradient Descent Regression is the most suitable Regression available, with the highest mean score of 0.69136 (out of 1.0) and ridge regression with a mean score of 0.68946 (out of 1.0). So, we have

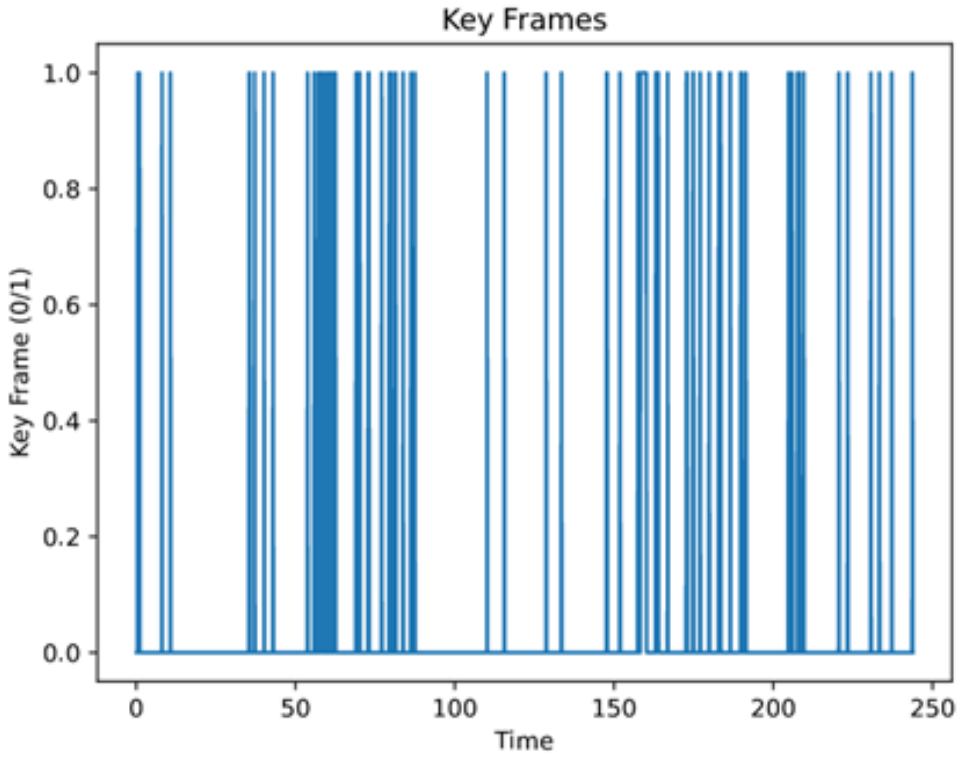


Figure 5.11: Frequency of key frames

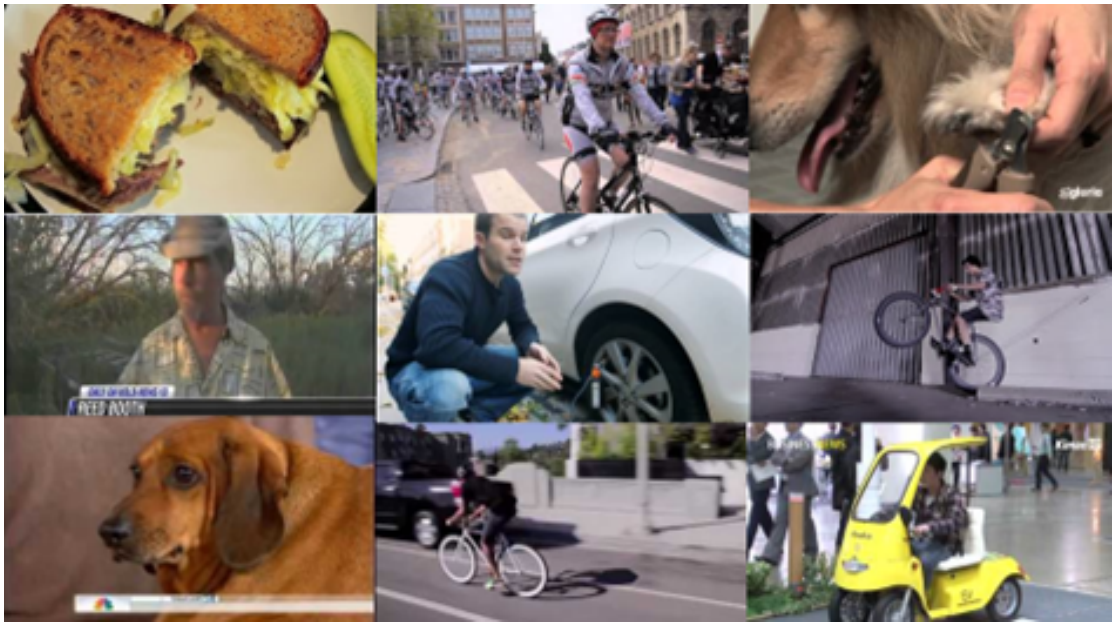


Figure 5.12: Partial Image of Dataset

used a combination of these two regression algorithms in the hybrid model. SVM and decision tree classifiers are not used in this. SVM performs worse when more features than training sets are available; hence, it is not appropriate for large datasets. Due to

Model	Maximum	Mean	Minimum
Linear Regression	0.73175	0.68529	0.65467
Stochastic Gradient Descent Regression	0.78787	0.69136	0.60714
Elastic Net Regression	0.70909	0.67458	0.64285
Ridge Regression	0.73333	0.68946	0.65467
Lasso Regression	0.70909	0.67344	0.64285
Random Forest Regression	0.71794	0.67584	0.62524
Gradient Boosting Regression	0.73214	0.68542	0.64891

Table 5.2: Average F-measures using Different Models

the abundance of trees, the performance of the summarization approach employing the decision tree classifier is poor. So, even a minor alteration to the decision tree could significantly impact prediction accuracy. Average F_measures is shown in Fig 5.13

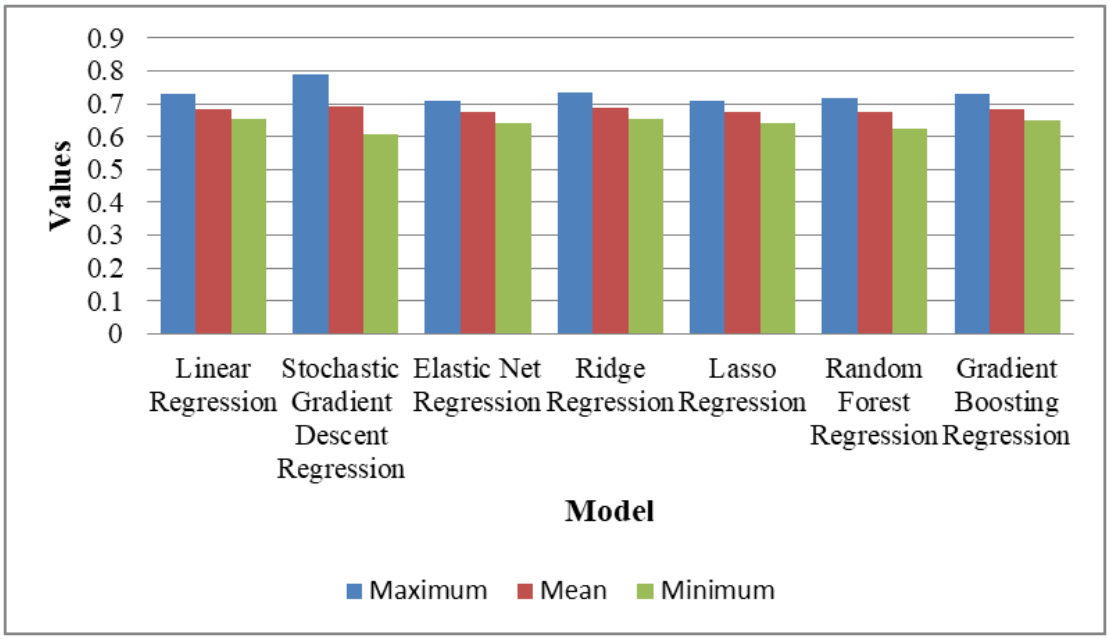


Figure 5.13: Average F measures

The F-measures for each approach using a machine learning model for the video's summary in the database are shown in Table 5.3. The table shows that VSEM outperforms the assessed methodologies, delivering competitive outcomes while retaining a balance between pace, duration, and quality.

The proposed algorithm is applied to different videos of the TVSum dataset, and it is found that there is a significant difference in the run time of the original video and video summary. The original video was 5 minutes and 54 seconds long, while the summarised video is 72 seconds long. Instead of watching the entire video, the consumer may see the summary, which saves time. It is found that VSEM saves 75-80 percent of the user's time.

Method	Year	F-Measure
M-AVS (115)	2017	61.0
VASNet (30)	2018	61.42
DSNet (116)	2020	62.1
PGLSUM (21)	2021	61.0
RRSTG (117)	2022	63.0
VSEM	**2022**	**69.6**

Table 5.3: Average F-measures of the Summaries Generated by each Technique

5.7 Summary

To improve the effectiveness of the summary, a hybrid model (VSEM) for the video summarising problem is proposed in this study. The suggested method's main contribution is:

1. Hybrid text summarization is proposed using text and frequency summarization and is compared with the state-of-the-art methods. The proposed hybrid text summarization shows better results.

2. Audio and image features are combined with text, and a hybrid model is proposed. The experimental results on TVSum show that the multimedia components—text, audio, and image can offer the summary task more information and accuracy than a single visual feature.

The performance of hybrid method can be achieved by providing more contextual understanding and reducing the possibility of evaluation bias. As these constraints could lead to video summaries that are less coherent, useful, and flexible than methods that make use of various modalities, as we cover in our upcoming chapter.

Publication

The work discussed in this chapter is published in:

Shambharkar, P. G., and Goel, R. (2023). Auto encoder with mode-based learning for keyframe extraction in video summarization. Expert Systems, 40(10), e13437.

CHAPTER 6

TAVM: A Novel Video Summarization Model based on Text, Audio and Video Frames

In the contemporary digital landscape, the task of video summarization has gained immense importance within the realm of multimedia analysis. This relevance is largely driven by the exponential expansion in multimedia content consumption, encompassing audio, video, and images, which is readily available on-demand through various digital platforms. Automatic video summary is the process of creating a brief synopsis that summarises the video by displaying its most useful and relevant elements, so consumers may rapidly comprehend the primary concept of a video without having to watch the entire material. Currently, the selection of the video segments to be included in the final summary is done in a variety of ways. The task is to analyze the video's multimedia data in search of relevant clues that will aid in decision-making. The proposed method TAVM (Text, Audio, and Video mode) will provide the video summary using different multimedia elements of text, audio, and frames. The proposed TAVM method can be separated into three parts. The process begins with video processing, where the BEiT vision transformer is employed to recognize objects within the chosen frame. Following that, Audio Processing comes into play, which uses speech-to-text converters to transcribe the audio content. Finally, in the last step, the Summary Builder utilizes the GPT-3-based OpenAI API to generate a summary of the content. The experimental analysis on the benchmark dataset SumMe demonstrates the effectiveness of the proposed approach.

6.1 Introduction

The fast expansion of multimedia data transmission over the internet requires the summarization of all the data. Multiple information modalities are used in video streams to communicate information. It can be challenging for users to efficiently gather crucial information since, for instance, visual events can comprise objects, gestures, and scene changes, auditory events might be changes in audio sources, and textual events can contain conversations, subjects, and key phrases. It is difficult for humans to get important information from all the data effectively. Manually extracting interesting parts

from video footage and processing them are time-consuming operations. Thus, there is a need for automated approaches to reduce duplication and extract usable information (13). Video summarization has emerged as a challenging challenge that aims to automatically analyze video footage. Video summarization gives the viewer a condensed version of the video that, incorporates all crucial details for comprehending the subject matter (118).

With the rise of video content as the primary source of data consumption for information, automation of the video summary process has taken centre stage. Video summary applications can be helpful in creating highlights for sporting events, movie trailers, medical diagnoses, and many other real-life applications (119). Different media organizations, such as sports or entertainment videos, develop teasers or previews for films and TV shows using highlights of events. Furthermore, video search engines can leverage video summarization for video indexing, browsing, retrieval, and recommendation (120).

Multimedia data are diverse and include more complicated information than plain text also it has a difficult time bridging the semantic gap between various modalities (34). A video has a hierarchical structure that records spatial and temporal data as frames, shots, and scenes. It is not possible to watch a video all the way as it would take a lot of time and effort. The process of producing a video summary from one or more raw videos is automated by video summarization. Whether static or dynamic, a video summary represents standardized content or user preferences (32).

