A
Mini-Project Report on

Enhancing Human–Machine Interaction through Multimodal Emotion Recognition

Submitted in partial fulfillment of the requirements
for the degree of
BACHELOR OF ENGINEERING
IN
Computer Science & Engineering
Artificial Intelligence & Machine Learning

By

Jeet Manjrekar (21106061)
Devesh Sali (21106016)
Sakshi Rajeshirke (22206002)
Prathamesh Mane (22206003)

Under the guidance of

Prof. Vijesh Nair

Department of Computer Science & Engineering
(Artificial Intelligence & Machine Learning)
A. P. Shah Institute of Technology
G. B. Road, Kasarvadavali, Thane (W)-400615
University Of Mumbai
2023-2024

A. P. SHAH INSTITUTE OF TECHNOLOGY

CERTIFICATE

This is to certify that the project entitled “**Enhancing Human–Machine Interaction through Multimodal Emotion Recognition”** is a bonafide work of Jeet Manjrekar (21106061), Devesh Sali (21106016), Sakshi Rajeshirke (22206002), Prathamesh Mane (22206003) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

Prof. Vijesh Nair
Mini Project Guide

Dr. Jaya Gupta
Head of Department

A. P. SHAH INSTITUTE OF TECHNOLOGY

Project Report Approval

This Mini project report entitled "**Enhancing Human–Machine Interaction through Multimodal Emotion Recognition***"* by Jeet Manjrekar (21106061), Devesh Sali (21106016), Sakshi Rajeshirke (22206002), Prathamesh Mane (22206003) is approved for the degree of ***Bachelor of Engineering*** in ***Computer Science &Engineering***, (AIML) ***2023-24***.

External Examiner: ___________________

Internal Examiner: __________________

Place: APSIT, Thane
Date:

Declaration

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission hasnot been taken when needed.

Jeet Manjrekar	Devesh Sali	Sakshi Rajeshirke	Prathamesh Mane
(21106061)	(21106016)	(22206002)	(22206003)

ABSTRACT

Enhanced human-machine interaction (HMI) is crucial for the advancement of various fields, including virtual assistants, robotics, and healthcare. Emotion recognition plays a pivotal role in making these interactions more intuitive and effective. This paper explores the potential of multimodal emotion recognition systems, which integrate data from multiple sources such as facial expressions, speech patterns, and physiological signals, to provide a more comprehensive understanding of human emotions. We review recent advancements in multimodal emotion recognition techniques, including machine learning algorithms and deep learning models, and discuss their applications in real-world scenarios. Furthermore, we analyze the challenges and opportunities associated with implementing multimodal emotion recognition systems, such as data privacy concerns, cross-cultural differences, and computational complexity. By leveraging multimodal emotion recognition, we can enhance the capabilities of HMI systems, enabling more natural and adaptive interactions between humans and machines, ultimately leading to improved user experiences and performance across various domains.

Index

Index			Page no.
Chapter-1			
	Introduction		1-2
Chapter-2			
	Literature Survey		3-7
	2.1	History	
	2.1	Review	
Chapter-3			
	Problem Statement		8-10
Chapter-4			
	Technology Stack		11-12
Chapter-5			
	Proposed system and Implementation		13-17
	5.1	Block Diagram of proposed system	
	5.2	Description of Block diagram	
	5.3	Implementation	
Chapter-6			
	Conclusion		18-20
References			21

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

In today's rapidly evolving technological landscape, the interaction between humans and machines has become increasingly prevalent and essential across various domains, including virtual assistants, gaming, healthcare, and customer service. While significant progress has been made in developing intelligent systems capable of understanding and responding to human inputs, there remains a critical gap in effectively interpreting and responding to human emotions. Emotions play a fundamental role in human communication, influencing decision-making, behavior, and overall well-being. Thus, incorporating emotion recognition capabilities into human-machine interaction (HMI) systems holds great promise for improving user experiences and performance.

Traditional approaches to emotion recognition have primarily focused on analyzing single modalities, such as facial expressions or speech patterns. However, human emotions are complex and multifaceted, often expressed through a combination of verbal and non-verbal cues. To address this complexity, researchers have increasingly turned to multimodal emotion recognition systems, which integrate data from multiple sources, including facial expressions, speech, physiological signals, and contextual information. By combining information from diverse modalities, multimodal emotion recognition systems offer a more comprehensive understanding of human emotions, enabling more natural and intuitive interactions between humans and machines.

