Visvesvaraya Technological University, Belagavi – 590018

MINI-PROJECT REPORT
ON

VIDEO DIGEST: A QUERY BASED DYNAMIC VIDEO SYNOPSIS SYSTEM

Submitted in partial fulfillment for the award of degree of

BACHELOR OF ENGINEERING
in
COMPUTER SCIENCE & ENGINEERING

Submitted by

Name 1	**4SO20CS075**
Student name 2	**4SO20CS080**
Third team member	**4SO20CS090**
Name 3	**4SO20CS095**

Under the Guidance of

Dr/Mr/Ms Mini-Project Coordinator Name
Associate/Assistant Professor, Department of CSE

DEPT. OF COMPUTER SCIENCE AND ENGINEERING
ST JOSEPH ENGINEERING COLLEGE
An Autonomous Institution
(Affiliated to VTU Belagavi, Recognized by AICTE, Accredited by NBA)
Vamanjoor, Mangaluru - 575028, Karnataka

2023-24

ST JOSEPH ENGINEERING COLLEGE

An Autonomous Institution

(Affiliated to VTU Belagavi, Recognized by AICTE, Accredited by NBA)

Vamanjoor, Mangaluru - 575028, Karnataka

DEPT. OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the Mini-project work entitled **"Video Digest: A Query Based Dynamic Video Synopsis System"** carried out by

Name 1	**4SO20CS075**
Student name 2	**4SO20CS080**
Third team member	**4SO20CS090**
Name 3	**4SO20CS095**

the bonafide students of VI semester Computer Science & Engineering in partial fulfillment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visvesvaraya Technological University, Belagavi during the year 2023-2024. It is certified that all suggestions indicated during internal assessment have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Mini-Project Coordinator Name	**Dr Sridevi Saralaya**
Mini-Project Coordinator	HOD-CSE

Abstract

Para 1 (Shall introduce the reader the subject matter of the project work. Shall discuss in brief the developments in the area of work/research so far based on the reference to the literature. Identification of the problem and defining exactly the purpose of the intended work or proposition of the novel solution. Research Methodology, brief discussion about the novel solution, experimental work, results and discussions. Conclusions drawn based on the work, advantages or benefits of using novel solution by way saving in cost, labor, space, energy or overall economy. Finally, write the proposition of the scope for future work.)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In article [5] Query based video summarisation is a crucial activity in many industries, including traffic management and surveillance. This has traditionally been accomplished through the laborious, error-prone, and time-consuming process of manual checking. Automatic object detection and summarisation is now possible thanks to computer vision and deep learning.

Creating a condensed and representative version of a longer video. It involves identifying and extracting key frames or scenes from the original video and arranging them in a way that conveys the essence of the video's content in a shorter amount of time. Video synopsis has become increasingly important due to the growing amount of video data being generated and the need to quickly analyze and understand this data. It has many practical applications in areas such as surveillance, law enforcement, and video search and retrieval.

Video synopsis has a wide range of potential users across different industries and applications. For law enforcement agencies, it can help to quickly analyze large amounts of surveillance footage, identify key events or suspects, and track their movements. Security companies can use it to monitor premises and identify potential threats or security breaches. Media companies can benefit from creating condensed versions of longer videos for news reports, documentaries, or other types of content. Marketing companies can use it to identify key moments in video content that are likely to be of interest to viewers and to create targeted advertisements or promotional videos. Overall, video synopsis offers a powerful tool for analyzing and summarizing video content for a variety of purposes.

1.2 Problem statement

Video digest revolves around developing algorithms and techniques to automatically create concise and coherent summaries of longer videos while retaining the essential content

and context of the original footage. The primary goal is to provide an efficient representation of the video, making it easier for users to understand the video's content quickly without having to watch the entire duration. In this project we develop a query based dynamic summarization tool using deept learning techniques.