Text and audio modality can help the visual modality comprehend the video's structure and content more thoroughly. To put it another way, the text, audio, and vision support the actions that are being done in the various modalities. For instance, the subtitle file as text helps to understand the video better, the wedding music conveys the festive mood of the setting, while the roars of support during cricket denote a successful run. However, videos frequently experience audiovisual and text inconsistencies as well. An illustration would be that the sounding item is not visible. The fundamental difficulty in text audiovisual video summarization will be exacerbated by interference with the visual modality. As a result of the potential and difficulties mentioned above, we suggest using a Text, Audio, and Visual model (TAVM) to combine text, audio, and visual data for video summarization.

6.2 Strategies Employed in the Architecture

Fig. 6.1 shows the proposed architecture in which video frames, text, and audio are taken from the input video. Audio and video processing are done separately, and mapping is done finally. The goal of the proposed approach is to choose the most informative video frames using text, audio, and keyframes. Given a video sequence X = [x1,

x2,..., xn], where x is the whole video and n is the total number of frames, The idea is to find a video summary.

$$Y = [(y1 + y2) + y3] \tag{6.1}$$

where y1 is selected text, y2 is audio, and y3 are important and non redundant keyframes. Combined Y1 and y2 processing is done, and y3 processing is done separately, and finally, all are combined to form Y. Y is a total video summary whose duration is less than X.

Algorithm 3 Steps for Proposed Approach

1. Input Video:
-Separating audio and video frames
2. Video Processing:
-Break the video into video frames.
-Extract the frames and remove redundant frames.
-Select candidate frames for summarization.
-Apply a vision transformer(e.g. BEiT) to identify objects in selected frames.
-Apply a Sparse Encoder to extract important features from the frames.
-Use K-means geometric filter to cluster similar frames based on object features.
-Obtain the object type, start time and end time of keyframes.
3. Audio Processing:
-Use a speech-to-text converter to transcribe the audio.
-Preprocess the text to remove stop words and extract important keywords for a summary generation.
4. Summary Builder:
-Use GPT-3 based OpenAI API to summarize the extracted content and keywords.
5. Mapping:
-Map the keyframes and summarized text to generate a concise and informative video summary.

6.2.1 Input Video

Videos are often regarded as the most informative and appealing way of data/content display. The input video is taken in MP4 format from the SumMe dataset and audio and video frames were extracted from the video. The input video's frames are divided into subsets by the shot segmentation algorithms, and each subset is made up of a collection of related frames that appear one after the other. Each subset's beginning and last frames signify a shift in content between shots. A significant content shift can be seen by keeping an eye on the size of the motion vectors between adjacent frames (121).

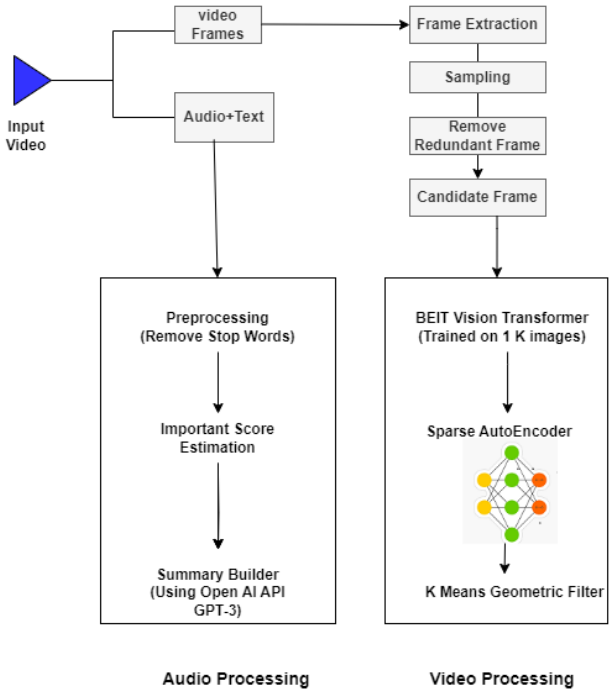


Figure 6.1: Proposed Architecture

6.2.2 Video Processing

More efficient methods are always urged because video summarising consumes a lot of computing time. If every frame in a video is looked at for possible selection, the summarising process can take a while, and processing resources are lost on redundant or similar frames. To expedite the process frame consideration is important. The video is broken into video frames. The video contains duplicate frames and these redundant frames needlessly lengthen the output video summary, making it less efficient. The amount of individual frames displayed to viewers is represented by frame rate, which is measured in frames per second (FPS). The sampling process seeks to remove superfluous frames by picking specific frames from the available video (122). Frames were extracted, redundant frames were removed and candidate frames were selected for summarization. Candidate frames are chosen from the resulting collection of frames using local thresholding of the magnitude of displacement vectors between successive frames. BEIT (BERT Pre-Training of Image Transformers) Vision Transformer is applied to identify objects in selected frames. Vision Transformer breaks a picture into fixed-size patches, embeds, adds positional embeddings and feeds the resultant vector sequence to a conventional Transformer encoder (123). BEiT models predict visual tokens using supervised pre-training from OpenAI's DALL-E codebook (124). A sparse encoder is applied to extract important features from the frames. The sparse Auto encoder provides a composite representation of feature vectors taken from CNN (125).

The SAE encoder feature vector contains 500 dimensions. K means geometric filter was applied to cluster similar frames based on object features and the object type, start time, and end time of keyframes are obtained. The K means method employs comparable pixel clustering and median placement (126). Similar frames were clustered based on object type, start, and end time.

6.2.3 Audio Processing

When the audio is merged with the video, it is possible to fully comprehend the information in the video. Audio and summarization account for elements of an audio stream that are helpful, attract human attention and offer the overall notion or concept of the stream. For audio processing audio and text are combined. A speech-to-text converter is used to transcribe the audio. First, preprocessing is done, in which stop words are removed and an important score is estimated. The input text is split into smaller chunks and a request for a summary is made. The length of the summary can be controlled using max_tokens. The audio summary is beneficial and helps people with visual difficulties.

6.2.4 Summary Builder

The audio-visual textual format of a video summary attracts the user's attention because it is more descriptive, simple to grasp, and graphic in nature (32). The video summarization method is used to extract important information from a video and construct a summary using keyframes or key shots.
The resulting video summary comprises multi-modal information and might be static or dynamic in nature. Summary of text data is built using GPT-3 based open AI because it needs a modest quantity of text as input to produce huge amounts of accurate and complex machine-generated text. Preprocessing of text is done and summary generated for each input chunk can be controlled and output summaries are concatenated to form a single string.

6.2.5 Mapping

Keyframes, summarized text, and audio are mapped to generate a concise and informative video summary. A video's start and end times are calculated, along with the time range of each statement. This data produces video chunks, which are subsequently assembled into the whole video.

6.3 Experimental Result and Analysis

Experiments were carried out using the benchmark dataset SumMe (127). The SumMe dataset comprises 25 videos on various subjects like cuisine, travel, sports, and so on. The majority of these are unedited videos. For each video, 15-18 users are employed to choose key shots and write video descriptions, and a total of 41 subjects participate. The annotation is binary, indicating whether or not a frame should be included in the summary. Each video is between one to six minutes in length.

6.3.1 Evaluation Metrics

Precision, Recall, and F-score are used to assess the efficacy and efficiency of the proposed strategy and to assess the summary quality by comparing the temporal consistency of produced versus human-created summaries as shown in Eq. (6.2), Eq. (6.3), Eq. (6.4). The following metrics are defined using the temporal overlap of the predicted summary Ps and the actual summary As. The F-measure is a metric that combines both Precision and Recall into a single value, providing insight into the accuracy of the experiment (128).

$$Precision(P) = \frac{Overlap(Ps, As)}{length(Ps)} \tag{6.2}$$

$$Recall(R) = \frac{Overlap(Ps, As)}{length(As)} \tag{6.3}$$

$$F = 2P * R/(P + R) * 100 \tag{6.4}$$

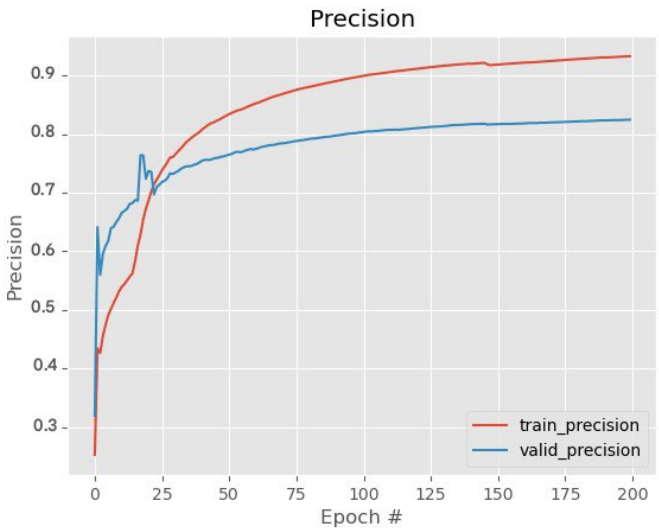


Figure 6.2: Precision

We have the following model performance on the SumMe dataset. Precision=0.974 as shown in Fig. 6.2 and Recall=0.977 as shown in Fig. 6.3 score and F1 Score=0.975

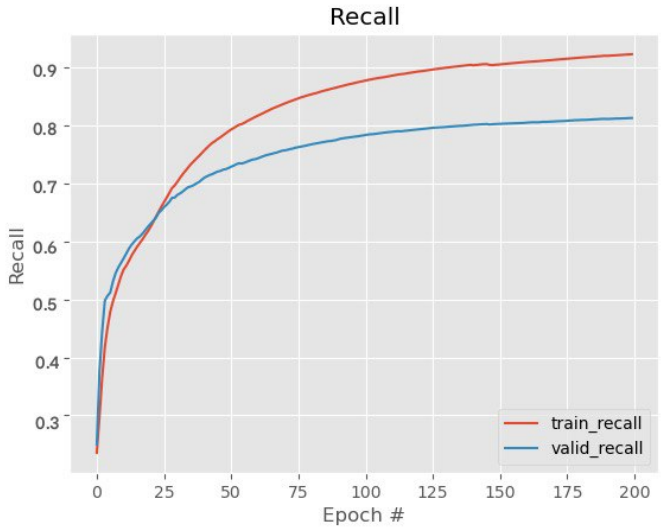


Figure 6.3: Recall

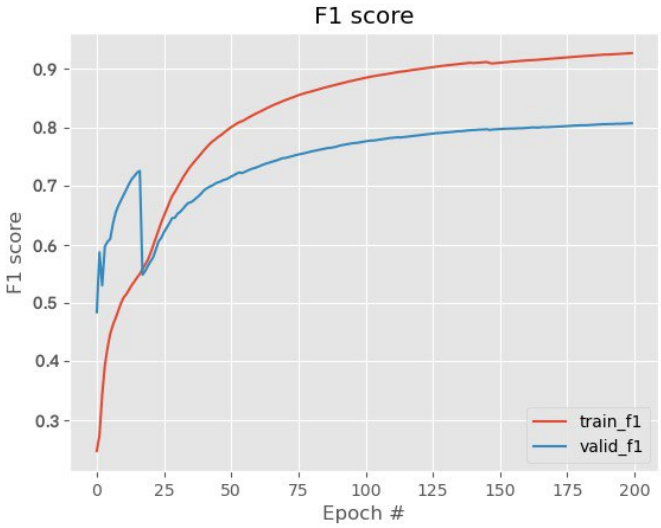


Figure 6.4: F1 Score

as shown in Fig. 6.4 for 200 Epoch was achieved on the proposed model. The experimental analysis of the SumMe dataset demonstrates the effectiveness of the proposed approach with an F1 score of 0.975. A higher F-measure indicates greater accuracy.

6.4 Conclusion

In a new technological era, it is simple for people to express their thoughts on various platforms. These platforms enable individuals to express themselves using a variety of representations, such as text, photographs, videos, and audio. Due to the subjective nature of video summarization, people have different preferences for summaries, which is the biggest challenge in video summarization. Automatically creating a summary for asynchronous data can assist consumers in keeping up with the increase of multi-modal content on the Internet. Due to the significance of video summarization in numerous domains, the study is growing rapidly, motivating scholars to investigate and create novel concepts and techniques to achieve the desired outcome. Travel videos are one of the most important choices made by consumers. uploaded videos address natural tourism, culture, gastronomy, and other things that may relate to the viewer on several levels. The narrative and visual presentation of tourist places using high-quality video increases the possibility of the audience developing good connections with the destination. Due to time constraints, it is not possible to watch the whole video, so the proposed model's performance on the SumMe dataset, will be applied in the travel department in real-time scenarios. The suggested method's contributions are:

1. The proposed model on the SumMe dataset demonstrates how the combined text, image, and audio aspects improve summarising information.

More accurate and comprehensive summaries can be obtained by combining different models and taking into account a broader range of multimedia elements and cues as discussed in the next chapter.