This report aims to explore the potential of multimodal emotion recognition in enhancing HMI systems. We will review recent advancements in multimodal emotion recognition techniques, including machine learning algorithms and deep learning models, and discuss their applications in real-world scenarios. Additionally, we will examine the challenges and opportunities associated with implementing multimodal emotion recognition systems, such as data privacy concerns, cross-cultural differences, and computational complexity. Through a thorough analysis of these factors, we will demonstrate the potential impact of multimodal emotion recognition on improving the capabilities of HMI systems, ultimately leading to more effective and personalized interactions between humans and machines.

.

CHAPTER 2

LITERATURE SURVEY

2. LITERATURE SURVEY

2.1 HISTORY

- Early Research (1970s-1990s): The exploration of emotions in human-computer interaction (HCI) began with early research efforts focused on understanding how users interact with computers. Emotions were initially considered secondary to task completion but gradually gained recognition as important factors influencing user experience.

- Introduction of Emotion Recognition (2000s): In the early 2000s, researchers started experimenting with methods to recognize emotions using facial expressions, speech patterns, and physiological signals. Early systems primarily focused on single-modal approaches, such as facial expression analysis or voice recognition.

- Advancements in Machine Learning (2010s): The proliferation of machine learning techniques, particularly deep learning, led to significant advancements in emotion recognition systems. Researchers began exploring multimodal approaches that combine data from multiple sources, such as facial expressions, voice tone, and body language, to improve accuracy and robustness.

- Commercial Applications (2010s-Present): As technology evolved, commercial applications of multimodal emotion recognition emerged in various industries, including healthcare, entertainment, and customer service. Companies started integrating emotion recognition capabilities into virtual assistants, gaming systems, and educational platforms to enhance user engagement and satisfaction.

- Integration with AI and Robotics (Present): In recent years, there has been a growing emphasis on integrating emotion recognition with artificial intelligence (AI) and robotics. This integration enables machines to not only detect human emotions but also respond empathetically, leading to more natural and effective human-machine interactions.

- Ethical and Privacy Considerations (Present): With the increasing deployment of emotion recognition technology in various domains, concerns regarding privacy, bias, and ethical implications have come to the forefront. Researchers and policymakers are actively addressing these concerns to ensure responsible development and deployment of emotion recognition systems.

- Future Directions (Ongoing): Looking ahead, researchers are exploring advanced techniques, such as affective computing and affective artificial intelligence, to further improve emotion recognition accuracy and sophistication. Additionally, efforts are underway to develop standards and guidelines for ethical and transparent use of emotion recognition technology in human-machine interaction.

2.2-LITERATURE REVIEW

1. **Multimodal Emotion Recognition using Deep Convolution and Recurrent Network : IEEE (2021)**

   Automatic human emotion recognition is one of the most important and growing field of research in Human Computer Interaction (HCI) domain. It has huge impact on applications like Automatic Human Behaviour Analysis and multimedia retrieval system. Recently, Deep Neural Networks have gain a lot of success in terms of accuracy related to machine learning tasks. Automatic human emotion recognition have also been solved using the deep learning based Convolution Neural Network (CNN). Currently, most of deep learning based algorithms for human emotion recognition only focuses on the particular direction like Vision, Text and Audio. These algorithms are trained on specific modality (visual data, textual data, acoustic data) and performs well in a controlled environment, but failed to achieve the good results in most of the real-life cases. It is due to the unpredictable behaviour of the human nature. So to tackle this problem, a novel deep learning based multi modal architecture have been proposed in this paper. This algorithms utilizes visual, textual as well as audio features to enhance the accuracy on automatic human emotion recognition. Extensive experiments have been performed to improve the accuracy and transparency of our proposed work. Results have proved that we achieved good results as compared to the sate-of-the-art results. [1]

2. **Facial emotion recognition system through machine learning approach : IEEE (2017)**

