Gesture Genie

Capstone Project Proposal

Submitted by:

(102053003) KESHAV GUPTA

(102053036) RITIK ARORA

(102003013) HARSH GUPTA

(102003756) VANSH MITTAL

(102017137) SUDHIT SONI

BE Third Year- COE/ CSE

CPG No. 51

Under the Mentorship of

Dr. Suresh Raikwar

Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

MONTH & 2023

TABLE OF CONTENTS

Mentor Consent Form

I hereby agree to be the mentor of the following Capstone Project Team

Gesture Genie

Roll No	Name	Signatures
102053003	Keshav Gupta	Keshav
102053036	Ritik Arora	Ritik
102003013	Harsh Gupta	Harsh.
102003756	Vansh Mittal	mittal
102017137	Sudhit Soni	Sudhit Soni

NAME of Mentor: Dr. Suresh Raikwar

SIGNATURE of Mentor:

NAME of Co-Mentor(if any): N/A

SIGNATURE of Co-Mentor: N/A

Project Overview

The motivation for creating a new type of AI assistant with a video avatar that emulates human-like gestures, facial expressions, and emotions is to provide a more natural and intuitive way for users to interact with technology. This type of AI assistant could also improve users' ability to understand and respond to their needs hence providing additional context and cues for the user. Additionally by providing a more natural and human-like interface for technology, we can create a more seamless integration between humans and machines, making technology more accessible and user-friendly for everyone.

There are several potential problems, solutions, and issues related to the development of an AI assistant with a video avatar that emulates human-like behavior out of which major one is creating a realistic video avatar that can accurately convey emotions and gestures is a complex and challenging task that requires advanced computer vision and machine learning techniques moreover the use of a video avatar could raise privacy and ethical concerns, as users may feel uncomfortable interacting with a virtual assistant that appears to be "watching" them. Lastly, use of this model could increase the computational resources required to run the AI assistant, which could impact the overall performance and user experience.

Our team would need to leverage state-of-the-art computer vision and machine learning technologies, as well as create realistic 3D avatars to solve such problems. Moreover, we would need to implement strict privacy and security protocols to ensure that users' data is protected and that the avatar's behavior is transparent and predictable. Additionally we need to optimize the algorithms used to run the AI assistant and leverage cloud-based computing resources to provide scalable and reliable performance. The benefits it would provide are like this technology more accessible for individuals with disabilities or language barriers. For example individuals who do not speak the language of the AI assistant may benefit from a visual interface. To develop an AI assistant with a video avatar that emulates human-like gestures, facial expressions, and emotions, the project would require several different types of datasets like to accurately model the video avatar's facial expressions, the project would require a dataset of images or videos that show a range of emotions and facial expressions. The dataset should include examples of different body positions, such as standing, sitting, and walking, among others.

Problem Statement

The current AI assistants lack human-like qualities, such as natural gestures, facial expressions, and emotions, which can make the user experience less engaging and intuitive. The problem is particularly acute in use cases where the AI assistant serves as a virtual assistant, customer service representative, or salesperson, where a human-like interface could lead to better customer engagement, increased customer satisfaction, and potentially higher sales.

Therefore, the problem is to develop an AI assistant with a video avatar that can emulate human-like gestures, facial expressions, and emotions, providing a more immersive and engaging user experience. This requires solving a range of technical challenges, such as accurate facial expression and gesture recognition, audio processing for speech recognition, and the development of a video avatar that can accurately mimic human movements and expressions. The solution to this problem has the potential to improve the user experience of AI assistants, leading to more natural and intuitive interactions between humans and technology.

