
RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

*Project Phase II Report On*

# Enhancing Accessibility for the Visually Impaired

*Submitted in partial fulfillment of the requirements for the award of the degree of*

## Bachelor of Technology

*in*

***Computer Science and Engineering***

**By**

**Elina Joy (U2003076)**

**Hemi Rose Sajeev (U2003093)**

**John Sherry (U2003107)**

**M. Aman Pradeep (U2003216)**

**Under the guidance of**

**Ms. Meenu Mathew**

**Department of Computer Science and Engineering**
**Rajagiri School of Engineering & Technology (Autonomous)**
**(Parent University: APJ Abdul Kalam Technological University)**
**Rajagiri Valley, Kakkanad, Kochi, 682039**
**April 2024**

# CERTIFICATE

*This is to certify that the project report entitled* ***"Enhancing Accessibility For The Visually Impaired"*** *is a bonafide record of the work done by* ***Elina Joy (U2003076)***, ***Hemi Rose Sajeev (U2003093)***, ***John Sherry (U2003107)*** *and* ***M Aman Pradeep (U2003216)*** *submitted to the Rajagiri School of Engineering & Technology (RSET) (Autonomous) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2023-2024.*

**Ms. Meenu Mathew**
Project Guide
Assistant Professor
Dept. of CSE
RSET

**Dr. Tripti C**
Project Coordinator
Assistant Professor
Dept. of CSE
RSET

**Dr. Preetha K. G.**
Head of the Department
Professor
Dept. of CSE
RSET

# ACKNOWLEDGEMENT

We wish to express our sincere gratitude towards **Prof. Dr. P. S. Sreejith**, Principal of RSET, and **Dr. Preetha K. G.**, Head of the Department of Computer Science and Engineering for providing us with the opportunity to undertake our project, "Enhancing Accessibility for the visually impaired-A human action recognition and speech conversion system".

We are highly indebted to our project coordinator, **Dr. Tripti C**, Assistant Professor, Dept of CSE, for her valuable support.

It is indeed our pleasure and a moment of satisfaction for us to express our sincere gratitude to our project guide **Ms. Meenu Mathew**, Assistant Professor, for her patience and all the priceless advice and wisdom she has shared with us.

Last but not the least, we would like to express our sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

**Elina Joy**

**Hemi Rose Sajeev**

**John Sherry**

**M. Aman Pradeep**

# Abstract

The project addresses a critical need for visually impaired individuals by developing an innovative system that seamlessly converts real-time human motion into descriptive text and, subsequently, into natural speech. Leveraging cutting-edge transformer model, computer vision techniques, and text-to-speech synthesis using Tacotron, this technology not only interprets intricate body movements but also generates human-like speech with exceptional clarity.

The system's core framework employs deep learning to analyze live human motion, extracting meaningful information from gestures, postures, and actions. This data is then translated into concise yet vivid textual descriptions, providing visually challenged users with a comprehensive understanding of their surroundings.

The second crucial component of this project involves the utilization of Tacotron-based text-to-speech synthesis. This technology ensures that the text descriptions are transformed into spoken words with remarkable naturalness and expressiveness, making the experience more immersive and intuitive for the end-users.

This holistic approach not only empowers visually impaired individuals with real-time, context-aware information but also enhances their overall quality of life by fostering greater independence and understanding of their environment

# Contents

# List of Abbreviations

Acronym - Expansion

- LSRT - Long Short-Term Relation Transformer
- G3RM - Global Gated Graph Reasoning Module
- LSTG - Long Short-Term Graph
- STFT - Short-Time Fourier Transform
- TTS - Text To Speech
- LSTM - Long Short Term Memory
- CBHG - Convolutional Banks + Highway Networks + Gated Recurrent Units
- GRU - Gated Recurrent Unit
- HAR - Human Action Recognition
- CNN - Convolution Neural Network
- RNN - Recurrent Neural Network

# List of Figures

# Chapter 1

# Introduction

Human Action Recognition (HAR) represents a crucial and dynamic domain within computer vision and artificial intelligence, dedicated to creating algorithms and systems proficient in comprehending and analyzing human actions depicted in visual data. This encompasses a wide range of applications, including surveillance, human-computer interaction, healthcare monitoring, sports analysis, and augmented reality.

## 1.1 Background

Human action recognition is a pivotal area within computer vision that seeks to understand and interpret human movements from visual data, covering a broad spectrum of applications, from surveillance to assistive technologies. The integration of both vision transformers and recurrent transformers represents an approach to enhance the accuracy and temporal understanding of recognized actions. Vision transformers excel at capturing spatial features within static images, while recurrent transformers specialize in modeling sequential dependencies over time. This fusion enables a more comprehensive analysis of dynamic human activities, making it particularly relevant for applications catering to the visually impaired. In current scenarios, where technological advancements are rapidly reshaping the landscape of accessibility solutions, human action recognition plays a crucial role. For the visually impaired, accurate identification of actions in real-time can significantly contribute to their situational awareness, aiding in navigation, social interaction, and overall engagement with the environment. As societies strive for greater inclusivity, the importance of projects like these lies in their potential to empower individuals with visual impairments, offering them a more informed and independent lifestyle by leveraging the capabilities of advanced machine learning models.

## 1.2 Problem Definition

The visually impaired face substantial challenges in comprehending and navigating their surroundings due to a lack of real-time information about the actions and events occurring in their environment.Existing solutions often fall short in providing comprehensive, instantaneous insights into dynamic situations, leaving a critical gap in the accessibility and independence of visually impaired individuals. Recognizing human actions in real-time and converting this information into a format that is readily understandable by the visually impaired represents a significant challenge that, if addressed, could significantly enhance their daily lives.

## 1.3 Scope and Motivation

Our project ambitiously addresses the aforementioned challenges by introducing a novel system that integrates cutting-edge technologies. The scope encompasses the development of a Human Action Recognition module, leveraging vision transformer and recurrent transformer models to identify and classify a wide range of human actions.The project also involves the implementation of a Text-to-Speech Conversion system to translate the identified actions into coherent and accessible audio output.The intended deployment of this system in the real world broadens its scope, allowing for the recognition of real-time actions and events in diverse environments. The project's scope extends beyond the technical implementation to encompass the integration of the system into the daily lives of visually impaired individuals,offering them an unprecedented level of awareness and understanding of their surroundings.

The motivation driving this project is rooted in a commitment to improving the quality of life for visually impaired individuals by addressing the fundamental challenges they face in understanding their surroundings. The lack of real-time, accessible information about human actions poses a significant barrier to their independence and engagement with the world. By developing a sophisticated system that seamlessly recognizes and translates human actions into audio output, we aspire to empower the visually impaired community.This motivation stems from a belief in the transformative potential of technology to break down barriers, fostering inclusivity, and providing individuals with visual

impairments the tools they need to navigate the world confidently and independently.

## 1.4 Objectives

- The project aims to create a sophisticated assistive technology system that not only recognizes a diverse range of human actions but also empowers visually impaired individuals to navigate independently. Key objectives include fine-grained action recognition, multi-modal input processing, obstacle detection, localization, and mapping. The system will allow user customization, continually improve through machine learning, and collaborate with smart infrastructure for enhanced contextual information. It prioritizes low-latency interaction, accessibility, and inclusivity, while community engagement initiatives focus on providing training and resources to ensure effective utilization of the technology within the visually impaired community.

## 1.5 Assumptions and Challenges

### 1.5.1 Assumptions

- **Data Distribution:** Assuming that the training and deployment data follow a similar distribution.
- **Model Generalization:** Assuming that the models generalize well to unseen scenarios and diverse user inputs.
- **Stable Internet Connection:** Assuming a stable internet connection for utilizing external APIs or cloud services.
- **Consistent User Behavior:** Assuming relatively consistent patterns in human actions for effective recognition.

### 1.5.2 Challenges

- **Labeling Complexity:** Challenges in obtaining accurate and comprehensive labels for human actions in diverse real-world scenarios.

- **Adaptability to Environmental Changes:** Adapting the models to varying environmental conditions, lighting, and background noise.
- **Hardware Limitations:** Potential limitations in hardware resources, especially for real-time processing on edge devices.
- **User Diversity:** Recognizing and accommodating diverse user behaviors, accents, and speech patterns in speech-related tasks.

## 1.6 Societal / Industrial Relevance

**Social Relevance**

The system holds profound social relevance as it addresses a critical aspect of inclusivity by empowering visually impaired individuals. By offering real-time audio insights into human actions and environmental events, the system enhances the daily lives of the visually impaired, fostering independence and promoting a more inclusive society. This technological innovation contributes to breaking down barriers that hinder accessibility, allowing individuals with visual impairments to engage more actively with their surroundings. In a broader social context, the project aligns with the principles of equal opportunities, advocating for a world where everyone, regardless of visual abilities, can participate fully in various aspects of life.

**Industrial Relevance**

From an industrial perspective, the system aligns with the growing demand for innovative solutions that integrate machine learning and artificial intelligence into real-world applications. The Human Action Recognition module, powered by advanced vision and recurrent transformer models, represents a cutting-edge approach to action classification. The integration of a Text-to-Speech Conversion system further positions the project at the forefront of accessibility technology. Industries focused on assistive technologies, artificial intelligence, and accessibility solutions stand to benefit from the development and implementation of such a system. Additionally, the potential for real-world deployment opens avenues for collaboration with organizations committed to improving the quality of life for individuals with visual impairments, enhancing the system's industrial relevance and societal impacts.

## 1.7 Organization of the Report

The report is divided into several essential sections. The introduction provides a background,followed by problem definition, outline of the scope and motivation behind the project. Following this, the assumptions made throughout and the challenges faced are mentioned. A thorough literature survey is conducted, analyzing five seminal papers relevant to the subject matter which includes the base paper. The subsequent section involves the requirements, system overview and architecture and the work schedule. Finally, the conclusion is made outlining the future scope and extensions.

## 1.8 Chapter Summary

The project focuses on Human Action Recognition (HAR), utilizing advanced vision and recurrent transformer models to enhance accuracy, particularly beneficial for the visually impaired. It ambitiously introduces a system recognizing a diverse range of human actions and implementing a Text-to-Speech Conversion system for real-time applications. Motivated by a commitment to improving the lives of visually impaired individuals, the project aims to empower them by providing real-time, accessible information about their surroundings. Objectives include fine-grained action recognition, multi-modal input processing, and continual improvement through machine learning. The report emphasizes societal and industrial relevance, outlining a systematic organization covering background, problem definition, scope, motivation, assumptions, challenges, and societal/industrial impact.

# Chapter 2

# Literature Survey

## 2.1 ViT-ReT For Human Action Recognition

The research [2] introduces and develops two transformer-based neural networks for recognizing human activities: a recurrent transformer (ReT), which is designed for making predictions on sequential data, and a vision transformer (ViT), which is optimized to efficiently extract important features from images. This approach aims to enhance the speed and scalability of activity recognition tasks.

### 2.1.1 Vision Transformer[1]

The feature extraction process using ViT (Vision Transformer) involves dividing the input video frames into fixed-size patches, which are then linearly embedded to create tokenized representations. These tokenized representations are then processed through a transformer encoder to capture spatial information and extract meaningful features from the video frames. Given below is a detailed explanation of the feature extraction process using ViT.