Publication

The work discussed in this chapter is published in:

Shambharkar, Prashant Giridhar, and Ruchi Goel "TAVM: A Novel Video Summarization Model Based on Text, Audio and Video Frames." In 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 878–882. IEEE, 2023. x

CHAPTER 7

Keyframe Extraction via Peak Wave Analysis with Integrated Human Presence and Face Recognition

Automatic video summarization serves the purpose of generating a concise and informative overview of lengthy video content. This summary plays a significant role in effectively categorizing and organizing videos within a video database. In this work, an innovative artificial intelligence (AI)-driven approach to address the challenges of identifying keyframes and summarizing video content efficiently is proposed. The methodology involves the integration of three models: Model 1 for human presence detection (HPD), Model 2 for face presence detection (FPD) using Peak Wave Time analysis (PWT) and Model 3, an advanced YOLO V face recognition model. The synergistic integration of these approaches is intended to meet the issues faced by the growing volume of video data. The proposed model aims to contribute to the field of video summarization by enhancing accuracy and reducing time requirements through a multi-model approach combining audio and video.

7.1 Introduction

The worldwide video market is becoming more and more prominent. Massive volumes of data are being generated on various social media platforms, such as Instagram, Facebook, YouTube, and others. Additional platforms, including news, sports, entertainment, and CCTV footage, further enhance this data flood. Videos have several duplicate events that viewers might not find interesting.The installation of CCTV cameras for tracking, security, and monitoring purposes has made this especially pertinent to many sectors. The massive volumes of data collected by surveillance videos—captured around the clock—make it difficult to identify individuals or incidents. Large volumes of video data bring on two main issues: 1) The incredibly high cost of data storage; and 2) The challenge of retrieving important information from video. Removing this extra content and concentrating on the important stuff will help viewers organize their memories and save time. The continuous surge in video data necessitates efficient video summarization methods. In the modern world, it is essential to summarise the main points of a long video. To overcome this difficulty, video summarization offers a condensed

summary that saves time and memory (129). Video summarization aims to provide a condensed depiction of the original video while maintaining its key elements (130). The user can efficiently organize and navigate large amounts of videos with the help of a video summary (131). Video summarization is conducted through two methods: static, which concentrates on important frame extraction, and dynamic, which involves video skimming by uniform sub-sampling (132). Authors (133) have proposed an approach that enhances content protection efficiency by combining deep learning models. In order to generate static keyframes, one or more sample keyframes from each original video shot must be chosen. Alternatively, time-ordered picture sequences with corresponding auditory information are combined with dynamic video skimming. The goal of these common techniques for video summarising is to produce summaries that are appropriate for every viewer.

Keyframes capture main objects and events, while video skimming, particularly through fast-forward methods, adjusts the frame rate for quicker content review. In contrast to keyframe extraction, video skimming can result in a more realistic and informative summary, but it can also create artifacts, distortions, or inconsistencies in audio synchronization, narrative flow, or video quality. Static summarization offers a more condensed and synchronized output, which is beneficial for browsing, indexing, and video analytics applications, addressing the challenge of processing numerous videos (134). Selecting the most important semantic frames or scenes from the source video while keeping the particular domain, like a person or object, in mind is necessary to create a representative video summary. A good video summary sums up the main points of the video and offers a concise synopsis as an alternative to seeing the entire thing (46). Efforts are focused on generating summaries that are both objective and comprehensive. Using summaries improves the effectiveness, usability, and accessibility of visual information. It makes it more controllable, which benefits both content producers and consumers. Important information in a video is represented by keyframes. They are also referred to as R-frames, representative frames, still-image synopses, and a group of noteworthy images that were extracted from the video (135). The process of choosing keyframes is important in video summarization because key moments, such as a player scoring a goal or the crowd applauding, are highlighted in sports video summaries. On the other hand, in schools or colleges, lengthy videos have been made by authorities during any event in campus. Every parent wants to see their child's performance or picture in that lengthy video. Every family member also wants to see their child's performance only instead of watching whole videos.

In short, people have different preferences for video content. The main difficulty lies in selecting keyframes so that redundant frames are not chosen, as crucial frames for one individual may not be relevant to others. A good video summary must be non redundant relevant and coherent (136).

This research proposes an AI-driven solution to automate key frame identification, introducing advancements in human presence and facial recognition using peak wave analysis. Peak wave analysis is a signal processing technique used to identify and analyze significant points, or peaks, in a waveform. It is commonly employed in various domains, including image processing and video analysis, to detect salient features. The analysis involves identifying points where the amplitude or intensity of a signal reaches a maximum, indicating points of interest. In the context of video summarization, peak wave analysis plays a crucial role in identifying keyframes—frames that capture important moments in the video.

7.1.1 Key Contributions

7.1.1.1 Precision in Keyframe Identification

The sequential application of models ensures precise identification of keyframes, reducing false positives and enhancing the accuracy of the summarization output.
Models focusing on human presence, face presence, and facial recognition collectively contribute to a comprehensive understanding of video content.

7.1.1.2 Efficient Time Utilization

The selective frame extraction process, combined with dynamic peak wave function analysis optimizes computational resources. The efficient utilization of time is further emphasized by the sequential and hierarchical application of models, streamlining the summarization process.

7.1.2 Combining Audio and Video

The inclusion of audio-visual features is a crucial aspect of the proposed model. Combining the two modalities, the approach aims to provide summaries that are more engaging and informative than those generated by traditional video summary techniques, which primarily focus on visual cues (75).

7.2 Strategies Employed in the Architecture

It's crucial to choose and combine representative and significant video segments in order to remove redundancy from the original video while summarizing it. One of the biggest challenges in artificial intelligence is automatically identifying keyframes using person detection or key part detection in any video. To effectively identify key parts in video footage, this effort involves both programmers and machines working together to train the system. This will help in building more efficient approaches for summarising

video content. The proposed approach can be broadly categorized into the following stages: a) Frame Extraction b) Audio Extraction c) Summary Generation. In the proposed approach, input video is taken from the benchmark dataset TVSum. Frames and audio are extracted from the input video. The whole approach is shown in Fig. 7.1.

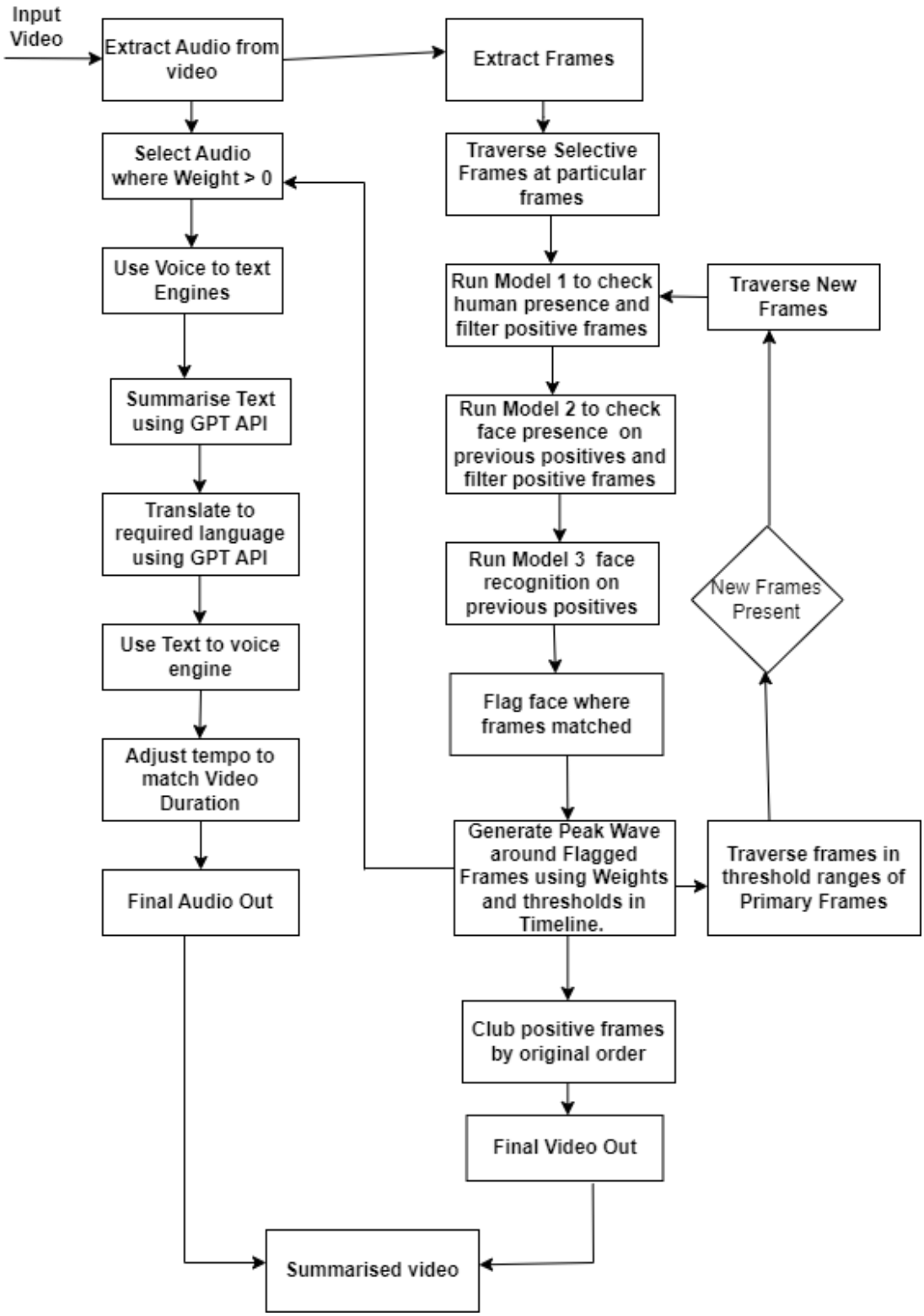


Figure 7.1: Video Frame Analysis

Selective frames are traversed at particular intervals. First, human presence and then facial presence are checked for that specific video. Peak wave rounds flagged frames are generated using weights and thresholds in the timeline. The peak wave is used to identify the frames where the particular face has been detected. For audio, that audio

is selected where weight 0. The uniqueness of the proposed approach is that voice-to-text engines are used, and then text is summarised using the GPT API. The tempo is adjusted to match the video duration. The final audio and video make a combined video summary.

7.2.1 Frame Extraction

Key-frame extraction is the process of identifying and preserving only those frames that accurately represent the video's original content (137) (138). The main idea behind key frame extraction is to reduce the enormous number of video frames from the whole video to only a few relevant frames (139). For the given video, frames were extracted at intervals from the input video.The frame extraction process as shown in Fig 7.2 involves utilizing the OpenCV library to capture frames. The cv2.videocapture function

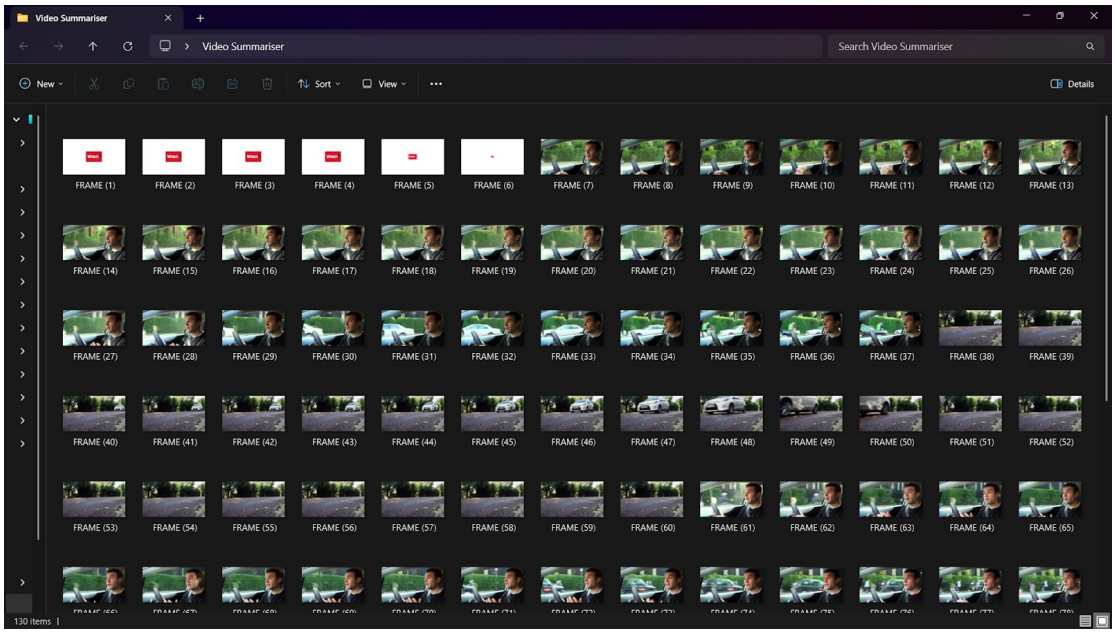


Figure 7.2: Frame Extraction from Sample Video

is employed to open the video file, and then frames are read using the read() function. The selective extraction is achieved by traversing the video at particular intervals and specifying the desired frame rate. Frames selected from the extracted frames are used as primary frames, as shown in Fig. 7.3. Here every fifth frame is chosen for reference, and frame selection criteria can be set by the user. The selective extraction of frames at specified intervals optimizes computational resources.