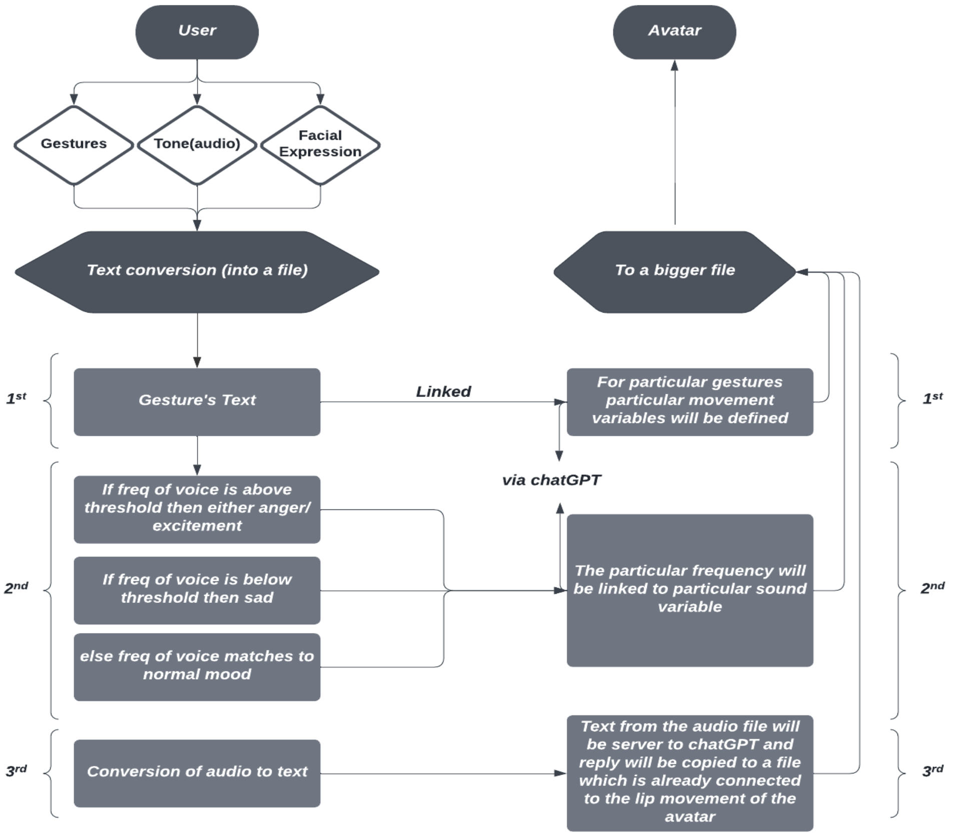
User
Avatar
Gestures
Tone(audio)
Facial Expression
Text conversion (into a file)
To a bigger file
1st
Gesture's Text
Linked
For particular gestures particular movement variables will be defined
1st
via chatGPT
2nd
If freq of voice is above threshold then either anger/ excitement
If freq of voice is below threshold then sad
else freq of voice matches to normal mood
The particular frequency will be linked to particular sound variable
2nd
3rd
Conversion of audio to text
Text from the audio file will be server to chatGPT and reply will be copied to a file which is already connected to the lip movement of the avatar
3rd

Fig1.1 WorkFlow

Need Analysis

The proposed project of a portal for face-to-face conversation with custom avatars is a unique and innovative solution to the problem of traditional communication methods lacking human touch. In today's digital age, communication is more important than ever, but it can often feel impersonal and disconnected. With the use of custom-made avatars that can detect human emotions, gestures, and voice amplitude, users will be able to have engaging and effective conversations that are more meaningful and personalized.

The portal's use of ML techniques and the ChatGPT model ensures that the responses generated are relevant and coherent. This has practical applications in many areas, such as customer service, education, healthcare, and more. In customer service, the portal can be used to provide a personalized and engaging experience for customers, enhancing customer satisfaction and loyalty. In education, the portal can be used to provide a more immersive and engaging learning experience for students. In healthcare, the portal can be used to provide a more personalized and empathetic experience for patients. For example, a paper written by V. Ribeiro on “Virtual Agent for Mental Health” proposed a virtual agent that uses speech, facial expressions, and gestures to interact with patients and detect depression levels. It also includes a machine learning algorithm to analyze the data collected and provide personalized recommendations. Paper written by M. El Ayadi on "Emotion Recognition using Speech Features" proposes a system for emotion recognition using speech features. The system uses machine learning algorithms to classify emotions and has applications in mental health and therapy. It has been determined that there has already been a great deal of work done on this side that could prove helpful for our project.

Overall, the proposed project has significant relevance in the real world, providing an innovative solution to the problem of traditional communication methods lacking human touch. With its practical applications in various fields, the portal for face-to-face conversation with custom avatars has the potential to improve communication and enhance experiences for users, making it an exciting and promising development.