   Data mining also sometimes called data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information. Image processing is related to Computer vision, which is a high-level image processing out of which a machine/computer/software intends to decipher the physical contents of an image or a sequence of images. One of the ways to do this is by comparing selected facial features from the image and a facial database. Recognizing emotion from images has become one of the active research themes in image processing and in applications based on human-computer interaction. This research conducts an experimental study on recognizing facial emotions. The flow of our emotion recognition system include the basic process in FER system. These include image acquisition, preprocessing of an image, face detection, feature extraction, classification and then when the emotions are classified the system assigns the user particular music according to his emotion. Our system focuses on live images taken from the webcam. The aim of this research is to develop automatic facial emotion recognition system for stressed individuals thus assigning them music therapy so as to relief stress. The emotions considered for the experiments include happiness, Sadness, Surprise, Fear, Disgust, and Anger that are universally accepted. [2]

3. CNN based Recognition of Emotion and Speech from Gestures and Facial Expressions : IEEE (2022)

The major mode of communication between hearing-impaired or mute people and others is sign language. Prior, most of the recognition systems for sign language had been set simply to recognize hand signs and convey them as text. However, the proposed model tries to provide speech to the mute. Firstly, hand gestures for sign language recognition and facial emotions are trained using CNN (Convolutional Neural Network) and then by training the emotion to speech model. Finally combining hand gestures and facial emotions to realize the emotion and speech. [3]

4. Automatic Recognition of Emotions in Speech With Large Self-Supervised Learning Transformer Models : IEEE (2023)

Speech Emotion Recognition (SER) is an important area of research in the realm of collaborative and social robotics, which aims to enhance human-robot interaction (HRI) and serves as a feedback mechanism for affective computing. Despite the recent progress in SER research area, it remains a challenging research problem due to the profound variations in the complexity, subjectivity, and contextual heterogeneity of human emotional expressions. Consequently, the inherent difficulties of modeling paralinguistic emotional information embedded in speech signals are further compounded when employing supervised learning, as it necessitates annotated labels for a large scale dataset for satisfactory model performance. To this end, self-supervised learning (SSL) approach is widely adopted in the speech domain to addresses this problem of limited availability of annotated data. Therefore, the focus of our research is to investigate and evaluate several state-of–the-art large attention-based self-supervised learning (SSL) models for the task of automatic speech emotion recognition (SER) on the challenging RAVDESS dataset.

CHAPTER 3

Problem Statement

3. Problem Statement

In the realm of human-machine interaction (HMI), understanding and responding to human emotions play a pivotal role in creating seamless and intuitive interfaces. However, existing methods predominantly rely on uni-modal emotion recognition, often overlooking the complexity and nuance of human emotional expression. To bridge this gap, there is a pressing need for advanced techniques that integrate multiple modalities, such as facial expressions, vocal intonations, gestures, and physiological signals, to accurately perceive and interpret human emotions.

This project aims to develop a robust multimodal emotion recognition system that enhances the efficacy of HMI across various domains, including virtual assistants, robotics, gaming, and healthcare. Key objectives include:

1. Data Fusion and Integration: Designing algorithms to seamlessly fuse information from diverse modalities, ensuring a holistic understanding of human emotions.

2. Feature Extraction and Representation: Identifying discriminative features within each modality and establishing effective representations to capture the subtle nuances of emotional states.

3. Machine Learning Models: Developing machine learning models capable of leveraging multimodal data for accurate emotion classification and continuous emotion tracking in real-time scenarios.

4. Adaptability and Personalization: Creating adaptive systems that can dynamically adjust to individual differences in emotional expression and user preferences, thereby enhancing user experience and engagement.

5. Robustness and Real-World Deployment: Addressing challenges related to environmental variability, noise, and inter-subject variability to ensure the reliability and practical applicability of the system in real-world settings.

By addressing these objectives, this project aims to propel the field of human-machine interaction towards a future where machines can not only comprehend human emotions but also respond in a nuanced and empathetic manner, ultimately fostering more natural and enriching interactions between humans and machines.

CHAPTER 4

Technology Stack

4. Technology Stack

This project delves into the intriguing realm of emotion detection, aiming to analyze humanemotions through text, audio, and video. We'll leverage the versatility of Flask, a Python web framework, to construct a user-friendly interface built with HTML. Machine learning, a powerful tool for pattern recognition, will be the engine driving this project. Specifically, we have train various machine learning models on meticulously curated datasets tailored to each modality – text, audio, and video. However, this ambitious project does present some technical hurdles. One challenge lies in acquiring a substantial amount of diverse data to train the models effectively, ensuring they can recognize emotions across a broad spectrum. Another hurdle involves achieving real-time performance for video analysis, as processing visual data can be computationally intensive. Despite these potential roadblocks, this project holds immense promise for its ability to interpret human emotions through various communication channels