Secondly It allows for efficient processing, focuses on relevant segments, and reduces redundancy in frame analysis. Model 1 is applied to detect human presence, filtering positive frames. Subsequently, Model 2 refines positive frames by checking for face presence. Finally, Model 3, utilizing YOLO V, performs face recognition on the refined frames, flagging frames where a face is detected. Peak waves are generated around

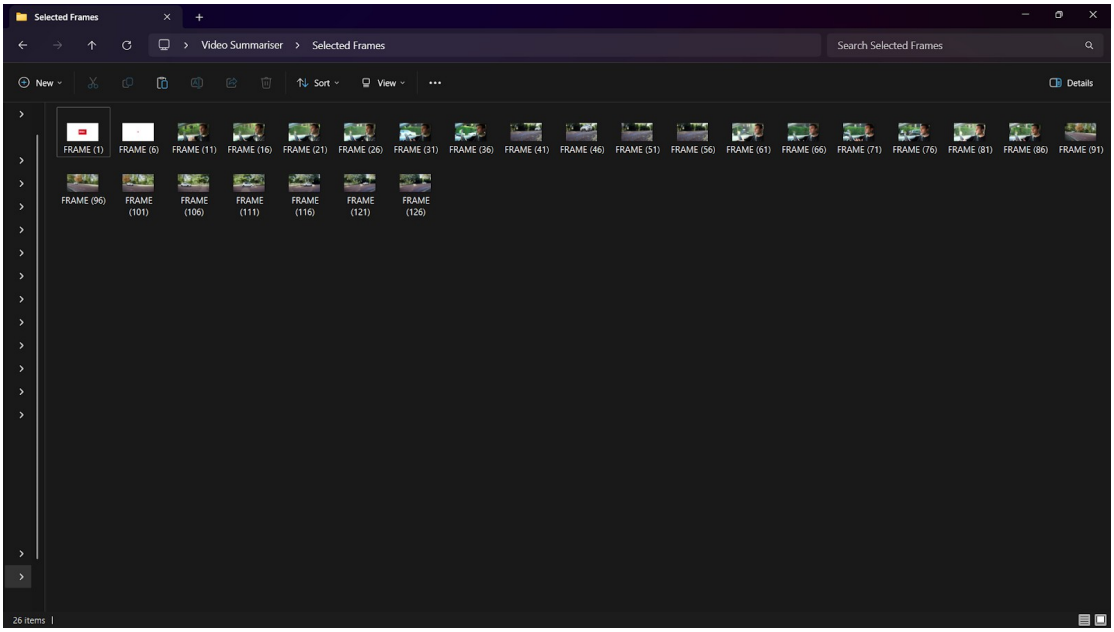


Figure 7.3: Frame Selection

flagged frames using weights and thresholds, facilitating efficient identification. It can be quickly trained to recognize diverse objects and has great generalization capabilities (140). The default weight is set to 0, and frames are traversed in threshold ranges for further analysis. If new frames are present, the process iterates, ensuring comprehensive coverage. Positive frames from the first iteration of frames through all three models sequentially (these frames are marked with positive weights in the peak-wave-time analysis) are shown in Fig. 7.4. A visualization of peak wave analysis is shown in Fig 7.5. Peak wave analysis is a signal processing technique used to identify and analyze significant points, or peaks, in a waveform (141). Peak wave analysis plays a role in enhancing the efficiency and effectiveness of video summarization by identifying and prioritizing significant moments within the video content. ”peak wave” refers to the peaks or high points in the visual or content-related aspects of the video. By analyzing peaks in the signal generated from video frames, the algorithm can pinpoint frames where specific events, such as the presence of a person or a recognizable face, occur.

The frame rate in the final output video has been adjusted to 24 frames per second (FPS). The number of video frames captured in a second is known as frames per second or FPS. This adjustment aims to balance visual quality and storage efficiency, ensuring essential frames are retained for a concise summary.

7.2.2 Audio Extraction

The audio processing phase of the proposed approach plays a crucial role in enhancing the overall video summarization experience. This phase involves extracting meaningful

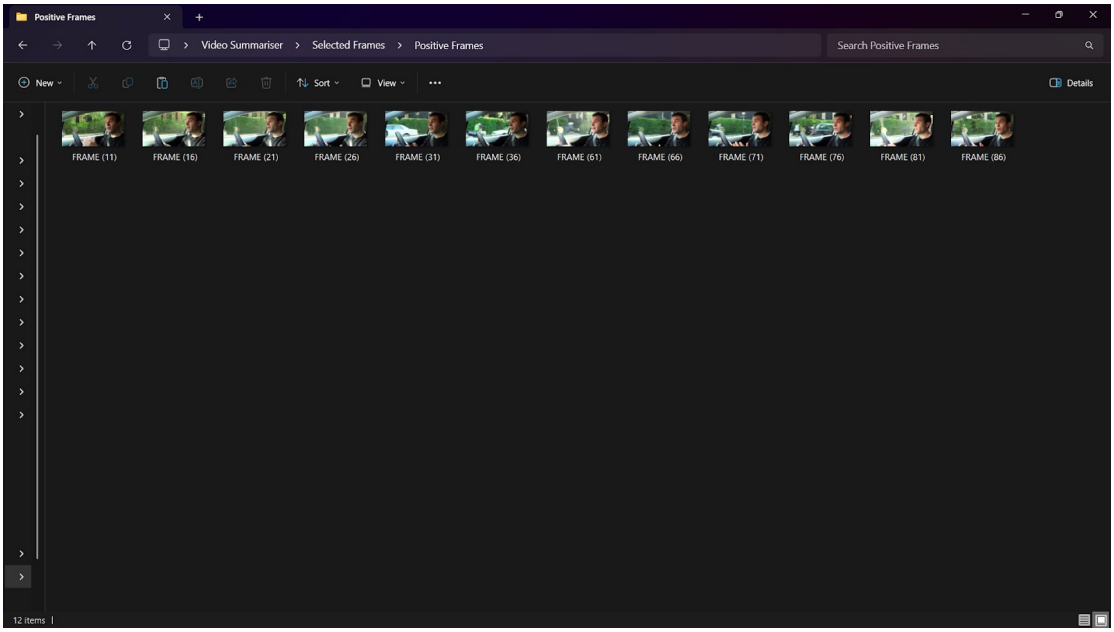


Figure 7.4: Frame Iteration

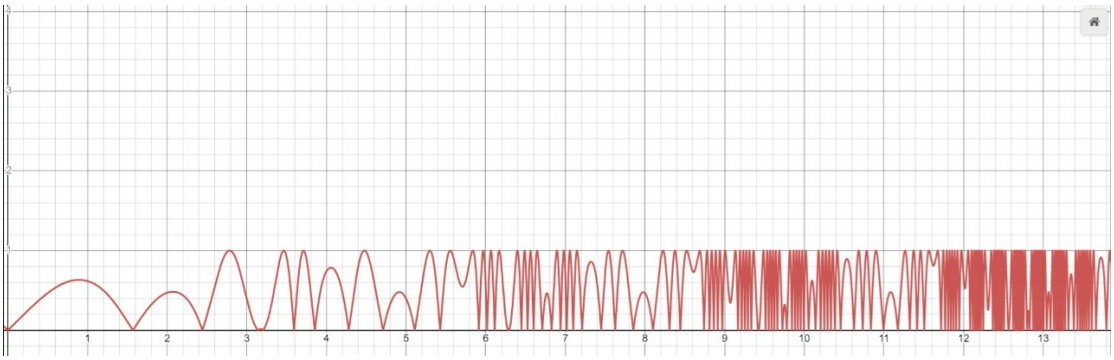


Figure 7.5: Peak Wave Analysis

information from the audio track of the video and incorporating it into the summarized content. Audio is extracted from video frames, where the associated weight is greater than 0. This weight parameter acts as a filter, ensuring that only relevant audio segments are considered for further processing. The audio extraction process is facilitated by leveraging OpenCV library, which enables the extraction of audio frames corresponding to the identified positive video frames.

7.2.2.1 Voice-to-Text Conversion

Once the relevant audio frames are extracted, voice-to-text engines are employed to convert the audio content into textual form. This conversion allows for the transformation of spoken words into a format that can be analyzed and summarized effectively.

7.2.2.2 Text Summarization using GPT API

The extracted text undergoes a summarization process using the GPT API. The reason for using it is that the model is able to anticipate words and produce cohesive writing by using the system to guess which word will best fit the context given on the left (142). The GPT API's natural language processing capabilities contribute to creating meaningful and contextually relevant textual summaries.

7.2.2.3 Translation

The summarized text can be further translated into the required language using the GPT API. This step ensures that the generated textual content is accessible to a broader audience, overcoming language barriers.
Translation enhances the inclusivity of the video summarization output, catering to diverse viewership.

7.2.2.4 Tempo Adjustment for Synchronization

To achieve synchronization between the audio and video components of the summary, tempo adjustment is implemented. The tempo is dynamically adjusted to match the duration of the video. This ensures that the audio narration aligns seamlessly with the visual content, providing a coherent and synchronized viewing experience.
The audio processing phase adds a layer of richness to the video output by incorporating spoken content into the textual summary. Voice-to-text conversion, language models, and translation services enhance the accessibility and global appeal of the summarized content. The tempo adjustment ensures a harmonious fusion of audio and visual elements in the final video summary (143).

7.2.3 Summary Generation

The final summary is created by combining keyframes and audio.

7.3 Experiment Results and Analysis

The effectiveness of the proposed approach is assessed using the TVSum dataset.

7.3.1 Dataset

50 different kinds of videos, including news, documentaries, and vlogs, at 30 frames per second are included in the TVSum dataset (57) The section demonstrates the efficacy through various experiments and comparisons conducted to validate the suggested method. The average video length is 4.2 minutes, with a range of 2 to 10 minutes.

In the experiments, all videos undergo down-sampling every 5 frames. Each video is summarized by 20 users for TVSum, which is annotated by 1000 responses on Amazon Mechanical Turk (46). Users rate the videos' relevance on a scale of 1 to 5, with segments lasting two seconds each. The Yahoo! WebScope application provides access to the human-annotated TVSum significance scores.

7.3.2 Result, Evaluation and Analysis

Quantitative metrics are employed that align with the criteria utilized in previous studies to ensure a fair comparison. The F-score, precision, and recall assessment criteria are used for the same.

$$P = \frac{NK_{\text{matched}}}{NGs} \tag{7.1}$$

$$R = \frac{NK_{\text{matched}}}{NMs} \tag{7.2}$$

F-Score is

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \tag{7.3}$$

Where NKmatched represents the number of keyframes matched between the summary generated from the proposed algorithm and the ground truth summary in the dataset, and NGs and NM_s show the number of keyframes in the generated summary and manual summary. Precision represents the ratio of true positive frames to the total generated frames, while recall indicates the ratio of true positive frames to all ground truth frames. Consequently, the F-score is utilised to gauge the averages of recall and precision. The F-score serves as a balanced measure, assessing the overall performance of video summarization. 85.8% F-Score is achieved using the proposed approach. The proposed approach is evaluated against a number of representative techniques on TVSum dataset, as shown in Table 7.1

Approach	**TVSum (F_Score)**
AC-SUM-GAN (144)	60.6%
ADSum (145)	64.3 %
CA-SUM (146)	61.4%
MHSCNET (147)	69.3%
Bi-Convolutional-LSTM-GAN (148)	71.6 %
Proposed Approach	**85.8%**

Table 7.1: F_Score

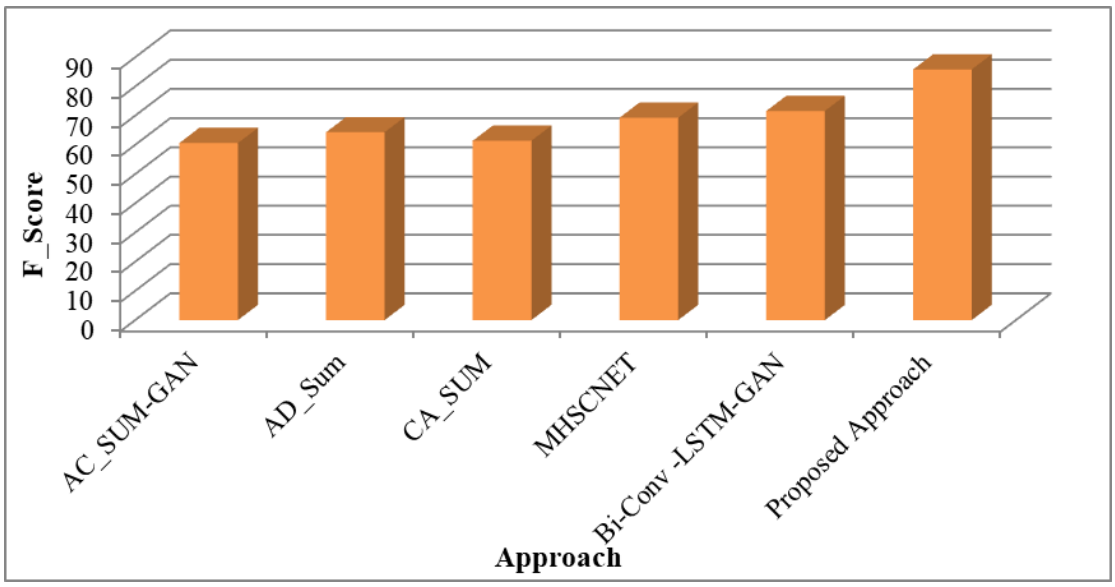


Figure 7.6: F_Score Analysis

The proposed approach has successfully condensed a 1-minute Full HD video into a 20-second summary, resulting in a storage reduction of approximately 60%. This reduction demonstrates the project's efficiency in optimizing storage space. Fig. 7.6 shows the F_ Score analysis of the proposed approach with other approaches. The chosen resolution, codec, and compression algorithm contribute to the storage reduction without compromising visual and audio quality. Despite the reduction in duration and frame rate, the final output video maintains high visual quality at 1920x1080 resolution. The advanced face recognition models contribute to the preservation of essential frames with significant facial presence, ensuring the coherence of the summary.