1. **Dividing video frame into patches**
   Initially the video frames is divided into fixed size patches. This is done to enable the model to process the video frames in a more efficient manner. The size of the patches is typically chosen to be small enough to capture fine-grained details in the video frames, but large enough to avoid excessive computational overhead.

2. **Linear Embedding of Patches**
   After dividing the video frames into patches, each patch is linearly embedded to create a tokenized representation. This is done by applying a linear projection to each patch, which maps the patch to a lower-dimensional space. The resulting

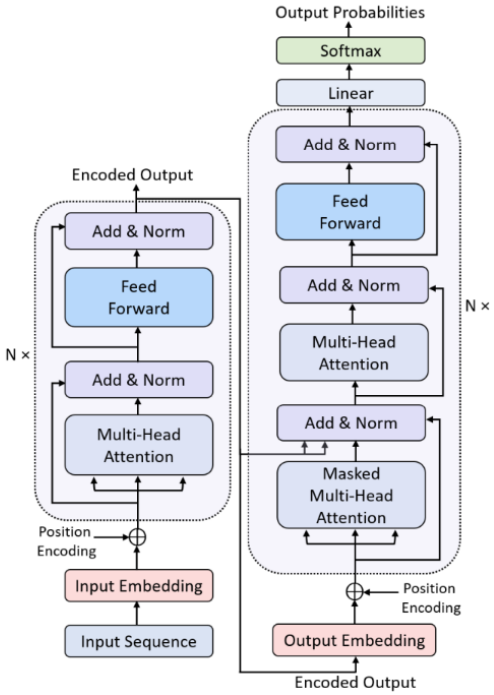



Figure 2.1: ViT-ReT model architecture

embeddings are then concatenated to form a sequence of tokenized representations.

3. **Processing Tokenized Representations using Transformer Encoder**

   The tokenized representations are then processed through a transformer encoder to capture spatial information and extract meaningful features from the video frames. The transformer encoder comprises several tiers of self-attention and feedforward neural network tiers. Self-attention layer enable the model to focus on various segments of the input sequence, while feedforward neural network layer aid in learning complex connections among tokenized representations.

4. **Aggregating Features from Transformer Encoder**

   After processing the tokenized representations through the transformer encoder, the resulting features are aggregated to form a fixed-length representation of the input video frames. This is typically done by taking the mean or max of the features across the sequence of tokenized representations. The resulting fixed-length representation can then be used as input to downstream tasks, such as sequence modeling using ReT.

### 2.1.2 Recurrent Transformer

Sequence modeling using ReT (Recurrent Transformer) involves leveraging the encoded representations from ViT for capturing temporal dependencies and sequential patterns within the video data. The ReT model applies recurrent transformer-based decoding to effectively model the temporal dynamics of human activities in the videos.Given below is a detailed explanation.

1. **Encoded Representations from ViT**

   The process begins with the encoded representations obtained from the ViT feature extraction stage. These representations capture spatial information and meaningful features from the video frames, providing a rich source of information for modeling temporal dependencies and sequential patterns.

2. **Recurrent Transformer-Based Decoding**

   The ReT model utilizes a recurrent transformer-based decoding architecture to capture temporal dependencies within the encoded representations. This involves processing the encoded representations through a series of recurrent transformer layers, each of which captures sequential patterns and temporal dynamics within the video data.

3. **Masked Multi-Headed Attention**

   Within the recurrent transformer layers, masked multi-headed attention mechanisms are employed to enable the model to focus on different segments of the input sequence while preventing information flow from future time steps. This enables the model to effectively capture dependencies within the sequence while maintaining causality.

4. **Positional Encoding**

   To incorporate positional information into the sequence modeling process, positional encoding is added to the encoded representations. This allows the model to differentiate between the positions of the encoded features within the sequence, enabling it to capture temporal dynamics and sequential patterns more effectively.

5. **Residual Connections and Feedforward Neural Networks**

   The recurrent transformer layers also incorporate residual connections and feedforward neural network layers to facilitate the learning of non-linear relationships and higher-level abstractions within the temporal dynamics of the video data. Residual connections help in mitigating the vanishing gradient problem, while feedforward neural network layers enable the model to capture complex temporal patterns.

In short, the ViT-ReT methodology outlined in the paper offers a potent technique for recognizing human activities in videos. The methodology involves using ViT for feature extraction and ReT for sequence modeling, allowing the model to effectively capture spatial information, extract meaningful features, and model temporal dynamics and sequential patterns within the video data.

## 2.2 A Novel Two-Stream Transformer-Based Framework

The paper[3] introduces a particular multimodal approach using RGB data and skeleton data along with fusion stratergies. The framework consists of two main streams: the RGB stream and the skeleton stream. Each stream processes data from its respective modality using a modified Transformer-based architecture tailored to the characteristics of the input data.

1. **RGB Stream:**

   - In this stream RGB videos is processed using a pure Transformer-based architecture. It takes input in the form of RGB frames at a low frame rate, capturing visual information from the videos.
   - The Transformer-based architecture for the RGB stream is designed to capture spatial and appearance features from the RGB frames, leveraging the strengths of the visual modality for action recognition.

2. **Skeleton Stream:**

   In this stream skeleton data is processed using a modified Transformer-based architecture with a specific focus on capturing motion and temporal information.It operates at a higher temporal resolution and focuses on motion-related features.This stream aims to capture the essential dynamics of human actions, complementing the spatial and appearance information processed by the RGB stream within the overall framework. The skeleton stream operates on heatmap frames derived from skeleton data. The proposed framework integrates the processed features from the skeleton stream with those from the RGB stream using fusion methods, allowing for the complementary characteristics of the two modalities to be effectively combined for improved action recognition performance

### 2.2.1 Fusion Methods

The RGBSformer framework incorporates two fusion methods for integrating information from the RGB and skeleton streams:

1. **Early Fusion:**

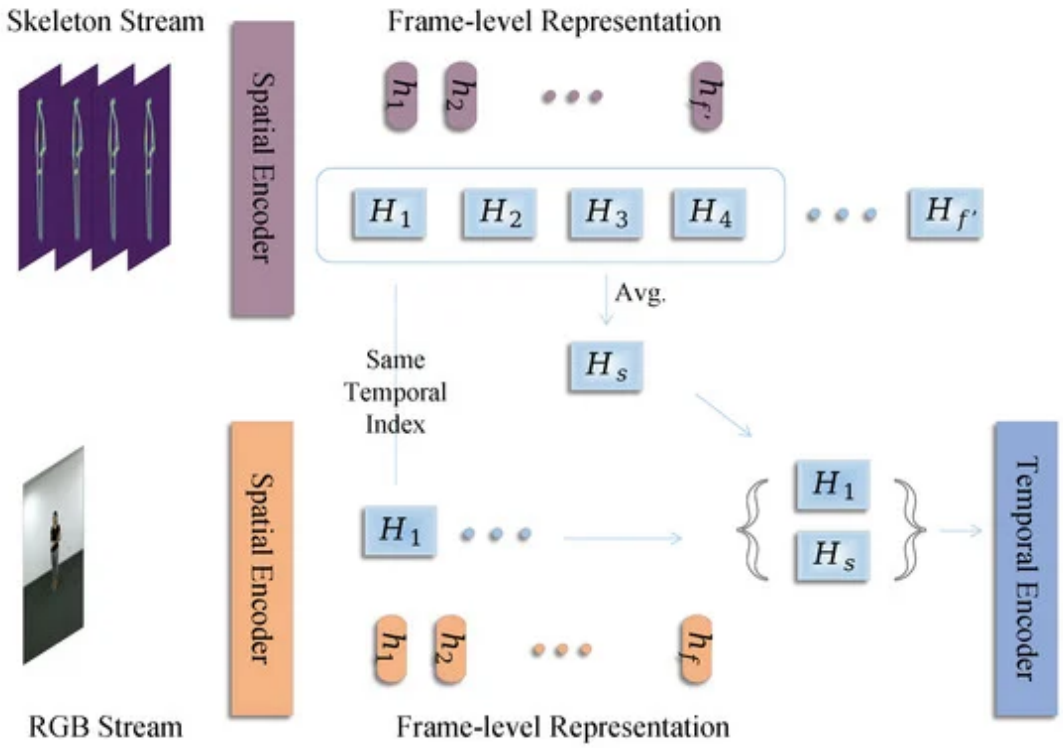



Figure 2.2: Late fusion of RGB and skeleton data

In early fusion, the features from the RGB stream and the skeleton stream are concatenated at the input level before being processed by the respective Transformer-based architectures. This allows for the combined information from both modalities to be considered from the beginning of the processing pipeline.

2. **Late Fusion:**

   In late fusion as shown in figure 3.5, the features from the two streams are concatenated at the output level after being processed by their respective Transformer-based architectures. This approach enables the integration of the processed features from both modalities at a later stage in the framework.

The proposed framework aims to utilize the combined strengths of RGB and skeleton modalities for action recognition. It seeks to harness the detailed visual information offered by RGB videos alongside the motion-specific data captured by skeleton representations. By effectively integrating these modalities using the two-stream Transformer-based approach and the fusion methods, the RGBSformer framework demonstrates state-of-the-art performance in action recognition tasks across multiple benchmarks.

## 2.3 Enhanced Video Captioning via Global Gated LSRT

The paper[4] discusses the importance of understanding spatio-temporal relationships among objects in videos for accurate video captioning.A Long Short-Term Graph to capture both short-term spatial semantic relations and long-term transformation dependencies in videos is designed. To reason about these relations, a Global Gated Graph Reasoning Module [4]that uses global context to control information propagation between objects and alleviate relation ambiguity is constructed. Finally, the Long Short-Term Relation Transformer is setup by incorporating G3RM into the Transformer model to fully explore object relations for caption generation.

### 2.3.1 Methodology

- **Transformer encoder with a Global Context**

  The Transformer global context encoder is responsible for enhancing the display of visual appearance and motion characteristics in the video. It is part of the LSRT model, which consists of three parts: a global context Transformer encoder, a graph reasoning encoder, and a coarse-to-fine language decoder. Transformer's global context encoder helps capture global context information that can be used to better understand video content. By incorporating the Transformer global context encoder, the LSRT model aims to improve the representation of appearance and motion features, resulting in better video caption generation performance.

- **Graph Reasoning Encoder**

  The component responsible for encoding the graph-based reasoning information is part of the Long Short Relation Transformer (LSRT) model for video subtitling. It replaces the self-observation mechanism with the Global Graph Reasoning Module (G3RM). A graph reasoning encoder takes as input the properties of an object and a previous long-short-term graph (LSTG). It performs relational reasoning over a graph, making it easy to capture both immediate spatial semantic relationships and persistent transformational dependencies. G3RM implements a new global gate mechanism using global context to control the flow of information between objects

and mitigate ambiguity in relationships. Incorporating the encoder's graph reasoning, the LSRT model focuses on extracting complex object relationships for caption generation, addressing issues such as redundant connections, over-smoothing, and ambiguity in relationships.