Literature Survey

Existing avatar-based conversation systems

Mitsuku is a popular chatbot designed by Steve Worswick. The system uses natural language processing and machine learning algorithms to provide personalized conversations with users. Mitsuku has won several awards, including the Loebner Prize, which is an annual competition for chatbots. The system is widely used in various domains, such as customer service, education, and entertainment.

Replika is an AI-powered personal chatbot designed to have natural conversations with users. The system uses machine learning algorithms to learn from the user's responses and create a personalized experience. Replika can be used for various purposes, such as mental health support, self-improvement, and daily conversations.

Woebot is a mental health chatbot designed to help individuals manage their mental health. The system uses cognitive-behavioral therapy (CBT) techniques and machine learning algorithms to provide personalized support to users. Woebot has been shown to be effective in reducing symptoms of depression and anxiety.

Talla is an AI-powered chatbot designed for customer service. The system uses natural language processing and machine learning algorithms to provide personalized support to customers. Talla can be used for various purposes, such as answering customer queries, providing product recommendations, and handling customer complaints.

Hugging Face is an AI-powered chatbot designed for natural language conversations. The system uses machine learning algorithms to understand the context of the conversation and provide relevant responses. Hugging Face is widely used in various domains, such as customer service, education, and entertainment.

Conversation systems based on avatars and emotion detection

One research paper on this topic is "Affective Computing with Emotion Recognition in Avatar-based Systems," by J. Althoff and J. Sidner. This paper explores the use of avatar-based conversation systems for affective computing and emotion recognition. The authors discuss the importance of recognizing emotions in user input to improve the effectiveness of conversation systems, and present a framework for emotion detection using multiple sources of information, such as text, voice, and facial expressions.

Another research paper on this topic is "Emotion Recognition in Text-Based Conversations with Virtual Agents," by C. van der Lee and E. M. A. G. van Dijk. This paper focuses on emotion recognition in text-based conversations with virtual agents, which are a type of avatar-based conversation system. The authors present a machine learning approach for detecting emotions in user input based on text-based features, such as sentiment analysis and lexical choice.

A third research paper on this topic is "Emotion Detection in Avatar-based Human-Computer Interaction: A Review," by S. Y. Park and H. J. Lee. This paper provides a comprehensive review of emotion detection in avatar-based human-computer interaction. The authors discuss the various methods used for emotion detection, including natural language processing, facial expression recognition, and physiological signals, and highlight the potential applications and challenges of this field.

Conversation systems using NLP techniques:

A research paper by Bing Liu and Ian Lane on "Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling" proposes a recurrent neural network (RNN) model with attention mechanism for joint intent detection and slot filling in conversation systems. The model uses an attention-based mechanism to attend to the relevant parts of the input sequence for intent detection and slot filling. The model is trained using a joint objective function that maximizes the probability of the correct intent and slot values. The model outperforms existing models on benchmark datasets for intent detection and slot filling.

Another research paper by Dongfang Xu, Chang Liu, Hua Xu, Ramakanth Kavuluru on "Contextual Intent Tracking in Intelligent Conversational Agents" proposes a contextual intent tracking (CIT) framework for intelligent conversational agents. CIT is a task that aims to identify the user's intent based on the current conversation context. The proposed framework uses a two-stage approach: the first stage extracts features from the input sequence using a bidirectional long short-term memory (BiLSTM) network, and the second stage uses a support vector machine (SVM) to classify the intent based on the extracted features. The framework also includes a confidence score to measure the certainty of the predicted intent. The framework is evaluated on benchmark datasets for intent tracking and outperforms existing models.

Text Summarization:

The paper proposed by Alexander M. Rush, Sumit Chopra, and Jason Weston on "A Neural Attention Model for Abstractive Sentence Summarization" talked about a new neural network model for abstractive sentence summarization that incorporates an attention mechanism. The model generates summaries by attending to important parts of the input sequence and using them to generate the output sequence. The authors demonstrate that their model outperforms previous state-of-the-art models on the Gigaword and DUC-2004 datasets.

A paper by Abigail See, Peter J. Liu, and Christopher D. Manning on "Get To The Point: Summarization with Pointer-Generator Networks" proposed a new neural network model for abstract text summarization that incorporates both a pointer mechanism and a generator mechanism. The pointer mechanism enables the model to copy words directly from the input text, while the generator mechanism allows the model to generate new words that do not appear in the input text. The authors demonstrate that their model outperforms previous state-of-the-art models on the CNN/Daily Mail dataset.

