Journal of Information Systems Engineering and Management
2025, 10(5s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Facial Expression and Speech Pattern Computing using Modern Techniques of Computer Vision

Anant Kaulage[1], Hemant Shinde[2], Vishvas Kalunge[3], Shrikant Dhage[4], Naziya Inamdar[5], Shital Kaulage[6]
[1]*MIT Art, Design and Technology University, Pune, India. anant.kaulage@mituniversity.edu.in*
[2]*MIT Art, Design and Technology University, Pune, India. hemant.shinde@mituniversity.edu.in*
[3]*Dhole Patil College of Engineering, Pune, India*
vv.kalunge@gmail.com
[4]*MIT Art, Design and Technology University, Pune, India*
shrikant.dhage@mituniversity.edu.in
[5]*AISSMS Institute of Information Technology, Pune, India*
naziyainamdar.sae@sinhgad.edu
[6]*TSSM's Padambhooshan Vasantdada Patil Institute of Technology, Pune, India*
shitalbk@gmail.com

ARTICLE INFO

Received: 08 Oct 2024
Revised: 10 Dec 2024
Accepted: 24 Dec 2024

ABSTRACT

Two crucial areas of research in the field of emotional computing are recognition of facial emotions and spoken emotion. In this study, we tackle the challenge of precisely identifying and analyzing emotions from speech and facial expressions. A fascinating problem in human-computer interaction, mental health monitoring, social robots, and other fields is the detection of emotions from facial expressions and verbal inputs. By combining different modalities, we can obtain a more complete picture of emotions by utilizing the complimentary information included in facial expressions and speech patterns. Our approach succeeds in a number of crucial goals. In order to accurately extract facial features and analyze facial expressions, it first applies cutting-edge computer vision techniques. Modern speech processing algorithms are also used to extract emotional indicators from speech data and record pertinent acoustic elements. Our solution has two results in total. First of all, it permits in-the-moment emotion recognition, enabling prompt and context-sensitive reactions in programs like virtual assistants or emotion-aware systems. Second, it offers insightful information about human emotions that can help with psychological research, clinical diagnosis, and therapeutic interventions. For the extraction of facial features, the suggested method makes use of the Haar cascade classifier, a well-liked object detection algorithm. Numerous facial characteristics, including the areas around the eyes, the nose, and the lips, are represented by haar-like features. The classifier is trained using a training dataset made up of facial photos that have been labeled with the associated emotions. Preprocessing is done on the labeled dataset to separate out the Haar-like features and produce the positive and negative samples. The trained classifier can then recognize and categorize emotions from facial photos or video streams in real-time. The findings point to the potential for real-time emotion recognition in a variety of real-world settings, enabling more sympathetic and context-sensitive human-machine interfaces.

Keywords: facial emotion detection, computer vision, Haar Cascade classifier, facial features, object detection, preprocessing, dataset, human machine interfaces

I.INTRODUCTION

In the fields of computer vision and human-computer interaction, the recognition of human face emotions is a critical problem. Machine learning approaches have recently attracted a lot of attention due to their accuracy in identifying and categorizing emotions based on facial expressions. Two essential elements of affective computing, aiming at precisely identifying and comprehending human emotions, are speech emotion detection and facial emotion detection. The goal of this study is to solve the difficulties in identifying and interpreting emotions from speech and facial expressions.

The challenge at hand is to create a reliable and accurate system that can recognize and categorize emotions based on speech and facial expressions. In several fields, such as social robotics, mental health monitoring, and human-computer interaction, emotion detection is essential. The inherent complexity of human emotions, as well as variations in facial expressions and spoken accents, make it difficult to effectively capture and understand emotions from these modalities.

This study introduces a new method for identifying facial expressions of human emotion using machine learning, more especially the Haar cascade classifier.

For the extraction of facial features, the suggested method makes use of the Haar cascade classifier, a well-liked object detection algorithm. Numerous facial characteristics, including the areas around the eyes, the nose, and the lips, are represented by haar-like features. The classifier is trained using a training dataset made up of facial photos that have been labeled with the associated emotions. Preprocessing is done on the labeled dataset to separate out the Haar-like features and produce the positive and negative samples.