7.4 Conclusion

Anyone who wants to learn more in less time can benefit from watching video summaries. They facilitate the process of swiftly determining the value of a certain video for learning about a particular subject and assist in overcoming language obstacles. With the current proliferation of huge video data over the Internet, more and more automatic video summary extraction is needed to provide more effective and interesting viewing experiences. Video summarization simplifies this process, so it has garnered much interest in recent years. The AI-based video summarization approach presented in this research demonstrates a significant advancement in the field, effectively identifying keyframes, contributing to both accuracy and time efficiency. The multi-model strategy, incorporating human presence detection, face presence detection, and facial recognition, enhances the overall summarization process by providing a nuanced and contextually rich analysis of video content. Peak wave analysis is essential for locating keyframes, or frames that record significant moments in the video, When it comes to

a video summary. Through the process of peak analysis, the algorithm is able to identify specific events, such the presence of a person or a recognizable face, in the signal created from video frames. Future research directions may explore the incorporation of additional modalities, such as sentiment analysis of audio content or context-aware summarization. Future directions for this video summarization approach may involve the adaptation of models for real-time applications. Implementing real-time summarization capabilities could enhance the system's utility in scenarios where immediate insights or Summaries are required. Fine-tuning and updates to the models based on evolving datasets and technological advancements can contribute to further improvements in accuracy.

Publication

The work discussed in this chapter is communicated in:

Shambharkar, Prashant Giridhar, and Ruchi Goel, "Keyframe Extraction via Peak Wave Analysis with Integrated Human Presence and Face Recognition, Journal of Visual Communication and Image Representation, Elsevier (2024).

CHAPTER 8

CONCLUSION, FUTURE WORK and SOCIAL IMPACT

8.1 Summary of the Research Work

This research work focused to improve the video summaries in lieu of solving challenging video summarization problems. Proposed methods show a variety of approaches and strategies for video summarization, such as subtitles, autoencoders, hybrid models, multi-modal fusion, and integrated analytic methods. Each method has various benefits and contributes to the efficacy and efficiency of video summarization algorithms.

8.2 Contributions and Major Findings

A thorough assessment of the literature on video summarization identifies several important conclusions that influence the state of the field today and point to intriguing avenues for further research.

- Video summarization analyzes videos, dividing them into meaningful segments, highlighting significant components, and producing textual summaries. This is done using techniques such as shot segmentation, feature extraction, keyframe extraction, machine learning and AI, and textual analysis.

- Deep learning, semantic understanding, and user personalisation are innovations in video summarization that aim to enhance the user experience, efficiency, and accuracy of user.

- Beyond saving viewers time, there are other uses for a video summary. It is applicable to many fields as it improves content management, and accessibility and has application in many fields like surveillance, hospital surgery, Research labs and education By enhancing search and organization systems. Spotting possible issues and producing succinct summaries for easier comprehension and review

can be applied to many fields.

- To assess the efficacy of video summarization techniques, precise measurements, succinctness, and human assessment are necessary.
- While multi-modal video summarization algorithms are still in stages of development, they mostly concentrate on audio text and visual data. For efficient training and assessment, the systems now in use lack the ability to integrate various modalities and a multimodal video summarizing dataset.
- The user-interaction-focused strategy is inadequate. However, user involvement or preference-based techniques are still in development. The system designed to gather online user preferences is quite complicated and has very little functionality.
- Many existing methods struggle to adapt to videos from specific domains. For example, summarizing a sports game requires a different approach than summarizing a scientific lecture.
- Existing techniques for summarizing videos are not adaptable enough to accommodate different user requirements. Frameworks are required that are able to produce summaries of different lengths according to user preferences and that can efficiently adjust to the distinctive features of various nature videos.

8.3 Implications of the Research

Video summarizing research has the opportunity to revolutionize how we deal with data from videos. It can help users explore the large ocean of video content more effectively, resulting in a more informed and simplified video experience. However, addressing ethical concerns and potential biases is still critical for the responsible development and deployment of this powerful technology. Video summarization research has major implications for a variety of domains. This can improve efficiency in industries like education, journalism, and research, where users must quickly discover important content among enormous video archives. In education, video summaries are an efficient way for teachers to extract key concepts and highlights from instructional videos. This promotes active learning approaches by allowing students to quickly comprehend key information and focus on important topic portions. In Smart Cities and Surveillance, it can speed up the analysis of large amounts of video data, allowing authorities to concentrate on anomalies or crucial occurrences more effectively. Video summarizing approaches enable individuals to manage their personal collections by automatically

creating condensed versions of their recordings. This makes it easier to organize and review personal memories and experiences captured on video. Summarized videos also improve accessibility for people with disabilities by offering alternative forms for viewing video content, such as text-based summaries or audio descriptions. This encourages diversity and ensures equal access to information and entertainment resources.

8.4 Limitations and Challenges

While the video summary provides interesting prospects, it currently encounters several constraints and challenges. Summarizing videos as they are taken or streamed necessitates efficient algorithms. Complex approaches may not be appropriate for real-time applications due to significant processing expenses. A live video summary for security purposes must be rapid and efficient. Second, people's preferences for significant frames differ from one another. One family member may be essential to one person while being unimportant to another. This lack of interpretability makes it impossible to improve the summarization process or confirm that the model is capturing the correct data.

8.5 Future Directions for Research

Video summary is a rapidly expanding field with promising prospects. Here are some important areas where future study is anticipated to focus:

1. Techniques should capture the semantic meaning of videos rather than merely visually appealing frames. This involves a greater understanding of complex information and storytelling.

2. Personalized customizing summaries to meet user preferences and information demands is an ongoing problem. Summarization systems could employ user profiles, video genres, or even real-time attention to personalize content.

3. The utilization of sensor data, such as heart rate in exercise videos, could yield valuable insights for specific summarization tasks.

4. Video summarization, especially in surveillance applications, raises issues related to privacy. More research can be done to look at ways to balance the advantages of summarization with user privacy.

5. Users can improve and personalize the videos with interactive summaries and augmented reality overlays, allowing them to focus on certain elements or provide opinions.

6. Developing robust evaluation metrics or more nuanced ways to assess the quality of video summaries. This could involve measuring how well summaries capture semantic meaning, factual accuracy, user satisfaction, and suitability for different tasks (e.g., browsing vs. in-depth understanding).

7. Techniques for producing real-time video summaries are being investigated to assist applications such as live streaming, and event monitoring.

8.6 Conclusion

Techniques such as subtitle analysis provide a practical solution for videos with correct captions, allowing for real-time summary. Autoencoders with mode-based learning demonstrate promise in automatically extracting keyframes that represent the primary content. Hybrid models, such as VSEM and TAVM, illustrate the ability to combine textual information with audio or video elements to produce deeper summaries. TAVM promotes a multimodal approach that incorporates text, audio, and video frames. Research on keyframe extraction looks into combining additional elements like as human presence and face recognition, which could be useful for a specific video summarizing jobs. Overall, this research illustrates the continued development of various video summary approaches, pointing to increasingly effective and complete methods for summarizing video content. This study sets the path for further advancements in summarizing different video summarization techniques and adapting to different user needs.

REFERENCES

[1] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, pp. 137–144, Apr. 2015.

[2] P. G. Shambharkar and M. N. Doja, "Movie trailer classification using deer hunting optimization based deep convolutional neural network in video sequences," *Multimedia Tools and Applications*, vol. 79, pp. 21197–21222, May 2020.

[3] V. Tiwari and C. Bhatnagar, "A survey of recent work on video summarization: approaches and techniques," *Multimedia Tools and Applications*, vol. 80, p. 27187–27221, May 2021.

[4] A. Senthil Murugan, K. Suganya Devi, A. Sivaranjani, and P. Srinivasan, "A study on various methods used for video summarization and moving object detection for video surveillance applications," *Multimedia Tools and Applications*, vol. 77, p. 23273–23290, Jan. 2018.

[5] Y. Li, Y.-F. Ma, and H.-J. Zhang, "Salient region detection and tracking in video," in *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, IEEE, 2003.

[6] Z. Xiong, R. Radhakrishnan, Y. Rui, A. Divakaran, T. Chen, and T. S. Huang, *A Unified Framework for Video Indexing, Summarization, Browsing, and Retrieval*, p. 437–472. Elsevier, 2009.

[7] P. K. Lai, M. Decombas, K. Moutet, and R. Laganiere, "Video summarization of surveillance cameras," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, Aug. 2016.

[8] M. Kini and K. Pai, "A survey on video summarization techniques," in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, vol. 1, pp. 1–5, IEEE, Mar. 2019.

[9] S. E. F. de Avila, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araújo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, p. 56–68, Jan. 2011.

[10] E. J. C. Cahuina and G. C. Chavez, "A new method for static video summarization using local descriptors and video temporal segmentation," in *2013 XXVI Conference on Graphics, Patterns and Images*, IEEE, Aug. 2013.

[11] S.-h. Zhong, J. Wu, and J. Jiang, "Video summarization via spatio-temporal deep architecture," *Neurocomputing*, vol. 332, p. 224–235, Mar. 2019.

[12] J. Ren and J. Jiang, "Hierarchical modeling and adaptive clustering for real-time summarization of rush videos," *IEEE Transactions on Multimedia*, vol. 11, p. 906–917, Aug. 2009.

[13] T. Hussain, K. Muhammad, W. Ding, J. Lloret, S. W. Baik, and V. H. C. de Albuquerque, "A comprehensive survey of multi-view video summarization," *Pattern Recognition*, vol. 109, p. 107567, Jan. 2021.

[14] N. Trieu, S. Goodman, P. Narayana, K. Sone, and R. Soricut, "Multi-image summarization: Textual summary from a set of cohesive images," *arXiv preprint arXiv:2006.08686*, 2020.

[15] V. Kaushal, S. Subramanian, S. Kothawade, R. Iyer, and G. Ramakrishnan, "A framework towards domain specific video summarization," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Jan. 2019.

[16] A. Sharghi, J. S. Laurel, and B. Gong, "Query-focused video summarization: Dataset, evaluation, and a memory network based approach," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, July 2017.

[17] S. Xiao, Z. Zhao, Z. Zhang, X. Yan, and M. Yang, "Convolutional hierarchical attention network for query-focused video summarization," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, p. 12426–12433, Apr. 2020.

[18] B. S. Phadikar, A. Phadikar, and G. K. Maity, "Content-based image retrieval in dct compressed domain with mpeg-7 edge descriptor and genetic algorithm," *Pattern Analysis and Applications*, vol. 21, p. 469–489, Nov. 2016.

[19] Y. Jiang, K. Cui, B. Peng, and C. Xu, "Comprehensive video understanding: Video summarization with content-based video recommender design," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, Oct. 2019.

[20] A. Chauhan and S. Vegad, "Smart surveillance based on video summarization: a comprehensive review, issues, and challenges," *Evolutionary Computing and Mobile Sustainable Networks: Proceedings of ICECMSN 2021*, pp. 433–449, 2022.

[21] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 3278–3292, Aug. 2021.

[22] M. Fei, W. Jiang, and W. Mao, "A novel compact yet rich key frame creation method for compressed video summarization," *Multimedia Tools and Applications*, vol. 77, p. 11957–11977, June 2017.

[23] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, p. 121–143, Feb. 2008.

[24] N. Hussein, E. Gavves, and A. W. Smeulders, "Videograph: Recognizing minutes-long human activities in videos," *arXiv preprint arXiv:1905.05143*, 2019.