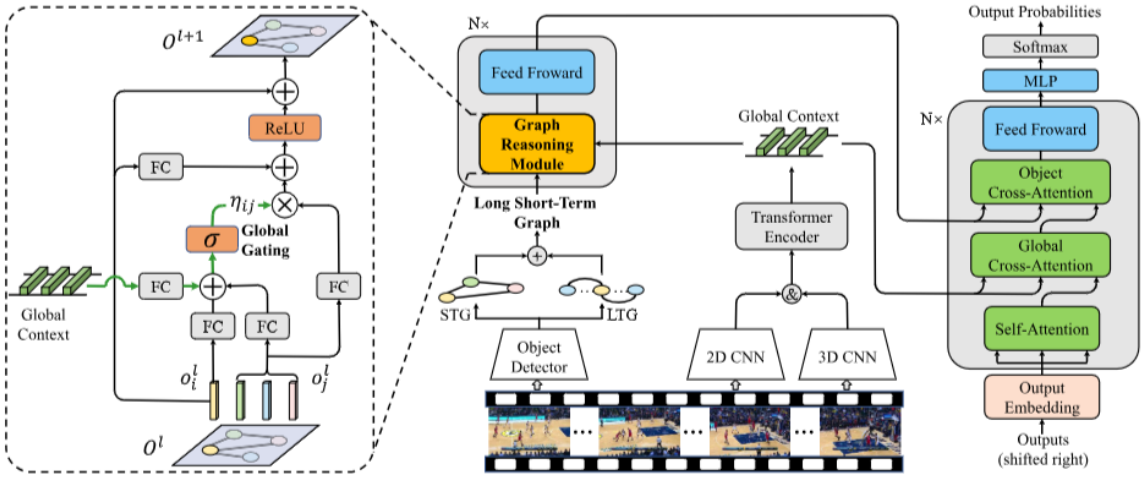



Figure 2.3: Architectural Overview of LSRT with Global Gating

- **Coarse-to-fine language Decoder**

  The decoder follows the principle of "coarse to fine" and uses both the properties of the wide scene and the properties of detailed objects to generate subtitles. The decoder uses a global cross-attention mechanism to generate attended scene features and a self-observation mechanism to generate a self-attended sentence representation. The attention-grabbing features of the scene are then used to obtain the relevant features of the object, and combining these two attended features is achieved by residual concatenation to form the current word. The decoder consists of several layers and the predicted probability of the current word is calculated based on the ground truth caption using cross entropy loss. Experimental results on the used datasets [5] demonstrate the effectiveness of the LSRT model, showing improved performance and mitigation of over-smoothing while increasing capacity for relational reasoning.

## 2.4 Text to Speech Synthesis using Tacotron 2

The study "Natural synthesis of TTS by adjusting WaveNet on the mel-spectrogram predictions "[6] describes the Tacotron 2, a neural network design adapted for speech synthesis directly from text input. It combines a network of feature predictions between individual sequences work with a modified WaveNet vocoder to generate mel-scale spectrograms and synthesize curves in the time domain.

### 2.4.1 Methodology

- **Intermediate Feature Representation**

  The intermediate feature representation used in the paper is mel-frequency spectrograms, which are derived from either the linear-frequency spectrogram or the magnitude of the short-time Fourier transform (STFT).Mel-spectrograms are acquired through a nonlinear transformation applied to the frequency axis of the STFT, drawing inspiration from the observed responses of the human auditory system. Mel spectrograms condense the frequency information into fewer dimensions, prioritizing details in lower frequencies crucial for speech intelligibility, while minimizing the emphasis on high-frequency details primarily comprising noise bursts. They serve as a more basic, fundamental acoustic representation of audio signals compared to the linguistic and acoustic features employed in WaveNet.The use of mel spectrograms as the conditioning input to WaveNet simplifies the acoustic representation and enables a substantial decrease in the size of the WaveNet architecture.Despite discarding phase information and presenting a challenging inverse problem, mel spectrograms can be used effectively as a neural vocoder to generate high-quality audio.

- **Network for predicting spectrograms**

  The spectrogram prediction network is a key component of the Tacotron 2 model.The structure comprises an encoder and a decoder incorporating an attention mechanism.The encoder transforms a character sequence into a hidden feature representation through the utilization of convolutional layers and a bi-directional LSTM layer. The decoder consumes the encoded features and predicts a mel spectrogram frame by frameutilizing an autoregressive recurrent neural network.The output of the de-

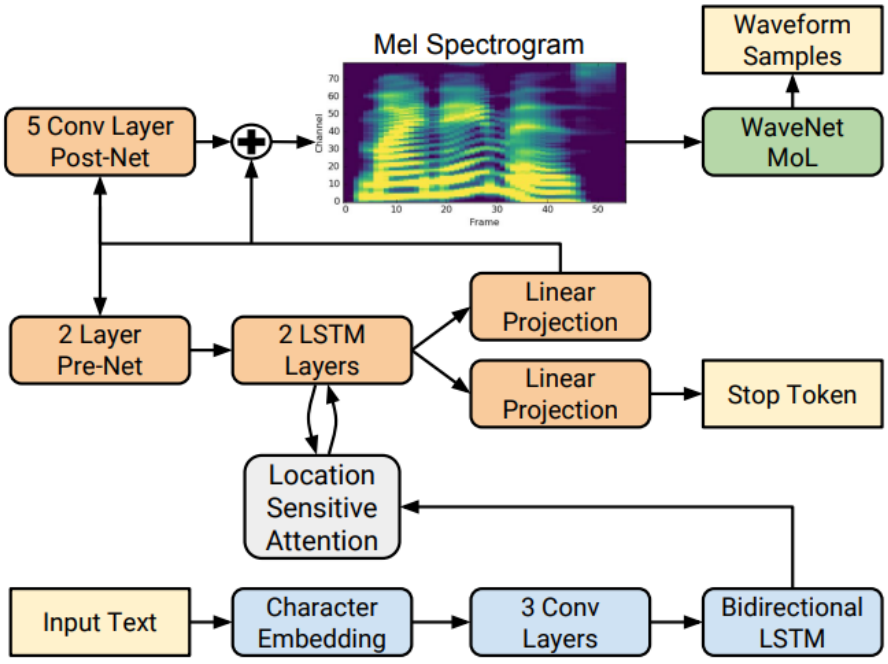



Figure 2.4: Tacotron 2 System Architecture

coder LSTM and the attention context vector are combined and processed through a linear transformation to forecast the target spectrogram frame.The forecasted mel spectrogram undergoes further enhancement through a post-net, which predicts a residual to refine the overall reconstruction.The network is regularized using dropout and zoneout techniques to introduce output variation and prevent overfitting .During training, the spectrogram prediction network operates in teacher-forcing mode, meaning each predicted frame is conditioned on both the encoded input sequence and the previous frame from the ground truth spectrogram.

- **WaveNet Vocoder**

  WaveNet is a modified version of the original WaveNet model used as a vocoder in the Tacotron 2 system.The WaveNet vocoder is responsible for synthesizing time-domain waveforms from the mel spectrograms predicted by the spectrogram prediction network.It is a deep generative model based on dilated causal convolutions, which allows it to capture the conditional distribution of the audio waveform based on the mel spectrogram.WaveNet generates audio samples autoregressively, meaning that it predicts one sample at a time conditioned on the previous samples and

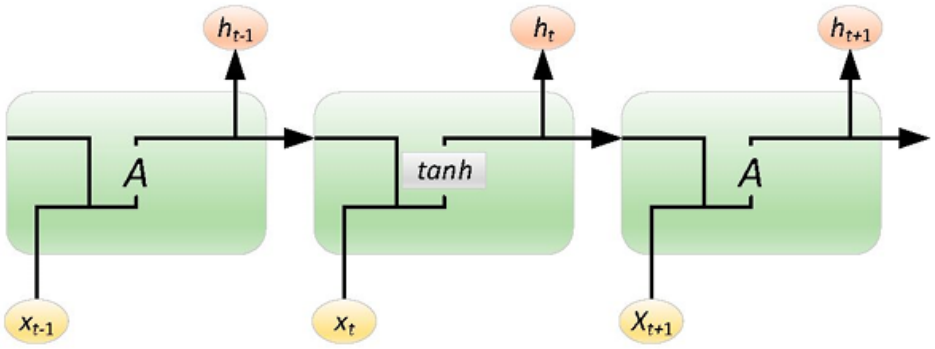



Figure 2.5: The repeating module in a standard RNN which has a single layer.

the mel spectrogram.The model is trained using a variant of the softmax-based categorical cross-entropy loss, where the target is the quantized waveform sample.By conditioning WaveNet on the predicted mel spectrograms, the system is able to synthesize natural-sounding speech with high audio quality.Leveraging mel spectrograms as the conditioning input to WaveNet streamlines the architecture, leading to a substantial reduction in its size. This adaptation of the WaveNet vocoder in Tacotron 2 attains sound quality approaching that of natural human speech, representing a state-of-the-art achievement.

## 2.5 Human Activity Recognition Using CNN and LSTM

The method[7] provides a comprehensive overview of Human Activity Recognition (HAR) and its diverse applications in fields like healthcare, personal fitness, gaming, and security. It underscores the increasing interest in Recurrent Neural Networks (RNN), particularly within HAR, citing their successful applications in speech recognition, language modeling, video processing, and time series analysis. The integration of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) is introduced as an effective approach for predicting human behaviors. Emphasis is placed on the pivotal role of sensors, such as accelerometers and gyroscopes, in capturing human activities and the necessity for efficient models to process and analyze the collected data.

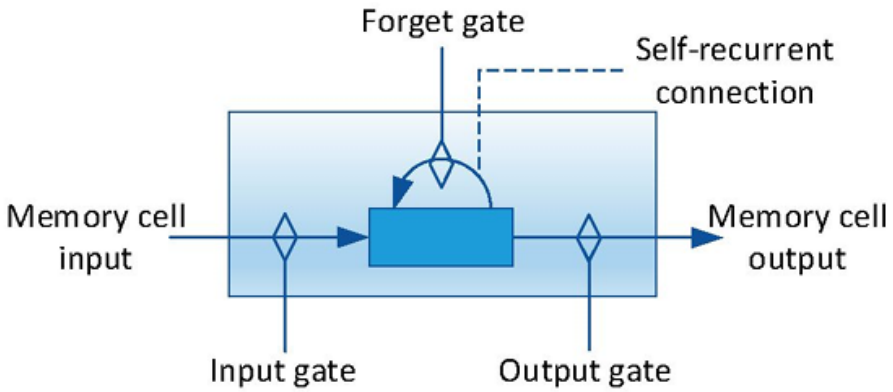



Figure 2.6: Illustration of an LSTM memory cell

### 2.5.1 Method

An effective approach for recognizing human activities involves utilizing a blend of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models. The methodology is initiated by constructing a foundational CNN model designed for activity recognition, followed by the development of a distinct LSTM network model tailored to their dataset. Subsequently, they propose a hybrid model merging CNN and LSTM architectures for the purpose of classifying and predicting six different activities. Furthermore, an innovative ConvLSTM model, an extension of the CNN LSTM model, is introduced to seamlessly integrate convolutions within the LSTM framework. The paper addresses challenges encountered by human activity recognition systems, including action variability, class similarity, time consumption, and a high proportion of null values. To tackle these issues, the authors advocate for the utilization of a deep convolutional network incorporating both CNN and LSTM components.