Extraction-based Summarization: One of the earliest approaches to text summarization is extraction-based summarization, which involves selecting the most informative sentences or phrases from a document to create a summary. Some key works in this area include "Automatic Summarization" by Edmundson (1969), "The Automatic Creation of Literature Abstracts" by Luhn (1958), and "TextRank: Bringing Order into Texts" by Mihalcea and Tarau (2004).

A paper "Get To The Point: Summarization with Pointer-Generator Networks" by See, Liu, and Manning (2017) talks about abstract summarization involving generating a summary by paraphrasing and synthesizing the content of a document. Abstractive summarization is a more challenging task than extraction-based summarization, as it requires a deep understanding of the text and the ability to generate new sentences that capture the essence of the document.

Objectives

- Study the research papers on github, kaggle and google scholar to learn which dataset and algorithm should be used depending on the accuracy, speed and linearity comparisons.

- The avatar should be capable of mimicking the same gestures and responding to emotions like sadness, anger, excitement and facial expressions as a human and to integrate ML and natural language processing (NLP) algorithms to enable the avatar to accurately interpret and respond to user requests and commands in real time.

- Implement cv algorithms to enable the avatar to analyze and interpret visual information, such as facial expressions and body language, to create a more immersive and engaging user experience. Train the avatar on large datasets of human gestures, facial expressions, and emotions to improve its accuracy and naturalness.

- Use ChatGPT to generate responses to user requests and commands, and convert them to audio to be spoken by the avatar and convert user speech to text and use NLP algorithms to summarize the text to enable faster and more efficient communication and to implement learning algorithms to allow the avatar to learn and improve over time based on feedback from user interactions.

- Testing the smooth and accurate functioning of the AI avatar along with various components like avatar testing, ML testing and more.

Methodology

- The first step is to collect a large dataset of text-based conversations to train the ML model. The dataset should be diverse and cover a wide range of topics. We'll preprocess the data to remove noise, clean the text, and convert it into a suitable format for the ML model. We need to select the appropriate algorithm based on the task at hand. For example, we can use algorithms such as Named Entity Recognition (NER), Part-of-Speech (POS) tagging, or Sentiment Analysis [4] to extract relevant information from the user's input.

- We'll use algorithms such as decision trees, support vector machines (SVMs), or deep neural networks (DNNs) to train the avatar [4]. Once we have selected the appropriate algo, we need to extract features from the preprocessed data. We can use various feature extraction techniques such as Principal Component Analysis (PCA), Local Binary Patterns (LBP), or Histogram of Oriented Gradients (HOG). We'll implement video and audio streaming capabilities to allow the user to have a real-time conversation with the avatar. This will require integrating video and audio codecs, as well as network protocols for efficient and reliable streaming

- Avatar Design: The first step is to design the avatar. This can be done using various 3D modeling software such as Blender, Maya, or 3DS Max [3]. We need to locate specific points on the face, such as the corners of the mouth, eyes, and eyebrows. We'll use machine learning algorithms such as Convolutional Neural Networks (CNNs) to detect and recognize emotions based on facial expressions [5]. This will enable the avatar to display the appropriate emotions in response to the user's input. We can use machine learning algorithms such as Hidden Markov Models (HMMs) [1] to recognize and track hand gestures made by the user. This will enable the avatar to respond appropriately based on the user's gestures..

- We'll integrate the ChatGPT API into our project, which will allow us to access its pre-trained language models to generate responses to user requests and commands. After generating a response using ChatGPT, we'll convert the text to speech using a text-to-speech (TTS) engine. There are several TTS engines available, such as Google Text-to-Speech or Amazon Polly, that can be integrated into our project. Once we have the audio file, we'll integrate it into the avatar in our project. The avatar will be designed to lip-sync with the audio and use natural gestures and facial expressions to make the conversation more engaging.

Work Plan

- You need to give a short work plan which you will set for achieving the set objectives, in this section.