[25] A. Kanehira, L. Van Gool, Y. Ushiku, and T. Harada, "Aware video summarization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7435–7444, 2018.

[26] M. S. Nair and J. Mohan, "VSMCNN-dynamic summarization of videos using salient features from multi-CNN model," *Journal of Ambient Intelligence and Humanized Computing*, June 2022.

[27] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, *Creating Summaries from User Videos*, p. 505–520. Springer International Publishing, 2014.

[28] O. Gorokhovatskyi, O. Teslenko, and V. Zatkhei, "Online video summarization with the kohonen som in real time," in *CEUR Workshop Proceedings*, pp. 1067–1078, 2020.

[29] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, June 2014.

[30] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Computer Vision – ACCV 2018 Workshops*, pp. 39–54, Springer International Publishing, 2019.

[31] N. Cooharojananone, S. Kasamwattanarote, S. Satoh, and R. Lipikorn, "Real time trajectory search in video summarization using direct distance transform," in *IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS*, IEEE, Oct. 2010.

[32] P. Narwal, N. Duhan, and K. K. Bhatia, "A comprehensive survey and mathematical insights towards video summarization," *Journal of Visual Communication and Image Representation*, vol. 89, p. 103670, 2022.

[33] H. Binol, M. K. K. Niazi, C. Elmaraghy, A. C. Moberly, and M. N. Gurcan, "Automated video summarization and label assignment for otoscopy videos using deep learning and natural language processing," in *Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications*, vol. 11601, pp. 153–158, SPIE, 2021.

[34] P. Saini, K. Kumar, S. Kashid, A. Saini, and A. Negi, "Video summarization using deep learning techniques: a detailed analysis and investigation," *Artificial Intelligence Review*, Mar. 2023.

[35] R. Zhong, R. Wang, Y. Zou, Z. Hong, and M. Hu, "Graph attention networks adjusted bi-lstm for video summarization," *IEEE Signal Processing Letters*, vol. 28, pp. 663–667, 2021.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[38] S. Rani and M. Kumar, "Social media video summarization using multi-visual features and kohnen's self organizing map," *Information Processing andamp Management*, vol. 57, p. 102190, May 2020.

[39] Y. Zhang, M. Kampffmeyer, X. Liang, M. Tan, and E. P. Xing, "Query-conditioned three-player adversarial network for video summarization," *arXiv preprint arXiv:1807.06677*, 2018.

[40] S. Xiao, Z. Zhao, Z. Zhang, X. Yan, and M. Yang, "Convolutional hierarchical attention network for query-focused video summarization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 12426–12433, 2020.

[41] S. Xiao, Z. Zhao, Z. Zhang, Z. Guan, and D. Cai, "Query-biased self-attentive network for query-focused video summarization," *IEEE Transactions on Image Processing*, vol. 29, pp. 5889–5899, 2020.

[42] S. A. Ansari and A. Zafar, "Multi video summarization using query based deep optimization algorithm," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 10, pp. 3591–3606, 2023.

[43] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proceedings 1998 international conference on image processing. icip98 (cat. no. 98cb36269)*, vol. 1, pp. 866–870, IEEE, 1998.

[44] J. Wu, S.-h. Zhong, J. Jiang, and Y. Yang, "A novel clustering method for static video summarization," *Multimedia Tools and Applications*, vol. 76, pp. 9625–9641, 2017.

[45] C. Qi, Z. Feng, M. Xing, Y. Su, J. Zheng, and Y. Zhang, "Energy-based temporal summarized attentive network for zero-shot action recognition," *IEEE Transactions on Multimedia*, 2023.

[46] M. Ma, S. Mei, S. Wan, J. Hou, Z. Wang, and D. D. Feng, "Video summarization via block sparse dictionary selection," *Neurocomputing*, vol. 378, pp. 197–209, 2020.

[47] C. Chai, G. Lu, R. Wang, C. Lyu, L. Lyu, P. Zhang, and H. Liu, "Graph-based structural difference analysis for video summarization," *Information Sciences*, vol. 577, pp. 483–509, 2021.

[48] J. Park, J. Lee, I.-J. Kim, and K. Sohn, "Sumgraph: Video summarization via recursive graph modeling," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 647–663, Springer, 2020.

[49] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," pp. 202–211, 2017.

[50] E. Apostolidis, A. I. Metsai, E. Adamantidou, V. Mezaris, and I. Patras, "A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization," in *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, pp. 17–25, 2019.

[51] Y. Yuan, H. Li, and Q. Wang, "Spatiotemporal modeling for video summarization using convolutional recurrent neural network," *IEEE Access*, vol. 7, pp. 64676–64685, 2019.

[52] A. Jangra, S. Mukherjee, A. Jatowt, S. Saha, and M. Hasanuzzaman, "A survey on multi-modal summarization," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–36, 2023.

[53] A. Verma, A. Soni, and A. Prajapati, "Video summarization using subtitles," *new arch-international journal of contemporary architecture*, vol. 8, no. 2, pp. 1077–1082, 2021.

[54] S. S. Alrumiah and A. A. Al-Shargabi, "Educational videos subtitles' summarization using latent dirichlet allocation and length enhancement.," *Computers, Materials and Continua*, vol. 70, no. 3, 2022.

[55] S. Garg, "Automatic text summarization of video lectures using subtitles," in *Recent Developments in Intelligent Computing, Communication and Devices: Proceedings of ICCD 2016*, pp. 45–52, Springer, 2017.

[56] C. Liu, M. Last, and A. Shmilovici, "Towards automatic textual summarization of movies," in *Recent Developments and the New Direction in Soft-Computing Foundations and Applications: Selected Papers from the 7th World Conference on Soft Computing, May 29–31, 2018, Baku, Azerbaijan*, pp. 481–491, Springer, 2021.

[57] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5179–5187, 2015.

[58] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[59] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR 2011*, pp. 3153–3160, IEEE, 2011.

[60] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3537–3545, 2019.

[61] S. M. Safdarnejad, X. Liu, L. Udpa, B. Andrus, J. Wood, and D. Craven, "Sports videos in the wild (svw): A video dataset for sports analysis," in *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, vol. 1, pp. 1–7, IEEE, 2015.

[62] N. Modani, P. Maneriker, G. Hiranandani, A. R. Sinha, Utpal, V. Subramanian, and S. Gupta, "Summarizing multimedia content," in *Web Information Systems Engineering – WISE 2016*, pp. 340–348, Springer International Publishing, 2016.

[63] P. G. Shambharkar and R. Goel, "Analysis of real time video summarization using subtitles," in *2021 International Conference on Industrial Electronics Research and Applications (ICIERA)*, IEEE, Dec. 2021.

[64] S. Yeung, A. Fathi, and L. Fei-Fei, "Videoset: Video summary evaluation through text," *arXiv preprint arXiv:1406.5824*, 2014.

[65] Y. Li and B. Merialdo, "VERT," in *Proceedings of the international conference on Multimedia - MM '10*, ACM Press, 2010.

[66] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, pp. 996–1009, May 2019.

[67] G. B. Martins, J. P. Papa, and J. Almeida, "Temporal-and spatial-driven video summarization using optimum-path forest," in *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, IEEE, Oct. 2016.

[68] M. Merler, K.-n. C. Mac, D. Joshi, Q.-b. Nguyen, S. Hammer, J. Kent, J. Xiong, M. N. Do, J. R. Smith, and R. S. Feris, "Automatic Curation of Sports Highlights using Multimodal Excitement Features," *IEEE Transactions on Multimedia*, vol. PP, no. c, p. 1, 2018.

[69] S. Huang, X. Li, Z. Zhang, F. Wu, and J. Han, "User-Ranking Video Summarization with Multi-Stage Spatio-Temporal Representation," vol. XX, no. X, pp. 1–11, 2018.

[70] A. Javed, A. Irtaza, H. Malik, M. T. Mahmood, and S. Adnan, "Multimodal framework based on audio-visual features for summarisation of cricket videos," vol. 13, pp. 615–622, 2019.

[71] B. Zhao, M. Gong, and X. Li, "AudioVisual video summarization," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–8, 2021.

[72] A. Savelieva, B. Au-Yeung, and V. Ramani, "Abstractive summarization of spoken and written instructions with bert," *arXiv preprint arXiv:2008.09676*, 2020.

[73] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Computer Vision – ECCV 2016*, pp. 766–782, Springer International Publishing, 2016.

[74] B. Zhao, X. Li, and X. Lu, "Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7405–7414, 2018.

[75] B. Zhao, M. Gong, and X. Li, "Audiovisual video summarization," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[76] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis, "Video event detection and summarization using audio, visual and text saliency," in *2009 IEEE international conference on acoustics, speech and signal processing*, pp. 3553–3556, IEEE, 2009.

[77] A. Raventos, R. Quijada, L. Torres, and F. Tarrés, "Automatic summarization of soccer highlights using audio-visual descriptors," *SpringerPlus*, vol. 4, no. 1, p. 301, 2015.

[78] M. Hesham, B. Hani, N. Fouad, and E. Amer, "Smart trailer: Automatic generation of movie trailer using only subtitles," in *2018 First International Workshop on Deep and Representation Learning (IWDRL)*, pp. 26–30, IEEE, 2018.

[79] A. Jangra, A. Jatowt, M. Hasanuzzaman, and S. Saha, "Text-image-video summary generation using joint integer linear programming," in *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pp. 190–198, Springer, 2020.

[80] M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas, and F. A. Shah, "Video summarization: Techniques and classification," in *Computer Vision and Graphics*, Lecture notes in computer science, pp. 1–13, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[81] N. Dimitrova, "Context and memory in multimedia content analysis," *IEEE MultiMedia*, vol. 11, p. 7–11, July 2004.

[82] G. Yasmin, S. Chowdhury, J. Nayak, P. Das, and A. K. Das, "Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework," *Neural Computing and Applications*, vol. 35, p. 4881–4902, June 2021.

[83] H. Bhaumik, S. Bhattacharyya, and S. Chakraborty, *Content Coverage and Redundancy Removal in Video Summarization*, p. 352–374. IGI Global, 2017.

[84] J. Sellers, *Evaluation of LSA and TextRank methods for automatic text summarization*. PhD thesis, Dissertation, University of Washington, Seattle, WA, 2019.

[85] B. Mirzasoleiman, S. Jegelka, and A. Krause, "Streaming non-monotone submodular maximization: Personalized video summarization on the fly," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[86] A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, and D. R. I. M. Setiadi, "Review of automatic text summarization techniques amp; methods," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, p. 1029–1046, Apr. 2022.

[87] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Video summarization using deep semantic features," in *Computer Vision – ACCV 2016*, pp. 361–377, Springer International Publishing, 2017.

[88] V. Rajpoot and S. Girase, "A study on application scenario of video summarization," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 936–943, IEEE, 2018.

[89] L. Moraes, R. M. Marcacini, and R. Goularte, "Video summarization using text subjectivity classification," in *Brazilian Symposium on Multimedia and Web*, ACM, Nov. 2022.

[90] D. Gupta and A. Sharma, "A comprehensive study of automatic video summarization techniques," *Artificial Intelligence Review*, Mar. 2023.

[91] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1113–1126, 2020.

[92] J. Lin, S. hua Zhong, and A. Fares, "Deep hierarchical LSTM networks with attention for video summarization," *Computers andamp Electrical Engineering*, vol. 97, p. 107618, Jan. 2022.

[93] S. Parri, V. Kosana, and K. Teeparthi, "A hybrid GAN based autoencoder approach with attention mechanism for wind speed prediction," in *2022 22nd National Power Systems Conference (NPSC)*, IEEE, Dec. 2022.

[94] S. S. Harakannanavar, S. R. Sameer, V. Kumar, S. K. Behera, A. V. Amberkar, and V. I. Puranikmath, "Robust video summarization algorithm using supervised machine learning," *Global Transitions Proceedings*, vol. 3, pp. 131–135, jun 2022.

[95] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2015.

[96] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila, "Rethinking the evaluation of video summaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7596–7604, 2019.

[97] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, and B. Kainz, "Video summarization through reinforcement learning with a 3d spatio-temporal u-net," *IEEE Transactions on Image Processing*, vol. 31, pp. 1573–1586, 2022.

[98] L. Lan and C. Ye, "Recurrent generative adversarial networks for unsupervised WCE video summarization," *Knowledge-Based Systems*, vol. 222, p. 106971, June 2021.

[99] M. U. Sreeja and B. C. Kovoor, "A multi-stage deep adversarial network for video summarization with knowledge distillation," *Journal of Ambient Intelligence and Humanized Computing*, Jan. 2022.

[100] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, July 2019.

[101] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *The Visual Computer*, vol. 38, pp. 2939–2970, June 2021.

[102] F. Amato, A. Castiglione, V. Moscato, A. Picariello, and G. Sperlì, "Multimedia summarization using social media content," *Multimedia Tools and Applications*, vol. 77, pp. 17803–17827, Jan. 2018.