## 2.6 Existing Systems

Various methods are employed in Human Activity Recognition (HAR). Some noteworthy approaches include:

1. **Recurrent Neural Networks (RNN)**

   Widely used for tasks like speech recognition and time series analysis, RNNs are neural network architectures designed to handle sequential data. They utilize feedback loops to process information in a sequential manner, making them suitable for recognizing patterns and dependencies over time.

2. **Deep Convolutional Networks with CNN and LSTM**

   Combining Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks has proven effective for sequential activity recognition. CNNs excel at spatial feature extraction, while LSTMs handle temporal dependencies. However, this approach may face efficiency challenges due to the computational complexity of both architectures.

3. **Statistical Learning Methods**

   Statistical learning methods, such as Naive Bayes and K-Nearest Neighbor (KNN)[8], have been employed for HAR. These methods often require expert knowledge for feature design, making them sensitive to the quality of handcrafted features. Naive Bayes is known for its simplicity, while KNN relies on proximity-based classification.

4. **Mobile Sensing Advances**

   Advancements in mobile sensing[9] leverage smartphone and sensor data to quantify diverse human behaviors. This approach is particularly attractive due to the ubiquity of smartphones. Through gathering data from diverse built-in sensors like accelerometers and gyroscopes, it becomes feasible to deduce human activities in a way that doesn't intrude upon the individual.

5. **Stacked LSTM Architectures**

   Stacked LSTM architectures[10] are implemented for feature-free activity classification using accelerometer and gyroscope signals. These architectures involve multiple

layers of LSTMs, enabling the model to learn hierarchical representations of sequential data. This approach has shown promise in capturing complex dependencies in activity data.

6. **CNN and LSTM Approaches**

   The combination of CNN and LSTM approaches has been employed for HAR from inertial sensor time series. This method showcases efficacy in handling sequential data, where CNNs extract spatial features from sensor data, and LSTMs model temporal dependencies. This approach is well-suited for capturing both short-term and long-term patterns in activity sequences.

These existing systems demonstrate the diversity of approaches in Human Activity Recognition, ranging from traditional statistical methods to advanced deep learning architectures. The selection of a method frequently relies on the particular attributes of the data and the demands of the application.

# Chapter 3

# Requirements

- **Hardware Requirements:**
  - **GPU:**

    A powerful GPU with CUDA support, e.g., NVIDIA GeForce RTX 30 series or A100.
  - **GPT-3 API:**
    - Internet Connection: Stable and fast internet connection.
  - **Speech-to-Text (STT) and Text-to-Speech (TTS):**
    - Microphone (for STT)
    - Speakers or headphones (for TTS)
  - **Camera:**
    - Webcam or camera with video capture capabilities.
- **Software Requirements:**
  - **Vision Transformer and Recurrent Transformer:**
    - Deep Learning Framework: TensorFlow or PyTorch.
    - Libraries: Hugging Face Transformers Library (for ViT).
  - **GPT-3 API:**
    - OpenAI API Key: Obtain a key from OpenAI for API access.
  - **Speech-to-Text (STT) and Text-to-Speech (TTS):**
    - SpeechRecognition library (for STT)
    - pyttsx3 library (for TTS)

# Chapter 4

# System Architecture

The project introduces a system architecture meticulously crafted to provide real-time assistance. At its core, the architecture comprises two integral modules: the Human Action Recognition module and the Text-to-Speech Conversion system. The Human Action Recognition module employs advanced video preprocessing techniques and leverages vision and recurrent transformer (ViT-ReT) models to identify and classify human actions, offering a comprehensive understanding of dynamic elements in the environment. Simultaneously, the Text-to-Speech Conversion system transforms identified actions into clear and coherent audio output, ensuring real-time and intelligible auditory feedback. These modules are seamlessly integrated, forming a cohesive system poised for real-world deployment, with a user-centric design philosophy that includes customization options and a continuous feedback mechanism. The architecture is designed to be adaptable, incorporating multimodal fusion techniques for a comprehensive understanding of the user's surroundings and setting the stage for collaborative innovation and scalability in the future.

## 4.1 System Overview

**1. Human Action Recognition :**

**1.1 Objective:**

The Human Action Recognition module serves as the foundational element, aiming to empower visually impaired individuals with real-time insights into their surroundings. The primary objective is to identify and classify human actions, facilitating a nuanced understanding of dynamic elements and events.

**1.2 Implementation:**

Harnessing the capabilities of cutting-edge vision transformer and recurrent transformer

models, this module operates in a dual capacity, capturing spatial intricacies and temporal dependencies. The vision transformer focuses on spatial features, while the recurrent transformer addresses temporal nuances, ensuring a comprehensive representation of human actions.

**1.3 Training Dataset:**

Critical to the success of this module is a meticulously curated training dataset. The datasets; HMDB 15 and UCF101, spans diverse scenarios and encompasses a broad spectrum of human actions, ensuring the models are trained to recognize and classify actions across various real-world contexts.

**2. Text-to-Speech Conversion System:**

**2.1 Objective:**

Following the identification of human actions, the Text-to-Speech Conversion system takes the textual representations and converts them into coherent and natural audio output. The primary aim is to provide visually impaired users with real-time, auditory descriptions of the identified actions.

**2.2 Implementation:**

Employing state-of-the-art text-to-speech conversion algorithms, this system is meticulously designed for real-time performance. The emphasis lies in delivering audio descriptions that are not only timely but also articulate and comprehensible, ensuring an enriched user experience.

**3. Real-world Implementation:**

**3.1 Integration:**

The seamless integration of the Human Action Recognition and Text-to-Speech Conversion modules is at the heart of the system. This integrated solution is poised for real-world deployment, where visually impaired individuals navigate daily.

**3.2 Hardware Considerations:**

To ensure widespread usability, the system is optimized to run efficiently on devices equipped with cameras and audio output capabilities. This consideration prioritizes portability and user-friendly interactions.

**4. User Interaction:**

**4.1 User Interface:**

A user-centric design philosophy is embraced, materializing in an intuitive interface that enhances user interaction. This interface not only provides additional information but also allows for customization, catering to individual preferences and needs.

**4.2 Feedback Mechanism:**

User feedback is not merely a post-deployment formality but an integral part of the system's ongoing refinement. The architecture is engineered to incorporate user input, ensuring continuous adaptation and improvement in user experience.

**5. Multimodal Fusion:**

**5.1 Enhanced Understanding:**

Multimodal fusion techniques take center stage, amalgamating information from various sources, including video frames and audio cues. This sophisticated approach enriches the system's understanding of the user's surroundings, contributing to a more comprehensive and contextually aware output.

**5.2 Adaptability:**

By incorporating multimodal fusion, the system becomes inherently adaptable to diverse environmental conditions. This heightened adaptability enhances the system's ability to provide accurate and meaningful information in various real-world scenarios.

**6. Future Scope:**

**6.1 Scalability:**

A modular architecture is intentionally crafted to accommodate future advancements and enhancements. This scalability ensures the system's longevity and relevance, making it amenable to future technological evolution and emerging challenges.

**6.2 Collaboration Opportunities:**

The modular design not only enables scalability but also fosters collaboration. The system is positioned for partnerships with organizations and stakeholders in the accessibility domain, creating avenues for innovation and continuous improvement.