1.3 Scope

The scope of query-based video synopsis is to create a condensed and representative version of a longer video that is specifically tailored to a user's query. This allows users to quickly and efficiently find relevant information within a large video dataset. Query-based video synopsis is important because it offers a solution to the problem of information overload in large video datasets. With the exponential growth of video data, it is becoming increasingly difficult for users to manually sift through large amounts of video content to find the information they need. Query-based video synopsis provides an automated and efficient way to extract and present the most relevant parts of the video data, making it easier for users to access and use the information they need. Another important aspect of query-based video synopsis is its potential applications in various industries and domains. For example, in law enforcement, query-based video synopsis can help investigators quickly identify relevant video footage related to a specific crime or suspect. In education, it can be used to create condensed versions of lecture videos for students who need to review specific topics or concepts. In marketing, it can be used to identify and extract key moments in promotional videos that are likely to be of interest to customers. Overall, query-based video synopsis has significant scope and importance in enabling users to efficiently and effectively access relevant information from large video datasets, and it is likely to continue to play an increasingly important role in a wide range of industries and applications.

Chapter 2

Software Requirements Specification

2.1 Introduction

paragraph contents...

2.2 Functional requirements

paragraph contents...

2.3 Non-Functional requirements

paragraph contents...

2.3.1 subsection heading

paragraph contents...

2.3.2 subsection heading

paragraph contents...

2.4 User Interface requirements

paragraph contents...

2.4.1 subsection heading

paragraph contents...

2.4.2 subsection heading

paragraph contents...

2.5 Software Requirements

paragraph contents...

2.5.1 subsection heading

paragraph contents...

2.5.2 subsection heading

paragraph contents...

2.6 Hardware Requirements

paragraph contents...

Chapter 3

System Design

paragraph contents...

3.1 Architecture Design

Figure 3.1: System Architecture Diagram

This Figure 5.1 illustrates a high-level overview of the audio visual speech separation system. It is important to note that the specific techniques, algorithms, and models used in each component can vary depending on the implementation approach and the requirements of the system.

3.2 Decomposition Description

Figure 3.2: Flow chart

Figure 5.2 represent the flow chart of the proposed system. In audio visual speech separation, the goal is to decompose an audio signal containing multiple overlapping speakers into individual speech signals corresponding to each speaker. The decomposition process involves separating the desired speech signals from the background noise and other interfering sounds.

3.3 Data Flow Design

The audio input undergoes pre-processing, while the visual input is processed to extract relevant cues. The pre-processed audio and processed visual data are then integrated. From the integrated representation, features are extracted. These features are utilized in the speech separation stage, where individual speech signals are separated from the mixture. Post-processing techniques are applied to enhance the quality of the separated speech signals. Finally, the individual speech signals are outputted as the result of the system. The data flow design ensures a sequential flow of operations, starting from capturing and processing the inputs, integrating the audio-visual information, extracting features, performing speech separation, applying post-processing, and generating the output. This design allows for effective processing and separation of audio visual data to obtain distinct speech signals from overlapping speakers.

Figure 3.3: Dataflow design

Chapter 4

Implementation

the chapter contains paragraph contents.(Pseudocode, Algorithm etc), for ex: Check the following data

4.1 Audio Extraction

Audio extraction is the process of isolating and extracting the audio content from a multimedia source, such as a video file. It involves separating the audio track from the accompanying video or other elements to obtain a standalone audio file representing the sound present in the source material.

Figure 4.1: code snippet for audio extraction

4.2 Speech Separation

SpeechBrain is an open-source framework

4.2.1 Sepformer

SepFormer is an algorithm for speech separation that utilizes self-attention mechanisms. It employs a transformer-based architecture to capture long-range dependencies and model the relationships between time-frequency points in the audio mixture, enabling the separation of multiple speech sources from the mixture.

Figure 4.2: code snippet for speech separation

4.3 Speech Enhancement

4.3.1 Lite Audio Visual Speech Enhancement

Lite AVSE algorithm is used for the separation and enhancement of the speech. The system includes two visual data compression techniques and removes the visual feature extraction network from the training model, yielding better online computation efficiency. As for the audio features, short-time Fourier transform (STFT) is calculated of 3-second audio segments. Each time-frequency (TF) bin contains the real and imaginary parts of a complex number, both of which used as input. Power-law compression used to prevent loud audio from overwhelming soft audio. The same processing is applied to both the noisy signal and the clean reference signal.