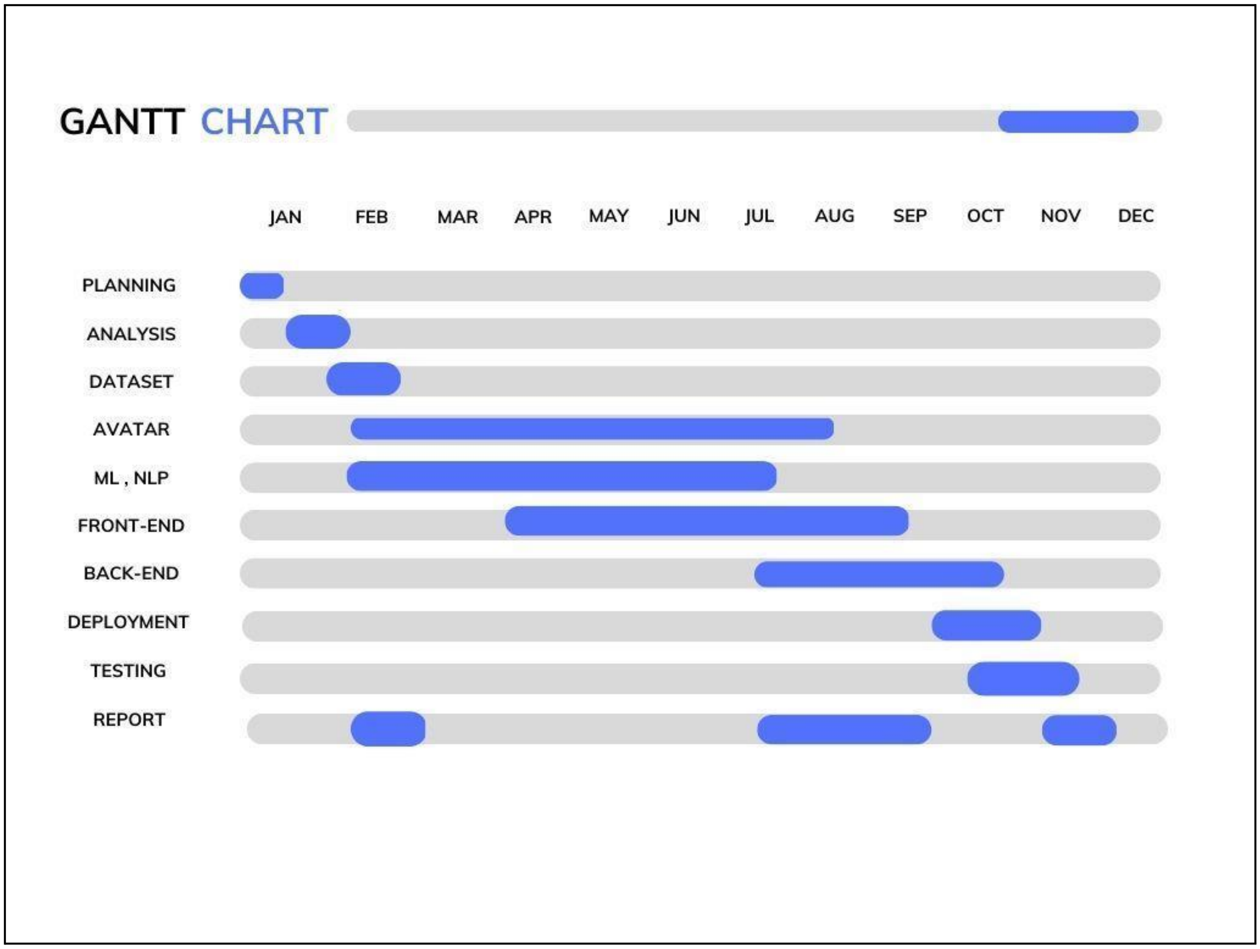

Fig1.2 Work Plan

Project Outcomes & Individual Roles

1. Video Recognition: One of the primary objectives of this project is to improve the accuracy and reliability of voice recognition along with the help of video. This can lead to more accurate responses and an improved user experience.
2. Personalization and Recommendations: This project has been trained on lots of data to provide tailored recommendations based on preferences. This will lead to increased customer satisfaction and loyalty.
3. Enhanced User Experience: This project consists of numerous Avatars that can provide users with a more natural and intuitive way to interact with them leading to increased user satisfaction and engagement.
4. Increased availability: Since this project is linked to ChatGPT so it can be accessed by users 24/7, making it a valuable resource for individuals seeking information or assistance at any time of day or night.
5. Expansion of Knowledge Base: This project is linked with ChatGPT which updates continuously and automatically with new information and knowledge, hence expanding its capabilities and usefulness to users

6. Overall, this project aims to provide users with a valuable resource for information, assistance and communication, leveraging the power of AI and natural language processing to enhance the user experience.