## 4.2 Architectural Design

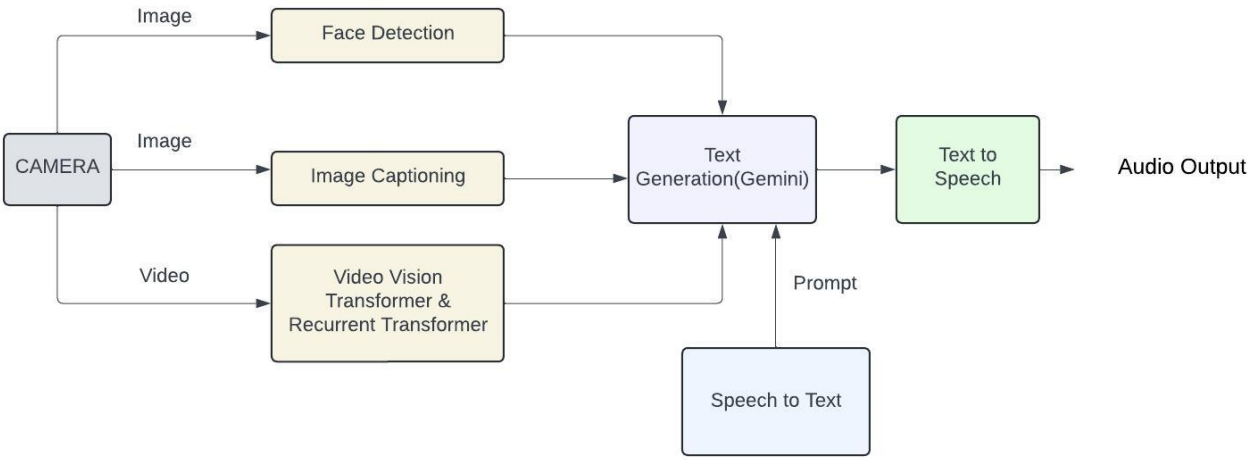



Figure 4.1: Architecture diagram

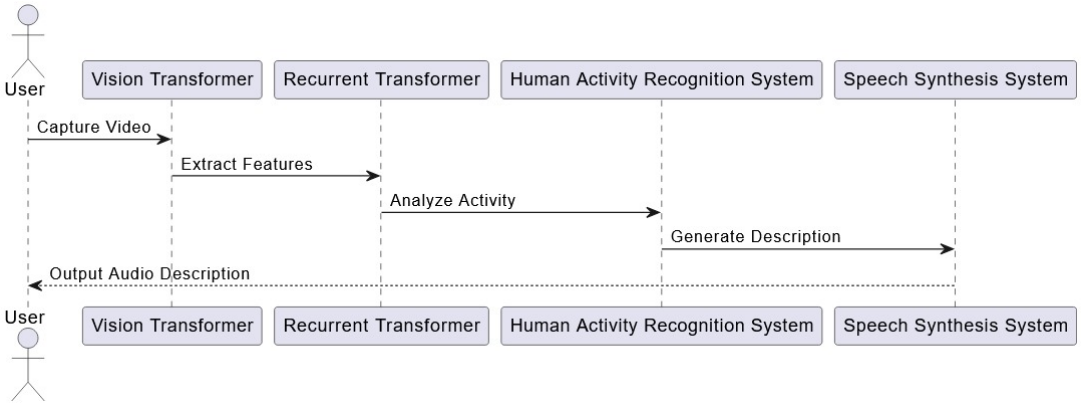



Figure 4.2: Sequence Diagram

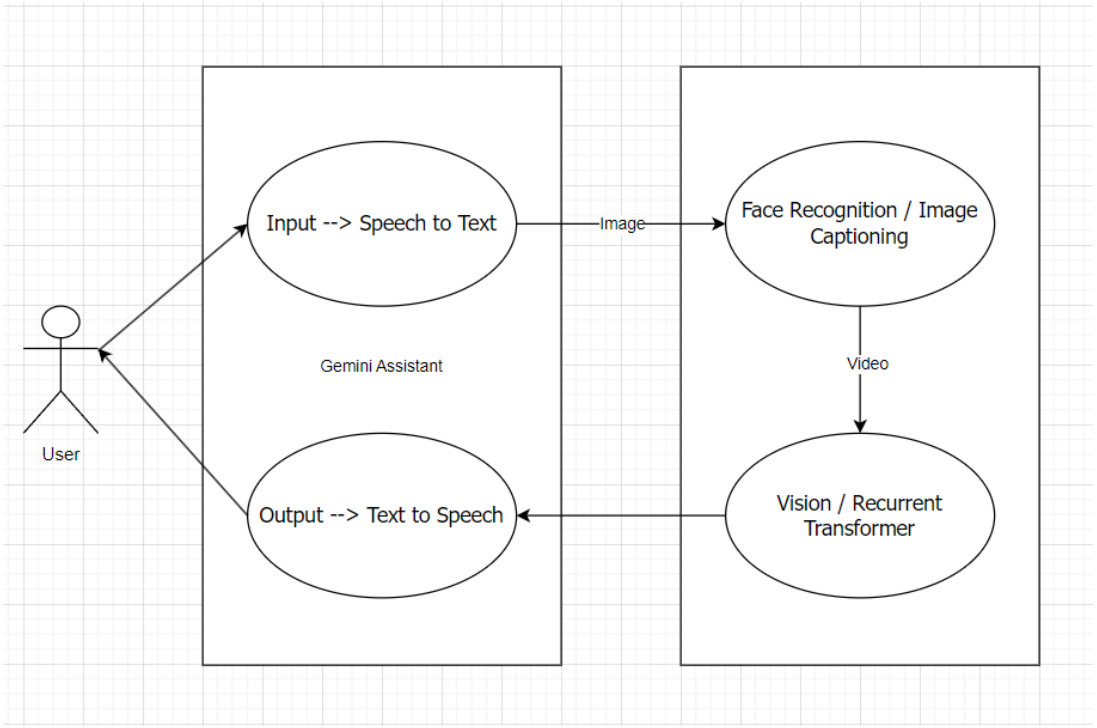



Figure 4.3: Sequence Diagram

## 4.3 Module Division

The entire project is divided into 4 modules:

1. **Video Preprocessing**

   In this the video is divided into frames.Each frame is converted into size 224*224 and into RGB colour format.

2. **Vision Transformer**

   In vision transformer, the frames are divided into patches and fed into the encoder.As output we get feature vectors.

3. **Recurrent Transformer**

   The feature vectors from ViT is fed into ReT.From this we get classification of human action as the output.

4. **Speech Synthesis**

   In this module we convert the caption generated in the previous module into speech.

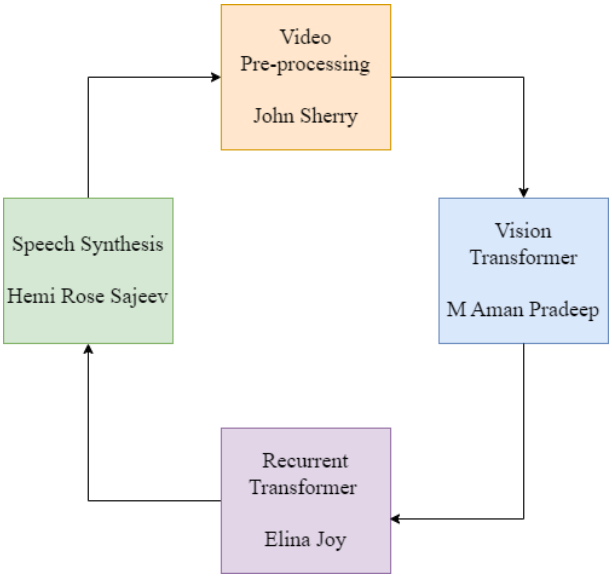



Figure 4.4: Module Division per team member

## 4.4 Work Breakdown and Responsibilities

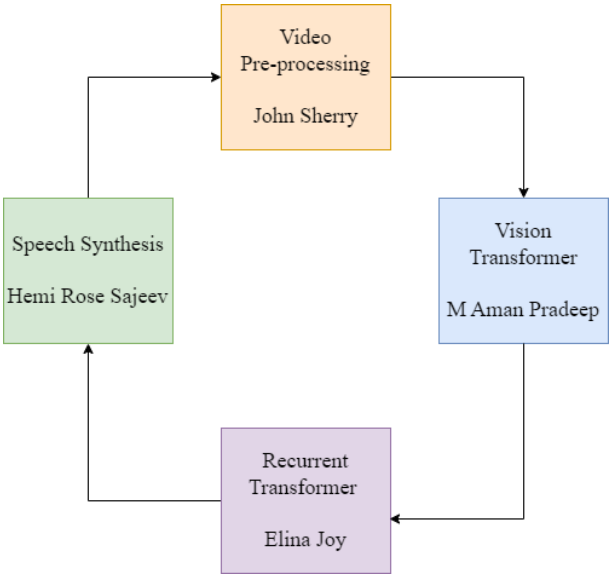



Figure 4.5: Workbreakdown and Responsibilities

## 4.5 Work Schedule - Gantt Chart

| TASKS | Month 1 | Month 2 | Month 4 | Month 6 | Month 8 |
|---|---|---|---|---|---|
| **Step 1:** Data Preprocessing | Completed dataset processing | | | | |
| **Step 2:** Vision Transformer | | Started model tuning and trainng | | | Still working on it as accuracy is not good |
| **Step 3:** Image Description | | Started to work on image descrpition model | Completed image captioning(60% output) | | |
| **Step 4:** Facial Recognition | | | started working on face recpgnition | Trained with group members facial data | |
| **Step 5:** Inference Model / Speech Synthesis | | | Started working on speech model | Completed speech model | |
| **Step 6:** Integration | | | Integration of face and image description model | Integrated speech/inference model | Final evaluation with video vision pending |

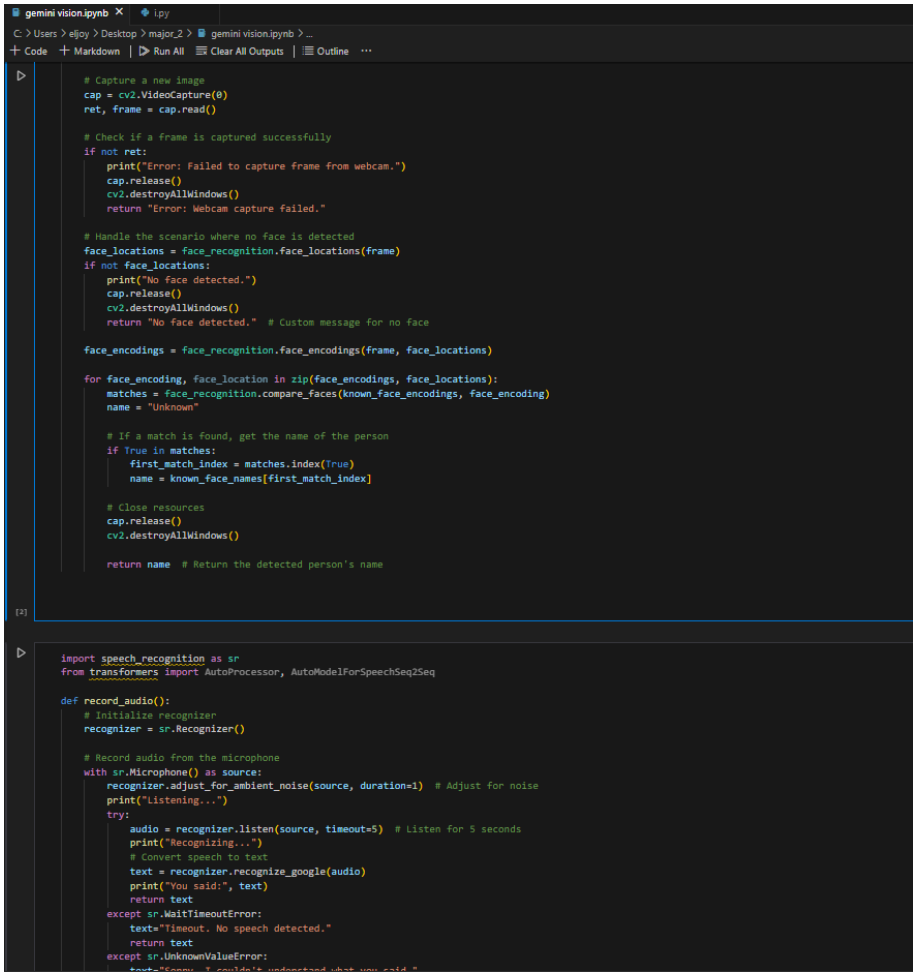
Figure 4.6: Project Timeline

# Chapter 5

# Results

gemini vision.ipynb

```
import face_recognition
import os
import pickle
import cv2

def add_new_face(person_name ):
    """
    Captures images of a person and saves them along with their extracted face encodings for training.
    """

    known_faces_dir = "known_faces"
    os.makedirs(known_faces_dir, exist_ok=True)  # Create directory if it doesn't exist


    person_dir = os.path.join(known_faces_dir, person_name)
    os.makedirs(person_dir, exist_ok=True)  # Create directory for the person

    num_images = 10

    known_face_encodings = []  # Initialize an empty list for face encodings
    known_face_names = []      # Initialize an empty list for person names

    for i in range(num_images):
        print(f"Taking picture {i+1} of {person_name}")
        cap = cv2.VideoCapture(0)  # Use default webcam
        ret, frame = cap.read()

        # Ensure a face is detected before saving the image
        while True:
            faces = face_recognition.face_locations(frame)
            if len(faces) > 0:
                break
            ret, frame = cap.read()

        cv2.imwrite(os.path.join(person_dir, f"{person_name}_{i}.jpg"), frame)
        cap.release()
        cv2.destroyAllWindows()

        # Extract face encoding from the captured image (within the loop)
        face_encoding = face_recognition.face_encodings(frame)[0]
        known_face_encodings.append(face_encoding)
        known_face_names.append(person_name)

    # Save face encodings and names to a file
    with open("known_faces.pkl", "wb") as f:
        pickle.dump((known_face_encodings, known_face_names), f)

    print(f"Training data collected and saved for {person_name}.")

def recognize_face():
    """
    Attempts to load known faces from a file and performs face recognition on a captured image.
    Returns the detected person's name (or a custom message if no face is found).
    """

    # Try loading face encodings from a file
    try:
        with open("known_faces.pkl", "rb") as f:
            known_face_encodings, known_face_names = pickle.load(f)
        print("Loaded known faces from file.")
    except FileNotFoundError:
        print("No known faces file found. Please add faces using the 'add_new_face' function.")
        return "Unknown"
```