CHAPTER 5

Proposed System & Implementation

5. Proposed system & Implementation

5.1 Block diagram of proposed system

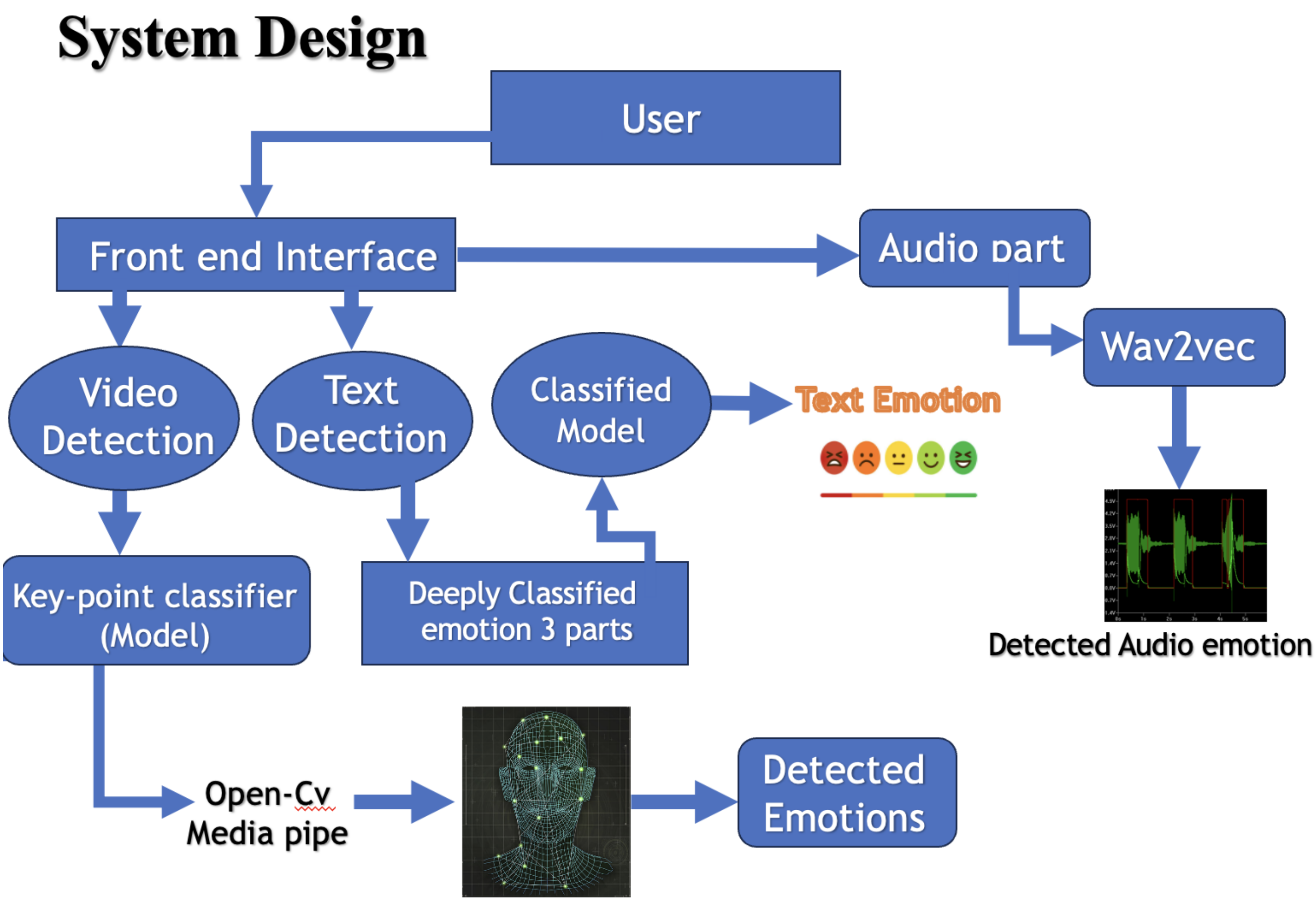
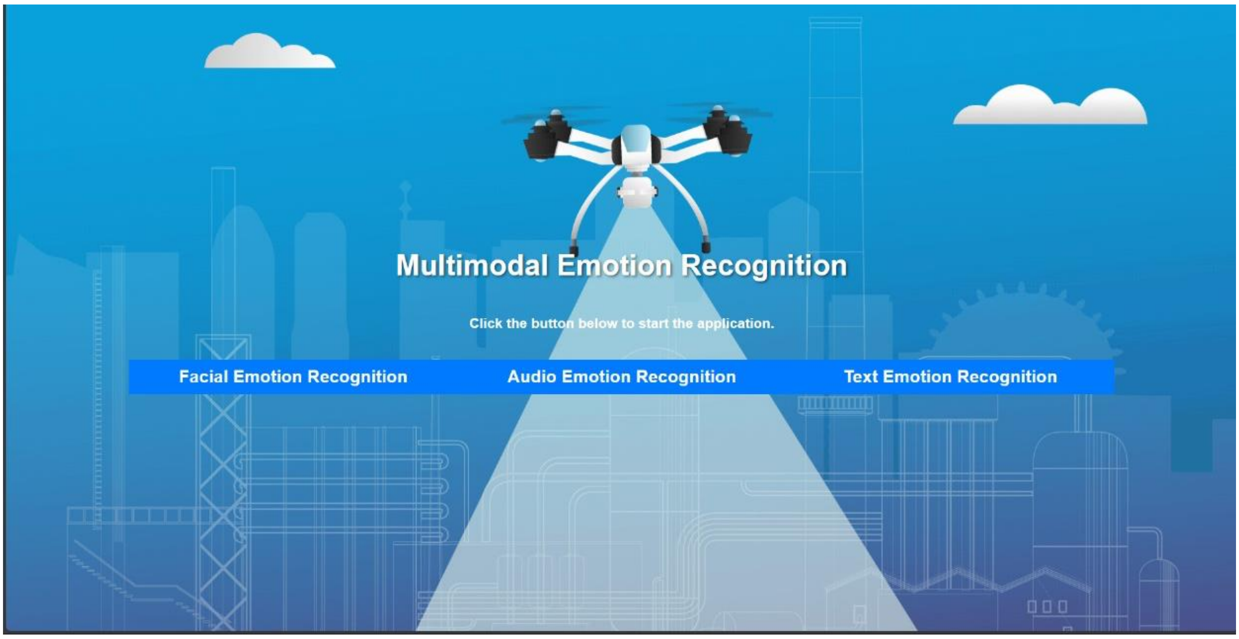