[103] C. Panagiotakis, H. Papadakis, and P. Fragopoulou, "Personalized video summarization based exclusively on user preferences," in *Lecture Notes in Computer Science*, pp. 305–311, Springer International Publishing, 2020.

[104] A. Bhowmik, S. Kumar, and N. Bhat, "Evolution of automatic visual description techniques-a methodological survey," *Multimedia Tools and Applications*, vol. 80, pp. 28015–28059, May 2021.

[105] L. Jin, Z. Li, and J. Tang, "Deep semantic multimodal hashing network for scalable image-text and video-text retrievals," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2020.

[106] J. Dong, X. Li, W. Lan, Y. Huo, and C. G. Snoek, "Early embedding and late reranking for video captioning," in *Proceedings of the 24th ACM international conference on Multimedia*, Oct. 2016.

[107] M. V. M. Cirne and H. Pedrini, "VISCOM: A robust video summarization approach using color co-occurrence matrices," *Multimedia Tools and Applications*, vol. 77, pp. 857–875, Jan. 2017.

[108] X. Lin, G. Bertasius, J. Wang, S.-F. Chang, D. Parikh, and L. Torresani, "Vx2text: End-to-end learning of video-based text generation from multimodal inputs," pp. 7005–7015, 2021.

[109] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, vol. 9, pp. 108069–108082, 2021.

[110] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *Journal of Big Data*, vol. 6, Oct. 2019.

[111] S. Sumana, "Towards automatically generating release notes using extractive summarization technique," in *Proceedings of the 33rd International Conference on Software Engineering and Knowledge Engineering*, KSI Research Inc., July 2021.

[112] G. Wu, S. Wang, and L. Liu, "Fast video summary generation based on low rank tensor decomposition," *IEEE Access*, vol. 9, pp. 127917–127926, 2021.

[113] A. Coutrot and N. Guyader, "Toward the introduction of auditory information in dynamic visual attention models," in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, IEEE, July 2013.

[114] N. Khamis, T. C. Sin, and G. C. Hock, "Segmentation of residential customer load profile in peninsular malaysia using jenks natural breaks," in *2018 IEEE 7th International Conference on Power and Energy (PECon)*, IEEE, Dec. 2018.

[115] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder–decoder networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, pp. 1709–1717, June 2020.

[116] W. Zhu, J. Lu, J. Li, and J. Zhou, "DSNet: A flexible detect-to-summarize network for video summarization," *IEEE Transactions on Image Processing*, vol. 30, pp. 948–962, 2021.

[117] W. Zhu, Y. Han, J. Lu, and J. Zhou, "Relational reasoning over spatial-temporal graphs for video summarization," *IEEE Transactions on Image Processing*, vol. 31, pp. 3017–3031, 2022.

[118] R. Jain, P. Jain, T. Kumar, and G. Dhiman, "Real time video summarizing using image semantic segmentation for cbvr," *Journal of Real-Time Image Processing*, vol. 18, pp. 1827–1836, 2021.

[119] Y. Zhu, W. Zhao, R. Hua, and X. Wu, "Topic-aware video summarization using multimodal transformer," *Pattern Recognition*, vol. 140, p. 109578, 2023.

[120] S. H. Emon, A. Annur, A. H. Xian, K. M. Sultana, and S. M. Shahriar, "Automatic video summarization from cricket videos using deep learning," in *2020 23rd International conference on computer and information technology (ICCIT)*, pp. 1–6, IEEE, 2020.

[121] J. Mohan and M. S. Nair, "Domain independent static video summarization using sparse autoencoders and k-means clustering," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 3, pp. 1945–1955, 2019.

[122] S. S. Harakannanavar, S. R. Sameer, V. Kumar, S. K. Behera, A. V. Amberkar, and V. I. Puranikmath, "Robust video summarization algorithm using supervised machine learning," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 131–135, 2022.

[123] M. I. H. Shihab, N. Tasnim, H. Zunair, L. K. Rupty, and N. Mohammed, "Vista: Vision transformer enhanced by u-net and image colorfulness frame filtration for automatic retail checkout," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3183–3191, 2022.

[124] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[125] M. S. Nair and J. Mohan, "Static video summarization using multi-cnn with sparse autoencoder and random forest classifier," *Signal, Image and Video Processing*, vol. 15, pp. 735–742, 2021.

[126] K. Deeparani and P. Sudhakar, "Efficient image segmentation and implementation of k-means clustering," *Materials Today: Proceedings*, vol. 45, pp. 8076–8079, 2021.

[127] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Automatic video summarization by graph modeling," in *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 104–109, IEEE, 2003.

[128] S. Kannappan, Y. Liu, and B. P. Tiddeman, "Performance evaluation of video summaries using efficient image euclidean distance," in *Advances in Visual Computing: 12th International Symposium, ISVC 2016, Las Vegas, NV, USA, December 12-14, 2016, Proceedings, Part II 12*, pp. 33–42, Springer, 2016.

[129] H. Khan, T. Hussain, S. U. Khan, Z. A. Khan, and S. W. Baik, "Deep multi-scale pyramidal features network for supervised video summarization," *Expert Systems with Applications*, vol. 237, p. 121288, 2024.

[130] G. Pan, Y. Zheng, R. Zhang, Z. Han, D. Sun, and X. Qu, "A bottom-up summarization algorithm for videos in the wild," *EURASIP Journal on Advances in Signal Processing*, vol. 2019, pp. 1–11, 2019.

[131] G. Yasmin, S. Chowdhury, J. Nayak, P. Das, and A. K. Das, "Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework," *Neural Computing and Applications*, vol. 35, no. 7, pp. 4881–4902, 2023.

[132] S. Rani and M. Kumar, "Social media video summarization using multi-visual features and kohnen's self organizing map," *Information Processing and Management*, vol. 57, no. 3, p. 102190, 2020.

[133] P. Saini, K. Berwal, S. Kashid, and A. Negi, "Stkvs: secure technique for keyframes-based video summarization model," *Multimedia Tools and Applications*, pp. 1–34, 2024.

[134] P. G. Shambharkar and R. Goel, "From video summarization to real time video summarization in smart cities and beyond: A survey," *Frontiers in big Data*, vol. 5, p. 1106776, 2023.

[135] N. Baghel, S. C. Raikwar, and C. Bhatnagar, "Image conditioned keyframe-based video summarization using object detection," *arXiv preprint arXiv:2009.05269*, 2020.

[136] H. Shakil, A. Farooq, and J. Kalita, "Abstractive text summarization: State of the art, challenges, and improvements," *Neurocomputing*, p. 128255, 2024.

[137] J. Sunuwar and S. Borah, "A comparative analysis on major key-frame extraction techniques," *Multimedia Tools and Applications*, pp. 1–46, 2024.

[138] T. H. Sardar, R. A. Hazarika, B. Pandey, G. Prasad, S. M. Hassan, R. Dodmane, and H. Gohel, "Video key concept extraction using convolution neural network," in *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, pp. 1–6, IEEE, 2024.

[139] H. Gharbi, S. Bahroun, and E. Zagrouba, "Key frame extraction for video summarization using local description and repeatability graph clustering," *Signal, Image and Video Processing*, vol. 13, pp. 507–515, Nov. 2018.

[140] D. Qi, W. Tan, Q. Yao, and J. Liu, "Yolo5face: why reinventing a face detector," in *European Conference on Computer Vision*, pp. 228–244, Springer, 2022.

[141] A. N. Indrawati, N. Nuryani, A. S. Nugroho, and T. P. Utomo, "Obstructive sleep apnea detection using frequency analysis of electrocardiographic rr interval and machine learning algorithms," *Journal of Biomedical Physics and Engineering*, vol. 12, no. 6, p. 627, 2022.

[142] F. Benites, A. Delorme Benites, and C. M. Anson, *Automated Text Generation and Summarization for Academic Writing*, p. 279–301. Springer International Publishing, 2023.

[143] C. Wall, P. McMeekin, R. Walker, V. Hetherington, L. Graham, and A. Godfrey, "Sonification for personalised gait intervention," *Sensors*, vol. 24, p. 65, Dec. 2023.

[144] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Ac-sum-gan: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3278–3292, 2020.

[145] Z. Ji, Y. Zhao, Y. Pang, X. Li, and J. Han, "Deep attentive video summarization with distribution consistency learning," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 4, pp. 1765–1775, 2020.

[146] E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras, "Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pp. 407–415, 2022.

[147] W. Xu, R. Wang, X. Guo, S. Li, Q. Ma, Y. Zhao, S. Guo, Z. Zhu, and J. Yan, "Mhscnet: A multimodal hierarchical shot-aware convolutional network for video summarization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[148] M. Sreeja and B. C. Kovoor, "A multi-stage deep adversarial network for video summarization with knowledge distillation," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 8, pp. 9823–9838, 2023.

Received: 16 May 2023 | Revised: 29 June 2023 | Accepted: 11 August 2023

DOI: 10.1111/exsy.13437

ORIGINAL ARTICLE

Expert Systems WILEY

Auto encoder with mode-based learning for keyframe extraction in video summarization

Prashant Giridhar Shambharkar | Ruchi Goel

Department of Computer Science and Engineering, Delhi Technological University, Delhi, India

Correspondence
Ruchi Goel, Department of Computer Science and Engineering, Delhi Technological University, Delhi, 110042, India.
Email: ruchigoel_phdco2k18@dtu.ac.in

Abstract

The exponential increase in video consumption has created new difficulties for browsing and navigating through video more effectively and efficiently. Researchers are interested in video summarization because it offers a brief but instructive video version that helps users and systems save time and effort when looking for and comprehending relevant content. Key frame extraction is a method of video summarization that only chooses the most important frames from a given video. In this article, a novel supervised learning method 'TC-CLSTM Auto Encoder with Mode-based Learning' using temporal and spatial features is proposed for automatically choosing keyframes or important sub-shots from videos. The method was able to achieve an average F-score of 84.35 on TVSum dataset. Extensive tests on benchmark data sets show that the suggested methodology outperforms state-of-the-art methods.

KEYWORDS
computer vision, keyframe extraction, NLP, spatiotemporal features, video summarization

1 | INTRODUCTION

The proliferation of advanced photographic devices and the improvement of internet connectivity have brought exponential growth in technology. We rely on social media websites for everything because they have become so pervasive in our daily lives. This includes daily news and updates on significant events, entertainment, attaching with dear ones, ratings and recommendations of products and services, discovering new places to travel, finding jobs, etc. Multimedia content is being produced and consumed at an increasing rate. Huge volumes of data have been generated as a result of this progression and all real-time practical applications require such vast data to be processed effectively. A typical computer vision task created for video analysis is the video because visuals convey new information regarding action, video enables a more in-depth analysis of the instance. Video is the most widely used type of visual information, which has grown in popularity swiftly. The task is to analyse the video's multimodal data in search of relevant cues (such redundant information) that can guide a conclusion (Moraes et al., 2022). High computational power and advanced vision techniques have increased the scope of video summarization techniques. Video summarization helps us to quickly review lengthy videos by the removal of pointless and unnecessary frames. The goal of video summarization is to extract the most significant and instructive segments from the full-length video in order to create a comprehensive and succinct description (Haq et al., 2022). A good video summary would condense the main points of the original video into a concise, viewable overview.

Video summarization is broadly divided into static and dynamic summary (Saini et al., 2023). Dynamic summaries also referred to as video skim, are produced by video segments that analyse the audio and visual content of the video. A static summary or keyframe-based summary is a grouping of the pertinent keyframes that are needed to produce the desired summary and are chosen in a sequential sequence (Tiwari & Bhatnagar, 2021). These can also be referred to as representative frames (R-frames), or a group of standout images gleaned from the video data. This kind of summary is static because the key-frames, which are spatially separated and distributed unevenly, prevent an adequate

Abbreviation: CLSTM, convolution Long short-term memory.

JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 40, 1253-1271 (2024)
DOI: 10.6688/JISE.202411_40(6).0007

VSEM: A Hybrid Model for Video Summarization

PRASHANT GIRIDHAR SHAMBHARKAR AND RUCHI GOEL+
Department of Computer Science and Engineering
Delhi Technological University
Delhi, 110042 India
E-mail: prashant.shambharkar@dtu.ac.in; ruchigoel06@gmail.com

In today's fast-moving digital era, video technology plays an important role. An effective video summarising approach is urgently needed to handle a lot of video data due to the ever-growing number of video content. In this paper, the authors have proposed, A hybrid summarization methodology for video summary evaluation using multimedia features (Text, Images, and Audio) that assess how well a video summary can keep the ranking of vital video frames, semantic data, and audio present in the original video. Video summary can be evaluated by ranking text, audio, and semantics of video frames, giving more accurate summarisation results. The proposed methodology works in three phases: The first part takes the text in the video, the second phase takes the audio to the file, and the last phase focuses on the video frames rather than images in the video. TVSum dataset has been used for the experimentation. F1 has been used as the evaluation metric for checking the efficacy and efficiency of the proposed methodology. The result shows that the proposed hybrid model achieves the highest F1 score of 69.9% and saves 75-80% of user time in watching video summaries instead of the whole video.