4.3.2 Spectral Subtraction

Spectral subtraction is a technique used in audio signal processing to reduce background noise from an audio signal. It involves estimating the noise spectrum from a noisy signal and subtracting it from the noisy spectrum to enhance the desired signal. The resulting spectrum is then transformed back into the time domain to obtain a cleaner audio signal.

Figure 4.3: code snippet for speech enhancement using LAVSE

Figure 4.4: code snippet for speech enhancement using spectral subtraction

4.4 Speaker Detection

The cv2 functions provide methods to load the pre-trained models, apply them to images or video frames, and draw bounding boxes around the detected faces. By leveraging cv2's face detection capabilities, you can automate tasks such as facial recognition, emotion analysis, or face tracking in various applications like surveillance, biometrics, or augmented reality.

Figure 4.5: code snippet for speaker detection

Chapter 5

Results and Discussion

this chapter contains the paragraphs as shown below

5.1 Face detection

Figure 5.1: Face detection

Above figure 8.1 shows initial face detection process using opencv and dlib. It convert the image to grayscale, apply the model using cv2.detectMultiScale(), and draw bounding boxes around the detected faces using cv2.rectangle(). Display or save the result using cv2.imshow() or cv2.imwrite().

5.2 Speaker recognition

Figure 5.2: Speaker recognition 1,person 1

Figure 5.3: Speaker recognition 1,person 2

Figure 5.4: Speaker recognition 2,person 1

Figure 5.5: Speaker recognition 2,person 2

Figure 5.6: Speaker recognition 3,person 1

Figure 5.7: Speaker recognition 3,person 2

Above figures from 8.2 to 8.7 shows speaker recognition process using opencv and dlib.Speaker detection using cv2 and dlib involves utilizing dlib's pre-trained models along with cv2 functions to detect and locate human faces. By combining face detection with additional techniques such as audio analysis or lip movement tracking, speaker detection can be achieved in various applications like video conferencing or surveillance.

Chapter 6

Conclusion and Future Work

[1] [6] The Project will help in narrowing the imprecise communication problem in real-time data using speech separation and speaker identification technique by Deep Learning and Image Processing algorithms. This will impact the communication and security sectors in a greater extent. Overall, this project aims to develop an application or method that can help to separate the audio-visual speech and enhance it based on speaker identification.
This project can be further developed as:

- By incorporating more real-world testing and gathering feedback from individual units.
- The system can be connected with communication devices or services to enable the users to communicate with others with ease.

[8] This project has a great potential to make a positive impact on communication and security situations. Its continuous improvement will be important to make this impact even greater

example for citing and bibtex for journal paper [2], conference paper [7], citing website [4], citing book [3]

References

[1] Muhammad Farrukh Bashir et al. “Subjective answers evaluation using machine learning and natural language processing”. In: *IEEE Access* 9 (2021), pp. 158972–158983.

[2] Florinel-Alin Croitoru, Vlad Hondru, and Radu Tudor Ionescu. “Diffusion models in vision: A survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[3] Paul Adrien Maurice Dirac. “The Principles of Quantum Mechanics”. In: International series of monographs on physics. Clarendon Press, 1981. ISBN: 9780198520115.

[4] Donald Knuth. *Knuth: Computers and Typesetting*. URL: `http://www-cs-faculty.stanford.edu/~uno/abcde.html`. accessed on 01.09.2016.

[5] Siyuan Li et al. “Matching Anything by Segmenting Anything”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 18963–18973.

[6] Himani Mittal and M. Syamala Devi. “Computerized Evaluation of Subjective Answers Using Hybrid Technique”. In: *Innovations in Computer Science and Engineering*. Ed. by H. S. Saini, Rishi Sayal, and Sandeep Singh Rawat. Singapore: Springer Singapore, 2016, pp. 295–303. ISBN: 978-981-10-0419-3.

[7] Omar Abdulwahabe Mohamad. “Smart Home Security based on optimal wireless sensor network routing protocols”. In: *2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. IEEE. 2015, SSS–17.

[8] Lee Smith. “City of Dreams”. In: *Tablet* 18 March (2008).