Figure 5.1Proposed System Design

5.2 Description of block diagram

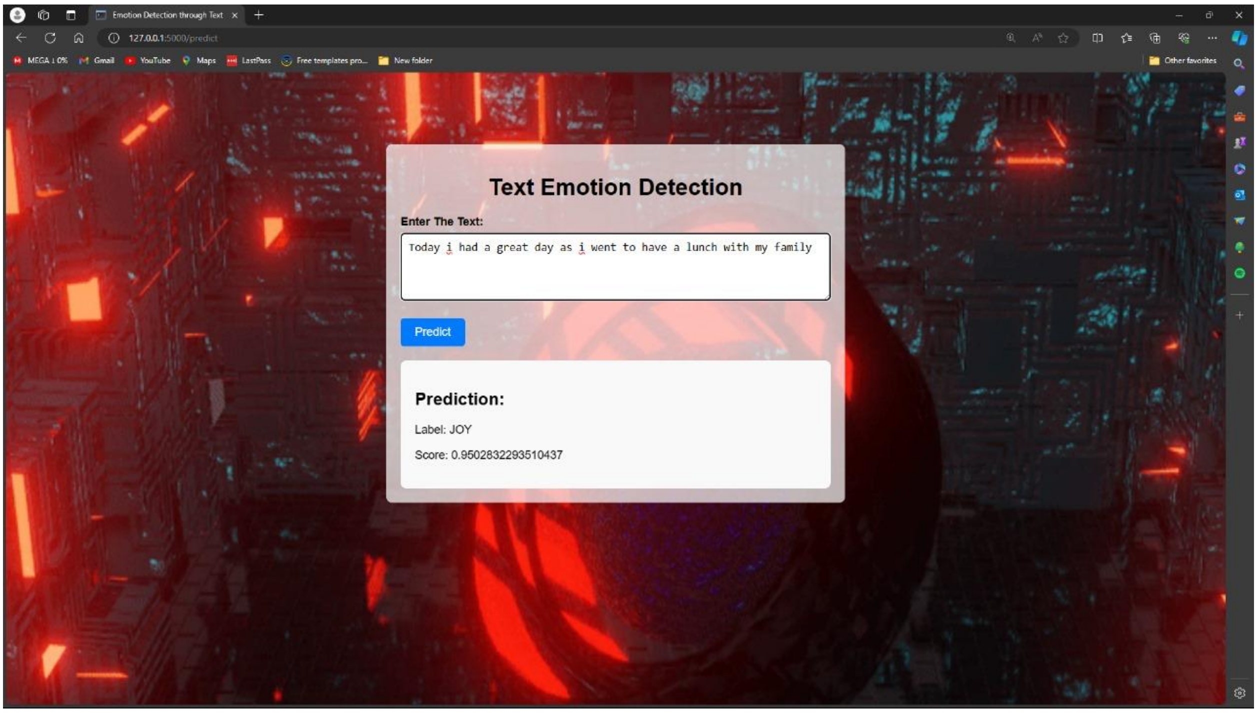
We've successfully integrated three advanced emotion detection systems, leveraging meticulously trained models with exceptional accuracy. Users are presented with a user-friendly homepage interface, featuring distinct buttons for video, text, and audio detection. These cutting-edge models have undergone rigorous testing, demonstrating their ability to reliably detect emotions across diverse media formats.