Keywords: video summarization, NLP, multimedia features, computer vision, multimodal representation

1. INTRODUCTION

Technology development has caused a quick increase in multimedia data on the Internet, making it difficult for consumers to access crucial information quickly [1]. Video is the most challenging multimedia (including text, pictures, graphics, and audio), as it incorporates all other media data into a single data stream and is difficult to access effectively due to its unstructured format and changing format length [2]. Video information is a sequential data type that gives unlimited data through its moving content [3]. Think about using YouTube to search for educational or tourist-related content, many individuals prefer not to invest their time in watching or listening to lengthy recordings. Instead, they often seek out concise video clips that provide a condensed and more digestible summary. It is inefficient to browse through the millions of returned results. It would be much simpler to view a brief description of each result. Secondly, because of the limited storage space, it is also necessary to summarise videos without losing much information. These issues can be solved by summarizing the essential information from the vast amount of

Received February 9, 2023; revised August 15 & October 1, 2023; accepted November 22, 2023.
Communicated by Jing-Ming Guo.
+ Corresponding author.

frontiers | Frontiers in Big Data

TYPE Systematic Review
PUBLISHED 09 January 2023
DOI 10.3389/fdata.2022.1106776

OPEN ACCESS

EDITED BY
Namita Gupta,
Maharaja Agrasen Institute of Technology, India

REVIEWED BY
Deepika Kumar,
Bharati Vidyapeeth's College of Engineering, India
Pooja Gupta,
Maharaja Agrasen Institute of Technology, India
Vidhi Khanduja,
University of Delhi, India

*CORRESPONDENCE
Ruchi Goel
✉ ruchigoel06@gmail.com

SPECIALTY SECTION
This article was submitted to Data Science, a section of the journal Frontiers in Big Data

RECEIVED 24 November 2022
ACCEPTED 14 December 2022
PUBLISHED 09 January 2023

CITATION
Shambharkar PG and Goel R (2023) From video summarization to real time video summarization in smart cities and beyond: A survey. *Front. Big Data* 5:1106776. doi: 10.3389/fdata.2022.1106776

From video summarization to real time video summarization in smart cities and beyond: A survey

Prashant Giridhar Shambharkar and Ruchi Goel*

Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

With the massive expansion of videos on the internet, searching through millions of them has become quite challenging. Smartphones, recording devices, and file sharing are all examples of ways to capture massive amounts of real time video. In smart cities, there are many surveillance cameras, which has created a massive volume of video data whose indexing, retrieval, and administration is a difficult problem. Exploring such results takes time and degrades the user experience. In this case, video summarization is extremely useful. Video summarization allows for the efficient storing, retrieval, and browsing of huge amounts of information from video without sacrificing key features. This article presents a classification and analysis of video summarization approaches, with a focus on real-time video summarization (RVS) domain techniques that can be used to summarize videos. The current study will be useful in integrating essential research findings and data for quick reference, laying the preliminaries, and investigating prospective research directions. A variety of practical uses, including aberrant detection in a video surveillance system, have made successful use of video summarization in smart cities.

KEYWORDS

computer vision, video summarization, real time video summarization (RVS), keyframes, summary

1. Introduction

Analyzing video content to extract valuable or intriguing information is difficult and time-consuming. Many videos are uploaded to YouTube, IMDB, tourism sites, Flickr, and other video-sharing sites every minute. Every minute, 300 h of video are posted to the YouTube channel and about one billion hours of video are watched every day (You Tube Stats, n.d). Millennial video cameras are installed in smart cities including public spaces, public transportation, banks, airfields, and other locations, resulting in a tremendous amount of data that is difficult to analyze in real time. There will be hundreds of suggestions for each search topic; navigating through these lengthy videos to find the essential video takes time, and also challenging to efficiently obtain this much data in a short amount of time.

Secondly, due to the abundance of videos, users must rely on metadata such as title, image, description, and remark to locate the video they want to see. This metadata,

Analysis of Real Time Video Summarization using Subtitles

Dr Prashant Giridhar Shambharkar
Department of CSE
Delhi Technological University
Delhi, India
prashant.shambharkar@dtu.ac.in

Ruchi Goel*
Department of CSE
Delhi Technological University
Delhi, India
ruchigoel06@gmail.com

2021 International Conference on Industrial Electronics Research and Applications (ICIERA) | 978-1-6654-3542-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIERA53202.2021.9726769

Abstract— **With the rapid growth of user-generated videos, being able to browse them effectively is becoming increasingly vital. Video summarization is seen to be a potential method for effectively realizing video content by identifying and selecting descriptive frames from the video. An automatic video summary would be advantageous for everyone who wants to save time and learn more in less time as video content continues to grow at a rapid rate. In this paper, essential components of a video can be summarized by the summation of subtitles in a video using text summarization and video mapping algorithms. The audio-generated version of the summary subtitle file will be played with the summarized video. The system's final output would be a summary video accompanied by a summarized audio-generated version.**

Index Term—Subtitle, Video summarization, Real Time Summarization

I. INTRODUCTION

With the advancement of the Internet and multimedia, an increasing number of videos are being produced. Video has a one-of-a-kind and significant effect on websites and social media platforms. Today, video is an extensively utilized multi-medium in a variety of applications (digital broadcasting, interactive television, video on demand, computer-based training, and multimedia devices) and video data is becoming increasingly significant used by a wide range of people. Managing, storing, and indexing massive amounts of video has become a pressing issue that must be addressed. By using keyframes or video skim to express the essential idea of videos, video summarization can solve this problem. As a result, in recent years, Proposing video summarization algorithms has sparked a great deal of research.

Video summarization is a technique for extracting important frames or sequences of frames from a video; it allows for quick browsing by condensing the original video into a synopsis and maintaining only the most important information [1]. In other words, the approach for constructing a summary of a video is video summarization, which might be in the form of a series of still images (keyframes) or moving images (video skims). Frames from a video are retrieved and processed into a static image in the keyframing technique [3]. Video skimming also means speeding up the frame rate or compressing a lengthier movie into a shorter one. The skims include audio and video parts., and dynamic summaries are formed by processing both the visual and auditory content of the video stream. [4] It is accomplished by deleting alternate frames or, in some cases, by employing an expression dependent on the trailer or summary's required length. Video summarization is classified into two types based on the amount of time it takes to summarise it: real-time and static, depending on whether it is performed on a live (real time videos) or recorded video [2]. Selecting critical frames when the video is being shot based on the video's context will be extremely useful in real-time circumstances.

This work focuses on selecting frames by summarising an automatically generated subtitle text file and mapping the summarised text to the video frames. The video's end output is a created audio track that plays in the background of the summary footage.

The next section discusses the literature survey. The proposed system is described in the III section, Part IV presents the experimental details and outcomes, while section V concludes the approach.

II. LITERATURE SURVEY

Even though the number of videos produced and indexed is rapidly increasing, the amount of time required to watch these video remains limited.

In [5] authors describes a video indexing and summarising method based on data taken from a DVD/DivX video's script file. The approach divides the script into segments, each of which is represented as a TF-IDF vector. Script matrix refers to the collection of these vectors. Two applications, video retrieval, and summarization are explained using machine learning techniques applied to the script matrix.

The necessity for semantic video indexing approaches has grown as the amount of available multimedia data repositories has grown. The authors of [6] presented a method for categorizing unsupervised video information using natural language processing techniques on the subtitles. It is also determined which WordNet domains correlate to the correct word senses.

Deep learning algorithms are being used to automate the creation of an internet trailer for any movie based just on its subtitle in [7]. The framework examines the movie subtitle file for significant textual elements that are utilized to categorize the film into its appropriate genre.

In [8] Authors has done text summarization of subtitle file using LSA then Video mapping is accomplished by selecting a video from the subtitle file that corresponds to the summary sentence.

In [9], the authors suggested a model for summarising subtitles based on LDA. LDA-generated keywords list was used to summarise the subtitles of instructive videos.

978-1-6654-3542-0/21/$31.00 ©2021 IEEE

2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS) | 979-8-3503-4233-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICTACS59847.2023.10389978

TAVM: A Novel Video Summarization Model Based on Text, Audio and Video Frames

Prashant Giridhar Shambharkar[1], Ruchi Goel[2]
[1,2]Department of Computer Science and Engineering, Delhi Technological University Delhi, India
E-mail: [1]prashant.shambharkar@dtu.ac.in, [2]ruchigoel06@gmail.com

***Abstract*—In today's digital world, the task of video summarization has gained immense importance within the realm of multimedia analysis. This relevance is largely driven by the exponential expansion in multimedia content consumption, encompassing audio, video, and images, which is readily available on-demand through various digital platforms. Automatic video summary is the process of creating a brief synopsis that summarizes the video by displaying its most useful and relevant elements, so consumers may rapidly comprehend the primary concept of a video without having to watch the entire material. Currently, the selection of the video segments to be included in the final summary is done in a variety of ways. The task is to analyze the video's multimedia data in search of relevant clues that will aid in decision-making. The proposed method TAVM (text, audio, and video mode) in this paper will provide the video summary using different multimedia elements of text, audio, and frames. The proposed TAVM method can be separated into three parts. The process begins with Video Processing, where the BEiT vision transformer is employed to recognize objects within the chosen frame. Following that, Audio Processing comes into play, which uses speech-to-text converters to transcribe the audio content. Finally, in the last step, the Summary Builder utilizes the GPT-3-based OpenAI API to generate a summary of the content. The experimental analysis on the benchmark dataset SumMe demonstrates the effectiveness of the proposed approach.**

Keywords: Video Summarization, Multimedia Analysis, Key-frames, BEiT(BERT Pre-Training of Image Transformers) vision transformer

I. Introduction

The fast expansion of multimedia data transmission over the internet needs the summarization of the whole data. Multiple information modalities are used in video streams to communicate information. It can be challenging for users to efficiently gather crucial information since, for instance, visual events can comprise objects, gestures, and scene changes, auditory events might be changes in audio sources, and textual events can contain conversations, subjects, and key phrases. It is difficult for humans to get important information from the whole data effectively. Manually extracting interesting parts from video footage and processing them are time-consuming operations, thus there is a need for automated approaches to reduce duplication and extract usable information [1]. Video summarization has emerged as a challenging challenge that aims to automatically analyse video footage. Video summarization gives the viewer a condensed version of the video that, incorporates all crucial details for comprehending the subject matter [2].

With the rise of video content as the primary source of data consumption for information, automation of the video summary process has taken centre stage. Video summary applications can be helpful in creating highlights for sporting events, movie trailers, medical diagnosis, and many real-life applications [3]. Different media organizations, such as sports or entertainment videos, develop teasers or previews for films and TV shows using highlights of events. Furthermore, video search engines can leverage video summarization for video indexing, browsing, retrieval, and recommendation [4]. Multimedia data are diverse and include more complicated information than plain text also it has a difficult time bridging the semantic gap between various modalities [5]. A video has a hierarchical structure that records spatial and temporal data as frames, shots, and scenes. It is not possible to watch a video all the way as it would take a lot of time and effort. The process of producing a video summary from one or more raw videos is automated by video summarization. Whether static or dynamic, a video summary represents standardized content or user preferences [6].

We contend in this work that the text and audio modality can help the visual modality comprehend the video's structure and content more thoroughly. To put it another way, the text, audio, and vision support the actions that are being done in the various modalities. For instance, the subtitle file as text helps to understand the video better, the wedding music conveys the festive mood of the setting, while the roars of support during cricket denote a successful run. But videos frequently experience audiovisual and text inconsistencies as well. An illustration would be that the sounding item is not visible. The fundamental difficulty in text audiovisual video summarization will be exacerbated by interference with the visual modality. As a result of the potential and difficulties mentioned above, we suggest using a text, Audio, and Visual model (TAVM) to combine text, audio, and visual data for video summarization. Figure 1 depicts the overall structure of our approach. The literature

Brief Profile

Ruchi Goel works as an Assistant Professor at Maharaja Agrasen Institute of Technology. She has submitted her thesis for a doctorate degree in the field of from Delhi Technological University. She completed her post-graduation in M.E (CSE) from DTU formerly Delhi College of Engineering in 2011, and her B.Tech (CSE) with from MDU University in 2003. She has published various research papers in peer-reviewed reputed International Journals, conferences. With 20 years of teaching experience and 1 year in the industry domain, her research interests include but are not limited to Software Testing, Image Processing, Artificial Intelligence, and Deep Learning. She has published 2 SCIE Papers and more than 15 International Conference and journals Papers.

Google Scholar -https://scholar.google.com/citations?user=VRGwMA4AAAAJ&hl=en

ORCID Profile - https://orcid.org/0000-0003-2961-8237