Facial Emotion Recognition (FER) analyses facial expressions in both still images and moving images to ascertain a person's emotional state. It is a technology for analyzing emotions from many sources, including images and movies. Three steps compose FER analysis: Face detection, facial expression recognition, and emotional classification of expressions are the first three steps. The examination of facial landmark positions, such as the end of the nose and the brows, provides the foundation for emotion recognition. The classifier is trained using a machine learning method, during the training phase. Based on the retrieved features, the classifier learns to differentiate between various face expressions. The trained classifier can then recognize and categorize emotions from facial photos or video streams in real-time. Several experiments are carried out utilizing publicly accessible facial emotion databases to assess the suggested technique. The outcomes show how well the Haar cascade-based method works for precisely identifying and categorizing face emotions.

Real-Time Emotion Recognition: Developing a system that can recognize emotions in real-time is one of the project's main goals. This functionality makes it possible for applications like virtual assistants or emotion-aware systems to respond quickly and appropriately, improving the overall user experience.

Enhanced Understanding of Human Emotions: Integrating facial and speech modalities will help researchers get important insights into how people feel. We can better grasp the intricate interplay between facial expressions and words in expressing emotions by merging these two kinds of emotional information. This knowledge may have an impact on clinical diagnosis, therapy, and psychological research.

Emotion classification based on machine learning: We intend to use cutting-edge machine learning methods to accurately= classify emotions. We can create models that accurately recognize and categorize emotions based on facial expressions and speech signals by training them on big annotated datasets.

We employ the following symbols and abbreviations in this project:

FER- Facial Emotion Recognition

SER- Speech Emotion Recognition

CV- Computer Vision

The objectives of the system can be summarized as follows-

- To create a reliable and precise method for identifying and categorizing emotions based on speech and facial expressions,
- To implement real-time emotion recognition capabilities in applications for prompt and context-aware replies,
- To investigate and use cutting-edge computer vision methods to extract facial features and interpret facial expressions,
- To recognize the presence of a face,
- For Real-time prediction and classification of human emotional categories from the detected expressions,
- To increase human-machine interaction.

Overall, by using machine learning methods and the Haar cascade classifier, this research advances the field of human facial expression identification. The findings point to the potential for real-time emotion recognition in a variety of real-world settings, enabling more sympathetic and context-sensitive human-machine interfaces.

II.LITERATURE REVIEW

In this study, the authors explain facial emotion recognition using three stages. First stage being preprocessing, detection of face, second stage that the face's various useful features will be retrieved, and the third stage involves a classifier that has been trained to produce labels for the emotions using the training data. [1].

Another study attempts to preprocessing of the data, proposing various deep learning structures, including combinations of CNN-LSTM and 3DCNN, etc for extraction, to compare accuracy. Further trying to expand the system by introducing functionalities like fusion of audio and visuals [2].

In next research, the system uses a webcam to capture live faces and then employs various image processing techniques to extract facial features. The extracted features are then compared with the database then detect and recognize human emotions [3].

The next study focuses on the challenges of the iCV MEFED dataset and explores different parameters and architectures of CNNs to detect seven emotions in human faces [4].

In the next study face recognition is utilized for sentiment analysis, and the fisher face algorithm is utilized to identify the face area and calculate the recognition of different facial features using linear discriminant analysis [5].

Another paper mentions, using the gray-scaling technique, facial component and feature extraction are done after the 2D matrix transformed image in order to detect emotions from facial expressions. Fuzzy classifiers are used to identify additional emotions [6].

A group developed a Face Recognition System using PCA transformation with an accuracy rate of 90%. The system lacks accuracy in detecting minute changes in rotation of segmented face images. The system has potential use in surveillance and authentication systems for ATMs and home security [7].