gemini vision.ipynb
i.py
C: > Users > eljoy > Desktop > major_2 > gemini vision.ipynb > ...
Code
Markdown
Run All
Clear All Outputs
Outline
# Capture a new image
cap = cv2.VideoCapture(0)
ret, frame = cap.read()
# Check if a frame is captured successfully
if not ret:
print("Error: Failed to capture frame from webcam.")
cap.release()
cv2.destroyAllWindows()
return "Error: Webcam capture failed."
# Handle the scenario where no face is detected
face_locations = face_recognition.face_locations(frame)
if not face_locations:
print("No face detected.")
cap.release()
cv2.destroyAllWindows()
return "No face detected." # Custom message for no face
face_encodings = face_recognition.face_encodings(frame, face_locations)
for face_encoding, face_location in zip(face_encodings, face_locations):
matches = face_recognition.compare_faces(known_face_encodings, face_encoding)
name = "Unknown"
# If a match is found, get the name of the person
if True in matches:
first_match_index = matches.index(True)
name = known_face_names[first_match_index]
# Close resources
cap.release()
cv2.destroyAllWindows()
return name # Return the detected person's name
[3]
import speech_recognition as sr
from transformers import AutoProcessor, AutoModelForSpeechSeq2Seq
def record_audio():
# Initialize recognizer
recognizer = sr.Recognizer()
# Record audio from the microphone
with sr.Microphone() as source:
recognizer.adjust_for_ambient_noise(source, duration=1) # Adjust for noise
print("Listening...")
try:
audio = recognizer.listen(source, timeout=5) # Listen for 5 seconds
print("Recognizing...")
# Convert speech to text
text = recognizer.recognize_google(audio)
print("You said:", text)
return text
except sr.WaitTimeoutError:
text="Timeout. No speech detected."
return text
except sr.UnknownValueError:


import textwrap
import pickle
from IPython.display import display, Markdown
import os
import PIL.Image
import google.generativeai as genai
def to_markdown(text):
text = text.replace('•', '*')
return Markdown(textwrap.indent(text, '> ', predicate=lambda _: True))
GOOGLE_API_KEY='AIzaSyAc1f275cooxoWBAM5G-c8pxWic3Rv4X0I'
genai.configure(api_key=GOOGLE_API_KEY)
# Check if the model is already saved
model_path = 'vision_model.pkl'
if os.path.exists(model_path):
with open(model_path, 'rb') as f:
vision_model = pickle.load(f)
print("Model loaded from", model_path)
else:
vision_model = genai.GenerativeModel('gemini-pro-vision')
# Save the model
with open(model_path, 'wb') as f:
pickle.dump(vision_model, f)
print("Model saved to", model_path)
Model loaded from vision_model.pkl
import pathlib import textwrap import google.generativeai as genai from IPython.display import display, Markdown def to_markdown(text): text=text.replace('•','*') return Markdown(textwrap.indent(text,'> ',predicate=lambda _: True)) import os GOOGLE_API_KEY='AIzaSyAc1f275cooxoWBAM5G-c8pxWic3Rv4X0I'
genai.configure(api_key=GOOGLE_API_KEY) import PIL.Image
vision_model=genai.GenerativeModel('gemini-pro-vision')

```python
from gtts import gTTS
import pygame
import io

def text_to_speech(text, lang='en'):
    tts = gTTS(text=text, lang=lang)

    # Create an in-memory file-like object
    audio_file = io.BytesIO()

    # Save audio to the in-memory file-like object
    tts.write_to_fp(audio_file)
    audio_file.seek(0)  # Move to the beginning of the file

    pygame.mixer.init()
    pygame.mixer.music.load(audio_file)
    pygame.mixer.music.play()

    while pygame.mixer.music.get_busy():
        pygame.time.Clock().tick(10)

    # Clean up
    pygame.mixer.quit()

# Example usage
text = "Hello, this is a test."
text_to_speech(text)
```

```
pygame 2.5.2 (SDL 2.28.3, Python 3.11.9)
Hello from the pygame community. https://www.pygame.org/contribute.html
```

```python
import cv2

def capture_image():
    # Open the first webcam device
    cap = cv2.VideoCapture(0)

    if not cap.isOpened():
        print("Error: Unable to open webcam.")
        return

    # Capture frame-by-frame
    ret, frame = cap.read()

    # Display the captured frame

    # Save the captured image
    cv2.imwrite('captured_image.jpg', frame)

    # Release the capture
    cap.release()
```

```
import pickle

# Define the path where the model is saved
model_path = 'vision_model.pkl'

# Load the model from the file
with open(model_path, 'rb') as f:
    vision_model = pickle.load(f)

# Now you can use vision_model as your loaded model
print("Model loaded successfully from", model_path)
```

```
Model loaded successfully from vision_model.pkl
```

```
while(1):
 audio_text = record_audio()
 if('add new face'==audio_text):
    text_to_speech("type the name of the person")
    name=input("person_name: ")
    text_to_speech("taking pictures of "+name)
    add_new_face(name)
 if(audio_text=='stop'):
    text_to_speech("thank you")
    break
 if(audio_text=="Sorry, I couldn't understand what you said."):
    print("Sorry, I couldn't understand what you said.")
    continue
 capture_image()
 name=recognize_face()
 img = PIL.Image.open(r"captured_image.jpg")

 if(name=='No face detected.'):
  response= vision_model.generate_content([audio_text,img],stream=True)
 else:
  response= vision_model.generate_content([audio_text+".name of the person in the image is "+name,img],stream=True)
 response.resolve()

 print("Assistant:"+response.text)
 text = response.text
 text_to_speech(text)
 to_markdown(response.text)
```

```
Listening...
Recognizing...
You said: add new face
Taking picture 1 of rohit
```


Figure 5.1: Image Captioning

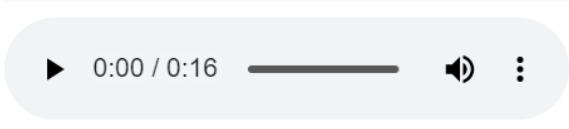



Figure 5.2: Text to voice


Figure 5.3: Signboard

```
Listening…

Recognizing…

You said: which way is Chavara Hall

Assistant: The sign for Chavara Hall is pointing to the
left.
```

Figure 5.4: Voice to Text and Sign Recognition


```
Listening...

Recognizing...

You said: What is in front of me

Assistant: The image shows the building of the Rajagiri School of Engineering and
Technology. The building has a large glass  facade and a red sign above the
entrance that says “RSET KURIAKOSE ELIAS BLOCK”. There are trees and shrubs in front
of the building, and a paved path leading up to the entrance.
```

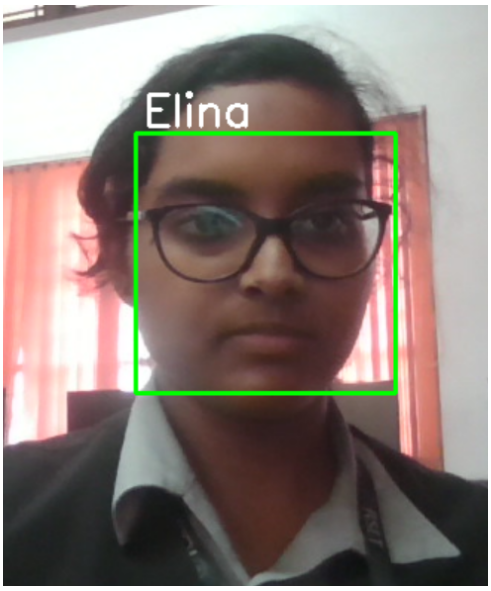



Figure 5.5: Face Recognition

# Chapter 6

# Conclusion

## 6.1 Conclusion

In conclusion, the project aims to create a comprehensive system for Human Activity Recognition (HAR) and multimodal interaction, incorporating models such as the Vision Transformer with a Reccurent Transformer (ViT-RET) for human action detection, , text-to-speech (TTS), and speech-to-text (STT). The project addressed several critical aspects of human-centric AI applications, and through the course of its development and implementation, various insights and achievements were realized.

### 6.1.1 Contributions

Our contributions include the successful integration of ViT-RET for robust human action detection, providing an effective solution for real-world applications where understanding complex human behaviors is crucial. The integration of text-to-speech (TTS) and speech-to-text (STT) components added a new dimension to the project, enabling seamless communication with users through both spoken and written language. This extended the usability of our system, making it accessible to a wider audience and facilitating interaction for individuals with diverse preferences and abilities.

### 6.1.2 Future Directions

Looking ahead, there are exciting opportunities for further enhancement and expansion of our system. Future work could focus on:

- Improving model interpretability to provide clearer insights into the decision-making processes of ViT-RET.

- Exploring advanced techniques for model optimization, scalability, and real-time processing to handle increased user loads.
- Conducting more extensive user studies to gather feedback and iteratively improve the user interface and overall user experience.
- Investigating additional modalities, such as incorporating other sensor data or exploring new AI models, to enhance the system's capabilities.

In conclusion, our project has laid a solid foundation for a versatile and user-friendly Human Activity Recognition system with multimodal interaction capabilities. The lessons learned and insights gained will undoubtedly contribute to the ongoing evolution of this project and inspire future endeavors in the field of AI-driven human-computer interaction.

# References


[1] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[2] J. Wensel, H. Ullah, and A. Munir, "Vit-ret: Vision and recurrent transformer neural networks for human activity recognition in videos," *IEEE Access*, vol. 11, pp. 72 227–72 249, 2023.

[3] J. Shi, Y. Zhang, W. Wang, B. Xing, D. Hu, and L. Chen, "A novel two-stream transformer-based framework for multi-modality human action recognition," *Applied Sciences*, vol. 13, no. 4, p. 2058, 2023.

[4] L. Li, X. Gao, J. Deng, Y. Tu, Z.-J. Zha, and Q. Huang, "Long short-term relation transformer with global gating for video captioning," *IEEE Transactions on Image Processing*, vol. 31, pp. 2726–2738, 2022.

[5] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[7] C. Shiranthika, N. Premakumara, H.-L. Chiu, H. Samani, C. Shyalika, and C.-Y. Yang, "Human activity recognition using cnn & lstm," pp. 1–6, 2020.

[8] S. Mohsen, A. Elkaseer, and S. G. Scholz, "Human activity recognition using k-nearest neighbor machine learning algorithm," in *Proceedings of the International Conference on Sustainable Design and Manufacturing*. Springer, 2021, pp. 304–313.

[9] X. Cheng, L. Zhang, Y. Tang, Y. Liu, H. Wu, and J. He, "Real-time human activity recognition using conditionally parametrized convolutions on mobile and wearable devices," *IEEE Sensors Journal*, vol. 22, no. 6, pp. 5889–5901, 2022.

[10] M. Ullah, H. Ullah, S. D. Khan, and F. A. Cheikh, "Stacked lstm network for human activity recognition using smartphone data," in *2019 8th European workshop on visual information processing (EUVIP)*. IEEE, 2019, pp. 175–180.

# Appendix A: Presentation

# Enhancing Accessibility For The Visually Impaired

## Human Action Recognition and Speech Conversion System

Group 14

April 30, 2024

Guide: Ms. Meenu Mathew
By: Elina Joy(U2003076,3)
Hemi Rose Sajeev(U2003093,20)
John Sherry(U2003107,35)
M Aman Pradeep(u2003216,52)





## Problem Definition

- People with visual impairments face daily challenges in understanding their surroundings and actions occurring in their environment.
- Existing assistive technologies have limitations in providing real-time feedback about human actions.

## Objective

- The objective of this project is to recognize and classify different human actions and to empower visually impaired individuals to interpret and navigate their surroundings independently.