5.3 Implementation



Multimodal Emotion Recognition

Click the button below to start the application.



Text Emotion Detection

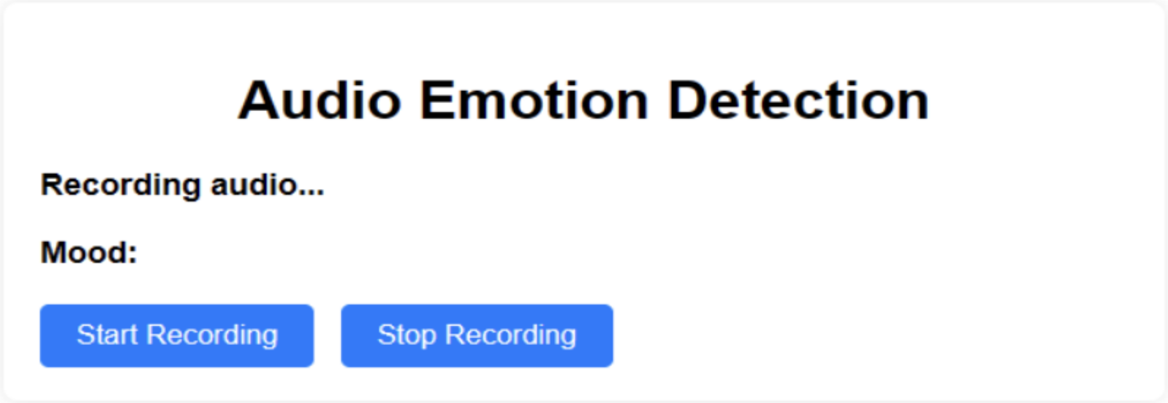
Enter The Text:

Today i had a great day as i went to have a lunch with my family

Prediction:

Label: JOY

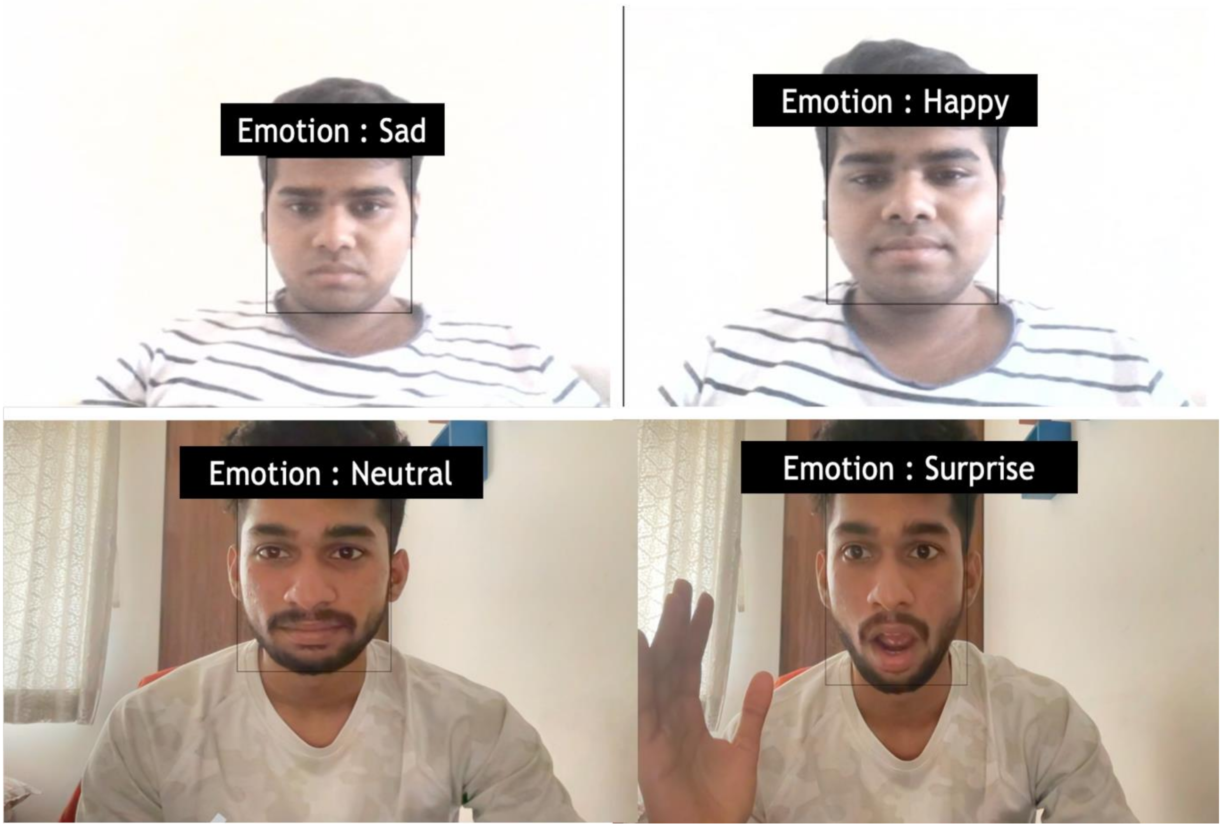
Score: 0.9502832293510437



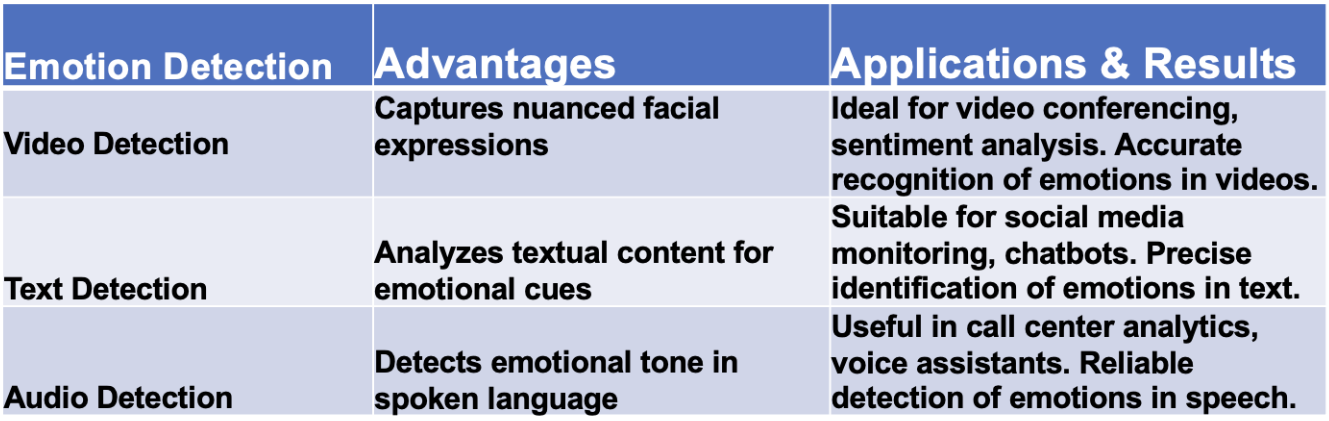
Audio Emotion Detection

Recording audio...