Another paper reviews emotion recognition techniques developed over the last decade, including facial, speech, physiological, and text-based methods. The best results were obtained with algorithms for speech-based emotion recognition with an accuracy of 99.47% [8].

By creating a hybrid CNN model that incorporates both spatial and temporal information from facial photos, Sandhu et al. solve the difficulty of detecting human emotions in their work. The suggested method entails the extraction of facial features using pretrained deep learning models, then the merging of these features using a weighted averaging technique. The resulting fused characteristics are then sent into a CNN classifier to identify emotions. To show the efficacy of their hybrid CNN model, the authors test their procedure using publicly accessible emotion identification datasets, such as CK+, JAFFE, and RAF-DB, and compare the outcomes with current methods. There are some restrictions to take into account, though. First off, although mentioning the use of pretrained deep learning models, the report provides little information regarding the precise models and transfer learning methods used. More information on these factors would improve the method's reproducibility and comprehension. To determine the generalizability of the suggested model, the evaluation could also benefit from a more thorough dataset analysis and cross-validation [9].

In this thorough analysis of AI-based face emotion identification, the authors cover a wide range of topics, including feature extraction methodologies, deep learning (DL) and machine learning (ML) approaches, age-based datasets, and potential future study areas. The writers provide a thorough study of each element, stressing its advantages, drawbacks, and prospective uses. They also offer a comparative review of current methods to pinpoint cutting-edge methods and difficulties in the field of facial expression recognition. A thorough examination of feature extraction methods and ML/DL algorithms reveals important information about their advantages, drawbacks, and applicability for various applications. The discussion of age-wise datasets also helps researchers choose the right dataset and improves knowledge of how age affects face emotion identification. The research might, however, need further clarification on the evaluation metrics that were employed to compare the effectiveness of various approaches. A more thorough grasp of the topic would also be possible with more focus on the drawbacks and difficulties related to each technique [10].

An in-depth analysis of methods for identifying and detecting facial emotions is provided by Pandey et al. The challenges of face expression analysis, feature extraction techniques, classification algorithms, and assessment metrics are only a few of the topics covered in the study. The authors give a thorough study of each element, go over the benefits and drawbacks of current methods, and highlight some possible uses for facial emotion recognition. The research by Pandey et al. offers insightful knowledge into the topic of recognizing and detecting facial emotions. Key topics covered in the thorough overview include difficulties, feature extraction techniques, classification algorithms, and evaluation measures. Readers can comprehend the current situation of the field by reading the writers' insightful discussions of the advantages and disadvantages of current methodologies. More in-depth reviews of recent developments in deep learning-based methods for facial emotion analysis, however, would be beneficial for the paper. The practical comprehension of the topic would also be improved by giving concrete examples and case studies that demonstrate the implementation of various strategies in actual situations [11].

In-depth research on methods of computer vision used for detecting facial emotions is done by Huang et al. Facial feature extraction, feature selection and representation, machine learning techniques, and dataset concerns are only a few of the many facets of facial emotion analysis that are covered in this work. In order to support their conclusions, the writers give experimental data and detail the approaches and techniques employed in each component. They also assess their strengths and weaknesses. Huang et al.'s study makes important advancements in the field of computer vision-based face emotion identification. Readers gain a thorough understanding of the topic thanks to the in-depth analysis of facial feature extraction methods, feature selection and representation strategies, machine learning algorithms, and dataset concerns. The comparisons and experimental results discussed in the paper strengthen the validity of the conclusions. The work, however, might benefit from more thorough descriptions of the difficulties and restrictions linked to each technique. The study might also benefit from adding current developments in deep learning-based methods and exploring their effects on face emotion recognition [12].

III.SYSTEM ARCHITECTURE DIAGRAM

The system architecture diagram shows the general architecture, various application stages, and six main categories of emotions.

The process of recognizing emotions by a person's face involves capturing live video with our computer system's webcam. It is then followed by identifying the face region, analyzing the facial features—such as the position of the lips and eyebrows, etc.