## Purpose and Need for HAR

- **Accessibility:** HAR can improve accessibility for people with disabilities.
- **Navigation:** HAR can help individuals with visual impairments navigate and interact with their environment safely. It can detect obstacles, identify pathways, and provide guidance.
- **Independence and Autonomy:** HAR promotes independence and autonomy among the blind population. It allows them to perform daily tasks without relying heavily on assistance from others.

## Novelty of Idea and Scope of Implementation

- Our project is unique in its combination of real-time human action recognition using vision transformer, text conversion via recurrent transformer, and text-to-speech functionality tailored for enhancing accessibility for the visually impaired.
- The project involves developing a Human Action Recognition module, implementing a Text-to-Speech Conversion system, and integrating the technology into the daily lives of visually impaired individuals, offering unprecedented awareness and understanding of their surroundings.





## Literature Survey - I

- ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos TNN(Vit-Ret Framework)
  - Method used:TNN(ViT-ReT)
  - Advantage:
    - Decreased complexity
    - Increased speed
    - Improved accuracy
    - More efficient and lightweight approach
  - Disadvantage
    - High computational cost.
    - Loss of spatial information.
- Novel Two-Stream Transformer-Based Framework for Multi-Modality Human Action Recognition
  - Method used:RGBSFormer
  - Advantage
    - Can achieve best fusion result.
    - More effective.

## Literature Survey - II

- Novel Two-Stream Transformer-Based Framework for Multi-Modality Human Action Recognition
  - Disadvantage
    - Requirement for a large number of training samples and a significant amount of computational overhead
- Human Activity Recognition Using CNN and LSTM
  - Method used:CNN and LSTM
  - Advantage :
    - Can capture spatial dependencies.
    - Can remember very long term dependencies.
  - Disadvantage :
    - Overfitting
    - Longer computational time.





## Literature Survey - III

- Long Short-Term Relation Transformer With GlobalGating for Video Captioning
  - Method used : LSTG and G3RM
  - Advantage :
    - Reduces over-smoothing
    - Less complex
    - Removes redundancy
  - Disadvantage :
    - Relies on most-widely used features only
    - Doesn't use linguistic prior knowledge
- Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions
  - Method used:RNN , LSTM
  - Advantage :
    - Robust to misspells
    - Efficient training
    - Reduced computational requirements

# Methodology

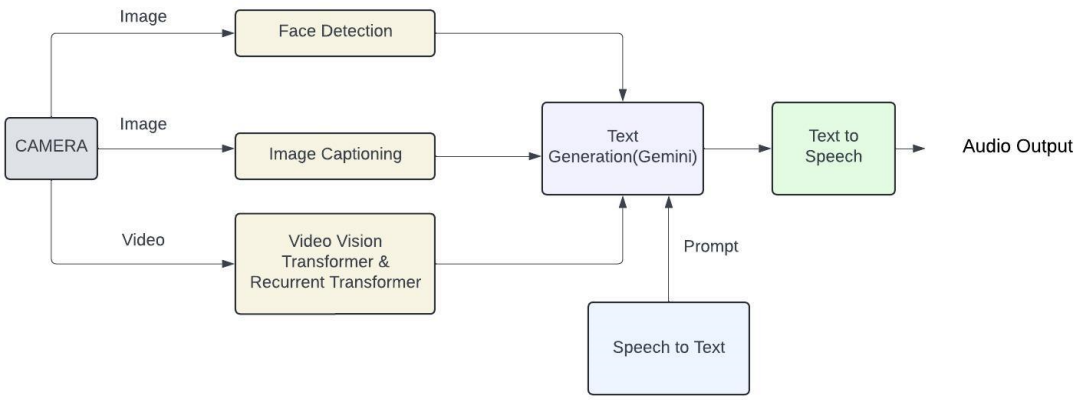
To help the visually impair navigate the environment we have introduced different modules .

1. Speech-to-text module
2. Face recognition module
3. Image captioning module
4. Human action recognition module
5. Text-to-speech module

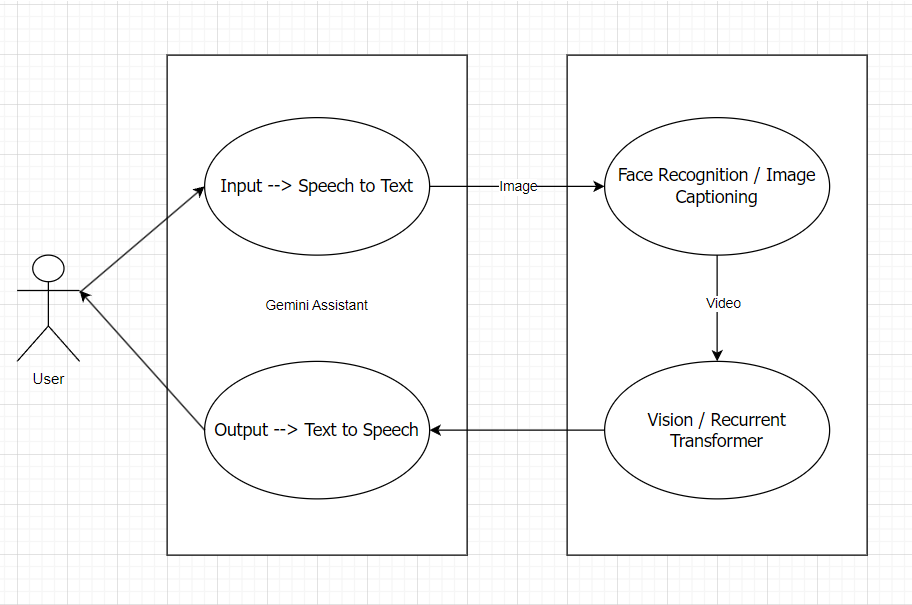




# Architecture Diagram



# UML Diagram








# Project Timeline

| TASKS | Month 1 | Month 2 | Month 4 | Month 6 | Month 8 |
|---|---|---|---|---|---|
| **Step 1:** Data Preprocessing | Completed dataset processing | | | | |
| **Step 2:** Vision Transformer | | Started model tuning and trainng | | | Still working on it as accuracy is not good |
| **Step 3:** Image Description | | Started to work on image descrpition model | Completed image captioning(60% output) | | |
| **Step 4:** Facial Recognition | | | started working on face recpgnition | Trained with group members facial data | |
| **Step 5:** Inference Model / Speech Synthesis | | | Started working on speech model | Completed speech model | |
| **Step 6:** Integration | | | Integration of face and image description model | Integrated speech/inference model | Final evaluation with video vision pending |

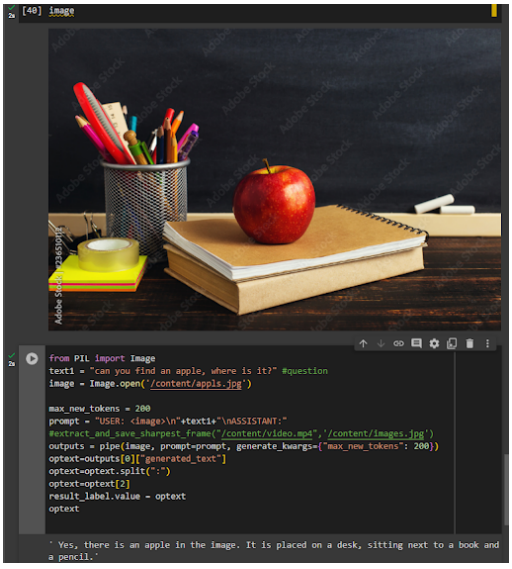
# 30% output and screenshot

```
/usr/local/lib/python3.10/dist-packages/torch/nn/modules/transformer.py:282: UserWarning: en
  warnings.warn(f"enable_nested_tensor is True, but self.use_nested_tensor is False because
<ipython-input-10-491c2dc0aa11>:43: UserWarning: Creating a tensor from a list of numpy.ndar
  frames_tensor = torch.tensor(frames, dtype=torch.float32).permute(0, 3, 1, 2) / 255.0
Features has been extracted.
```

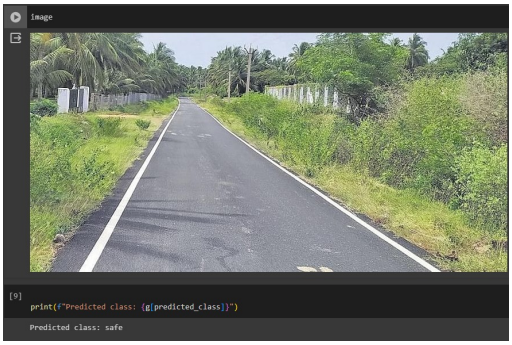




# 30% output and screenshot



## 30% output and screenshot








## Work Progress(60 Percent)

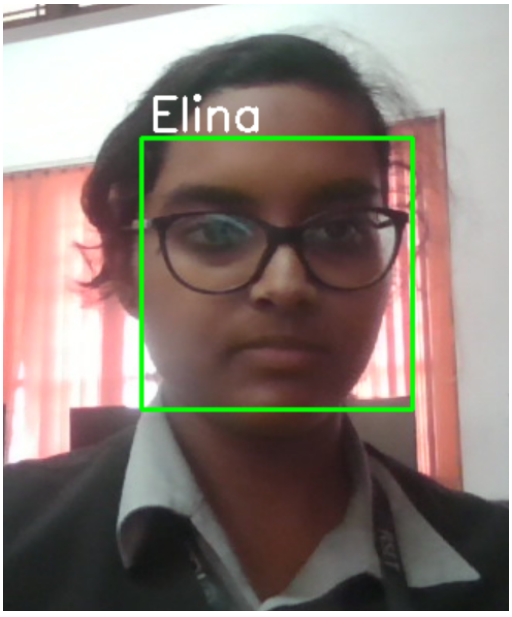



Figure: Face Recognition

# Work progress(60 Percent)




Figure: Image description








# Interim Results

- Real-time Face Recognition
- Scenic description of image
- Audio to Text
- Text to Audio

## Work Progress


```
Listening…

Recognizing…

You said: which way is Chavara Hall

Assistant: The sign for Chavara Hall is pointing to the
left.
```

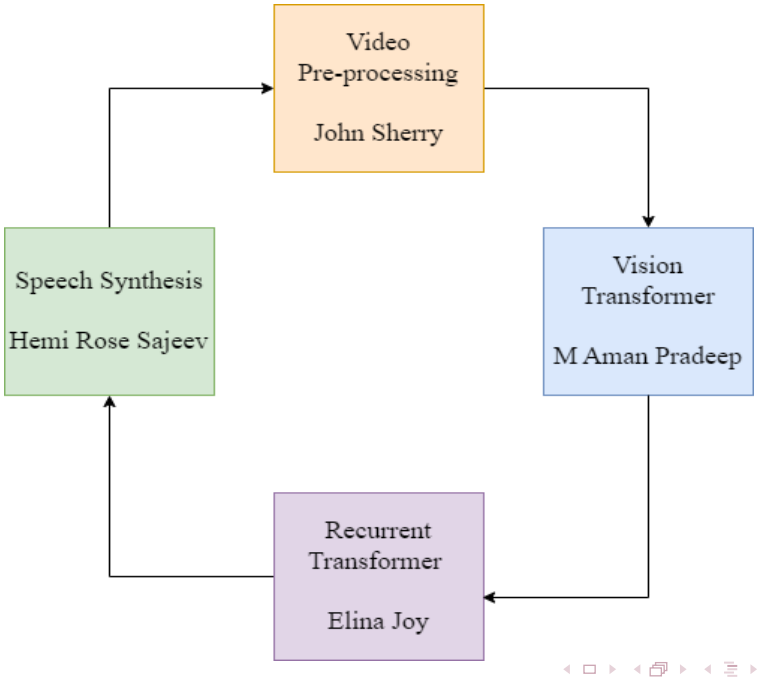




## 100 Percent Results

- Speech-to-text conversion
- Face Detection and recognition
- Image Caption generation
- Video Scene Description
- Text-to-Speech conversion

## Task Distribution








## Conclusion

- This project represents a significant step toward improving the lives of visually impaired individuals.
- By leveraging technology to detect and convey human actions through speech, we can empower the blind to navigate their world with greater confidence and independence.