Mood:



5.4 Advantages/ Application/ result table are included in the table below



Emotion Detection	Advantages	Applications & Results
Video Detection	Captures nuanced facial expressions	Ideal for video conferencing, sentiment analysis. Accurate recognition of emotions in videos.
Text Detection	Analyzes textual content for emotional cues	Suitable for social media monitoring, chatbots. Precise identification of emotions in text.
Audio Detection	Detects emotional tone in spoken language	Useful in call center analytics, voice assistants. Reliable detection of emotions in speech.

CHAPTER 6

Conclusion

6.Conclusion

This project embarks on a captivating journey into the realm of multi-modal emotion detection, aiming to unveil the hidden emotional tapestry woven through text, audio, and video. We envision a future where technology transcends mere word recognition, delving into the depths of human sentiment. Here, we propose a comprehensive system that leverages the power of machine learning and user-centric design to achieve this ambitious goal.

At the heart of this project lies the potent combination of Flask, a versatile Python web framework, and HTML, the language that constructs user interfaces. This synergy allows us to create a user-friendly platform where individuals can interact with the system. But the true magic lies beneath the surface, where machine learning models take center stage. These models, meticulously trained on carefully curated datasets specific to each modality (text, audio, and video), will be the driving force behind emotion recognition.

However, the path to success is not without its challenges. Acquiring a substantial amount of diverse data is paramount to training robust models capable of recognizing a wide spectrum of emotions across various communication styles. This data collection presents a significant hurdle, as it requires capturing a representative sample of human interactions to ensure the models' generalizability. Another technical challenge lies in achieving real-time performance for video analysis. Processing visual data can be computationally intensive, and ensuring a seamless user experience with minimal latency requires optimization strategies.

Despite these potential roadblocks, the potential rewards are truly transformative. Imagine a future where customer service interactions are no longer sterile exchanges, but rather empathetic experiences informed by real-time analysis of a customer's emotional state. Educational environments could leverage this technology to personalize learning approaches based on student emotions, fostering a more engaging and effective learning experience. In the healthcare domain, emotion detection could revolutionize patient-doctor interactions, aiding in early diagnoses of mental health conditions or tailoring treatment plans to individual needs.

In conclusion, this multi-modal emotion detection project transcends the realm of mere technical exploration. It presents a thrilling opportunity to bridge the gap between human expression and machine comprehension. As we overcome the technical hurdles and refine the system, we pave the way for a future where technology can not only understand our words but also the emotions that resonate beneath them. This project holds the potential to reshape human-computer interaction, fostering a future where technology becomes a more empathetic and insightful companion.

References

[1] M. Sajid, M. Afzal and M. Shoaib, "Multimodal Emotion Recognition using Deep Convolution and Recurrent Network," 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 128-133, doi: 10.1109/ICAI52203.2021.9445262. IEEE..

[2] R. S. Deshmukh, V. Jagtap and S. Paygude, "Facial emotion recognition system through machine learning approach," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2017, pp. 272-277, doi: 10.1109/ICCONS.2017.8250725. IEEE..

[3] H. Avula, R. R and A. S Pillai, "CNN based Recognition of Emotion and Speech from Gestures and Facial Expressions," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 1360-1365, doi: 10.1109/ICECA55336.2022.10009316. IEEE..

[4] M. P. Gavali and A. Verma, "Automatic Recognition of Emotions in Speech With Large Self-Supervised Learning Transformer Models," 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings), Mount Pleasant, MI, USA, 2023, pp. 1-7, doi: 10.1109/AIBThings58340.2023.10292462. IEEE..

[5] Park, Seo-Hui, Byung-Chull Bae, and Yun-Gyung Cheong. "Emotion recognition from text stories using an emotion embedding model." In 2020 IEEE International Conference on big data and smart computing (BigComp), pp. 579-583. IEEE, 2020.

[6] Ristea, Nicolae-Cătălin, Liviu Cristian Duţu, and Anamaria Radoi. "Emotion recognition system from speech and visual information based on convolutional neural networks." In 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 1-6. IEEE, 2019.

[7] Pattun, Geeta, and Pradeep Kumar. "Emotion Classification using Generative Pre-trained Embedding and Machine Learning." In 2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), pp. 1-6. IEEE, 2023.

[8] Kwon, Junhwan, Kyeong Teak Oh, Jaesuk Kim, Oyun Kwon, Hee Cheol Kang, and Sun K. Yoo. "Facial Emotion Recognition using Landmark coordinate features." In 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 4916-4918. IEEE, 2023.