Roles and Responsibilities

S.No.	Name	Role	Objectives
1	Sudhit Soni	1. UI/UX of web app 2. Video processing	1. Build of a web portal providing a face-to-face convo with avatars. 2. Ensure a smooth flow of cross communication without any lag.
2	Keshav Gupta	1. Implementation of various ML and DL algorithms 2. Increase the efficiency using different ml models.	1. Implement ML and DL models to recognize gestures, facial expressions, and emotions. 2. To improve accuracy, train the avatar with ml algos on large datasets..
3	Vansh Mittal	1. Formation of avatars 2. Incorporation of its movements and animation	1. Building of avatars that'll respond to every user's request and gesture, 2. Integration of animations to make it more accessible and user-friendly.
4	Ritik Arora	1. Backend of Web app 2.Testing	1. Linking of website with backend. 2. Testing of avatars response with backend
5	Harsh Gupta	1.API Integration 2. Backend of Web app	1. Chat Gpt integration via OpenAPI. 2. Text to Speech conversion.

Table 1.1

Course Subjects

- The Various Core subjects that will be used includes -

1. Machine Learning : Machine learning provides us with a workflow describing how various features are implemented chronologically, such as first data preprocessing, data reduction, analysis, model selection and so on. The code snippets in the slides provided a basic understanding of how to work on various libraries such as pandas, numpy, and matplotlib etc. which was very helpful during implementation of our project.

2. Cloud Computing : Cloud Computing provides us the skills and knowledge to make real life applications and scale the systems for handling heavy traffic and real time requests. Cloud Computing provides an IAAS (infrastructure as a service) for developers to deploy websites and hence we have used AWS as a cloud service provider.

3. Full Stack Development : Full Stack Development provides us to Develop various aspects of a website or a software including frontend and backend. Frontend enables us to style our website and provide a good user experience and an interactive UI/UX for the users. Backend provides us functionalities to work with various REST API'S (Application Programming Interface) like ChatGpt OpenAI API'S to connect our software to Chatgpt.

4. Database Management System (DBMS) : Database Management System allows us to manage and query our information in our database effectively and efficiently and NOSQL Database helps us to scale our systems horizontally and provide data sharding for managing heavy read and writes to and from a database and hence MongoDB and AmazonS3 as databases.

5. Natural Language Processing : The AI assistant must be able to accurately interpret and respond to user requests and commands, which requires sophisticated natural language processing (NLP) algorithms. These algorithms enable the AI assistant to understand and respond to spoken or written language, including slang, colloquialisms, and variations in grammar and syntax.

REFERENCES

[1] Zhiwen Yu, Shiguang Shan (2007) ‘Emotion Recognition from Facial Expressions Using Multilevel HMM’, Recognizing emotions from facial expressions using multilevel Hidden Markov Models (HMM). Available: https://ieeexplore.ieee.org/document/4161366.

[2] Yasser Mahgoub, Mohamed Ally, Saul Greenberg (2008),‘Designing Avatars for Face-to-Face Communication in Virtual Worlds’.Available : https://dl.acm.org/doi/10.1145/1358628.1358817

[3] Sakshi Sharma, Akanksha Tiwari (2009),"Chatbot Design and Implementation Using Dialog flow" Available : https://link.springer.com/chapter/10.1007/978-981-15-3312-2_9

[4] Li (2018),’Speech Synthesis Based on Deep Learning and Its Applications’ Reviews speech synthesis techniques based on deep learning.Available: https://www.mdpi.com/1424-8220/18/5/1468

[5] Li et al (2018) , ‘Deep Learning-Based Facial Expression Recognition: A Comprehensive Review’. :Reviews Deep learning-based facial expression recognition methods. Available: https://www.mdpi.com/1424-8220/18/12/4160