## Future Scope

- Extend support for multiple languages and implement personalized voice synthesis capabilities to cater to diverse user preferences and linguistic backgrounds.
- Implement mechanisms for soliciting user feedback and preferences to adapt the system's behavior and output according to individual requirements, enhancing user experience and satisfaction.





## References

1 Shi, J.; Zhang, Y.; Wang, W.; Xing, B.; Hu, D.; Chen, L. A Novel Two-Stream Transformer-Based Framework for Multi-Modality Human Action Recognition. Appl. Sci. 2023, 13, 2058.

2 L. Li, X. Gao, J. Deng, Y. Tu, Z. -J. Zha and Q. Huang, "Long Short-Term Relation Transformer With Global Gating for Video Captioning," in IEEE Transactions on Image Processing, vol. 31, pp. 2726-2738, 2022, doi: 10.1109/TIP.2022.3158546.

3 J. Wensel, H. Ullah and A. Munir, "ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos," in IEEE Access, vol. 11, pp. 72227-72249, 2023, doi: 10.1109/ACCESS.2023.3293813.

# References

4 C. Shiranthika, N. Premakumara, H. -L. Chiu, H. Samani, C. Shyalika and C. -Y. Yang, "Human Activity Recognition Using CNN  LSTM," 2020 5th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2020, pp. 1-6, doi: 10.1109/ICITR51448.2020.9310792.

5 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenbor, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image worth 16 $\times$ 16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent., 2021, pp. 1–22.

# Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

# Vision, Mission, Programme Outcomes and Course Outcomes

**Institute Vision**

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

**Institute Mission**

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

**Department Vision**

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

**Department Mission**

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

**Programme Outcomes (PO)**

Engineering Graduates will be able to:

**1. Engineering Knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

**2. Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

**3. Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

**4. Conduct investigations of complex problems**: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

**5. Modern Tool Usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

**6. The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

**7. Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

**8. Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

**9. Individual and Team work**: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

**10. Communication**: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

**11. Project management and finance**: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

**12. Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

**Programme Specific Outcomes (PSO)**

A graduate of the Computer Science and Engineering Program will demonstrate:

**PSO1: Computer Science Specific Skills**
The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

**PSO2: Programming and Software Development Skills**
The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

**PSO3: Professional Skills**
The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

**Course Outcomes (CO)**

**Course Outcome 1:** Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: Apply).

**Course Outcome 2:** Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: Apply).

**Course Outcome 3:** Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: Apply).

**Course Outcome 4:** Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms (Cognitive knowledge level: Apply).

**Course Outcome 5:** Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level: Analyze).

**Course Outcome 6:** Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: Apply).

# Appendix C: CO-PO-PSO Mapping

## CO-PO AND CO-PSO MAPPING

| | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO 10 | PO 11 | PO 12 | PSO 1 | PSO 2 | PSO 3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CO1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 3 | | |
| CO2 | 2 | 2 | 2 | | 1 | 3 | 3 | 1 | 1 | | 1 | 1 | | 2 | |
| CO3 | | | | | | | | | 3 | 2 | 2 | 1 | | | 3 |
| CO4 | | | | | 2 | | | 3 | 2 | 2 | 3 | 2 | | | 3 |
| CO5 | 2 | 3 | 3 | 1 | 2 | | | | | | | 1 | 3 | | |
| CO6 | | | | | 2 | | | 2 | 2 | 3 | 1 | 1 | | | 3 |

3/2/1: high/medium/low

## JUSTIFICATIONS FOR CO-PO MAPPING & CO-PSO MAPPING

| MAPPING | LOW/MEDIUM/ HIGH | JUSTIFICATION |
|---|---|---|
| 100003/ CS722U.1-PO1 | M | Knowledge in the area of technology for project development using various tools results in better modeling. |
| 100003/ CS722U.1-PO2 | M | Knowledge acquired in the selected area of project development can be used to identify, formulate, review |

| | | |
|---|---|---|
| | | research literature, and analyze complex engineering problems reaching substantiated conclusions. |
| 100003/ CS722U.1-PO3 | M | Can use the acquired knowledge in designing solutions to complex problems. |
| 100003/ CS722U.1-PO4 | M | Can use the acquired knowledge in designing solutions to complex problems. |
| 100003/ CS722U.1-PO5 | H | Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions. |
| 100003/ CS722U.1-PO6 | M | Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices. |
| 100003/ CS722U.1-PO7 | M | Project development based on societal and environmental context solution identification is the need for sustainable development. |
| 100003/ CS722U.1-PO8 | L | Project development should be based on professional ethics and responsibilities. |
| 100003/ CS722U.1-PO9 | L | Project development using a systematic approach based on well defined principles will result in teamwork. |

| | | |
|---|---|---|
| 100003/ CS722U.1-PO10 | M | Project brings technological changes in society. |
| 100003/ CS722U.1-PO11 | H | Acquiring knowledge for project development gathers skills in design, analysis, development and implementation of algorithms. |
| 100003/ CS722U.1-PO12 | H | Knowledge for project development contributes engineering skills in computing & information gatherings. |
| 100003/ CS722U.2-PO1 | H | Knowledge acquired for project development will also include systematic planning, developing, testing and implementation in computer science solutions in various domains. |
| 100003/ CS722U.2-PO2 | H | Project design and development using a systematic approach brings knowledge in mathematics and engineering fundamentals. |
| 100003/ CS722U.2-PO3 | H | Identifying, formulating and analyzing the project results in a systematic approach. |
| 100003/ CS722U.2-PO5 | H | Systematic approach is the tip for solving complex problems in various domains. |
| 100003/ CS722U.2-PO6 | H | Systematic approach in the technical and design aspects provide valid conclusions. |

| | | |
|---|---|---|
| 100003/ CS722U.2-PO7 | H | Systematic approach in the technical and design aspects demonstrate the knowledge of sustainable development. |
| 100003/ CS722U.2-PO8 | M | Identification and justification of technical aspects of project development demonstrates the need for sustainable development. |
| 100003/ CS722U.2-PO9 | H | Apply professional ethics and responsibilities in engineering practice of development. |
| 100003/ CS722U.2-PO11 | H | Systematic approach also includes effective reporting and documentation which gives clear instructions. |
| 100003/ CS722U.2-PO12 | M | Project development using a systematic approach based on well defined principles will result in better teamwork. |
| 100003/ CS722U.3-PO9 | H | Project development as a team brings the ability to engage in independent and lifelong learning. |
| 100003/ CS722U.3-PO10 | H | Identification, formulation and justification in technical aspects will be based on acquiring skills in design and development of algorithms. |
| 100003/ CS722U.3-PO11 | H | Identification, formulation and justification in technical aspects provides the betterment of life in various domains. |
| 100003/ CS722U.3-PO12 | H | Students are able to interpret, improve and redefine technical aspects with mathematics, science and |

| | | |
|---|---|---|
| | | engineering fundamentals for the solutions of complex problems. |
| 100003/ CS722U.4-PO5 | H | Students are able to interpret, improve and redefine technical aspects with identification formulation and analysis of complex problems. |
| 100003/ CS722U.4-PO8 | H | Students are able to interpret, improve and redefine technical aspects to meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations. |
| 100003/ CS722U.4-PO9 | H | Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions. |
| 100003/ CS722U.4-PO10 | H | Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools for better products. |
| 100003/ CS722U.4-PO11 | M | Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices. |
| 100003/ CS722U.4-PO12 | H | Students are able to interpret, improve and redefine technical aspects for demonstrating the knowledge of, and need for sustainable development. |

| | | |
|---|---|---|
| 100003/ CS722U.5-PO1 | H | Students are able to interpret, improve and redefine technical aspects, apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice. |
| 100003/ CS722U.5-PO2 | M | Students are able to interpret, improve and redefine technical aspects, communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions. |
| 100003/ CS722U.5-PO3 | H | Students are able to interpret, improve and redefine technical aspects to demonstrate knowledge and understanding of the engineering and management principle in multidisciplinary environments. |
| 100003/ CS722U.5-PO4 | H | Students are able to interpret, improve and redefine technical aspects, recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change. |
| 100003/ CS722U.5-PO5 | M | Students are able to interpret, improve and redefine technical aspects in acquiring skills to design, analyze and develop algorithms and implement those using high-level programming languages. |
| 100003/ CS722U.5-PO12 | M | Students are able to interpret, improve and redefine technical aspects and contribute their engineering skills in |

| | | |
|---|---|---|
| | | computing and information engineering domains like network design and administration, database design and knowledge engineering. |
| 100003/ CS722U.6-PO5 | M | Students are able to interpret, improve and redefine technical aspects and develop strong skills in systematic planning, developing, testing, implementing and providing IT solutions for different domains which helps in the betterment of life. |
| 100003/ CS722U.6-PO8 | H | Students will be able to associate with a team as an effective team player for the development of technical projects by applying the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems. |
| 100003/ CS722U.6-PO9 | H | Students will be able to associate with a team as an effective team player to Identify, formulate, review research literature, and analyze complex engineering problems |
| 100003/ CS722U.6-PO10 | M | Students will be able to associate with a team as an effective team player for designing solutions to complex engineering problems and design system components. |
| 100003/ CS722U.6-PO11 | M | Students will be able to associate with a team as an effective team player, use research-based knowledge and research methods including design of experiments, analysis and interpretation of data. |

| | | |
|---|---|---|
| 100003/ CS722U.6-PO12 | H | Students will be able to associate with a team as an effective team player, applying ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice. |
| 100003/ CS722U.1-PSO1 | H | Students are able to develop Computer Science Specific Skills by modeling and solving problems. |
| 100003/ CS722U.2-PSO2 | M | Developing products, processes or technologies for sustainable and socially relevant applications can promote Programming and Software Development Skills. |
| 100003/ CS722U.3-PSO3 | H | Working in a team can result in the effective development of Professional Skills. |
| 100003/ CS722U.4-PSO3 | H | Planning and scheduling can result in the effective development of Professional Skills. |
| 100003/ CS722U.5-PSO1 | H | Students are able to develop Computer Science Specific Skills by creating innovative solutions to problems. |
| 100003/ CS722U.6-PSO3 | H | Organizing and communicating technical and scientific findings can help in the effective development of Professional Skills. |