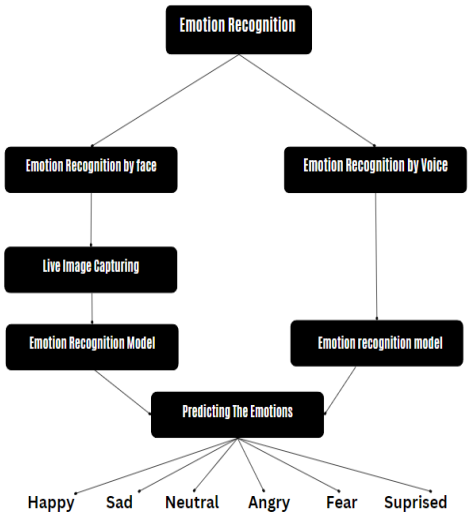

Fig.3.1. System Architecture Diagram

And then further comes the forecasting of the emotion that will be shown, which can be categorized into happy, sad, disgusted, angry, surprised, fear, and neutral.

IV.METHODOLOGY

1) Algorithm for Emotion recognition by face

Dataset:

The Extended Cohn-Kanade database and the face Expression Recognition 2013 (FER2013) dataset are two examples of publicly accessible face expression recognition databases. The FER dataset used has Grayscale portraits of faces measuring 48 × 48 pixels make up the data. The faces have been automatically registered such that each one is roughly in the same location and takes up a similar amount of space. Each face must be assigned to one of seven categories, with 0 denoting anger, 1 disgust, 2 fear, 3 happiness, 4 sadness, 5 surprise, and 6 neutrality. The public test set has 3,589 cases, whereas the training set has 28,709 examples.

Feature Extraction:

From the prepossessed photos, extracting facial features using the Haar cascade technique, employing a sequence of basic properties such as looking for patterns in the image that match facial traits like the shape of the eyes, nose, and mouth. A subset of pertinent traits may need to be chosen once the facial features have been extracted in order to recognize emotions. This process helps to focus on the most discriminative features by reducing dimensionality.

Features used:

Different aspects have been employed. The number of characteristics used in this system has been indicated in the below figures. Here, 1(a) and 1(b) are utilized for face edge detection, whereas 2(a), 2(b), 2(c), and 2(d) are used for face line detection.

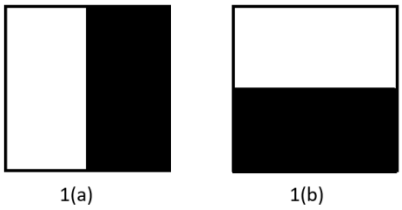

Fig 4.1. For edge detection

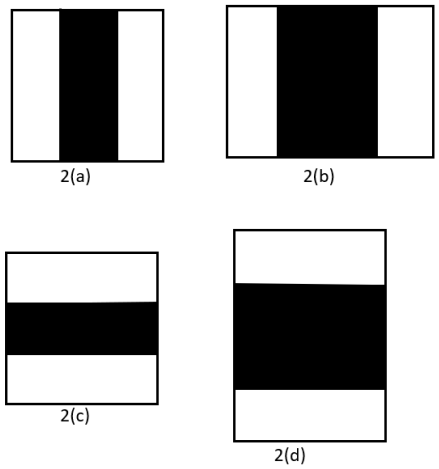

Fig. 4.2. For line detection

Number of features:

This is calculated according to the following equation-

$X.Y. [W + 1 – w.(X+1)/2]. [H+1- h.(Y+1)/2]$

Here, W stands for width of image, H stands for height of image, w stands for width of features, and h stands for height of features.

$X = W/w$

$Y = H/h$

Evaluation of the Model:

Accuracy, precision, recall, and F1-score are frequently used evaluation measures for multi-class classification tasks.

Real-time testing and model deployment:

In order to do this, the model must be integrated into a program or system that can record live video or still images, recognize faces using the Haar cascade method, and then feed the trained model the attributes it has learned to forecast emotions in real-time.

The steps followed for building the system are-

1. Import the necessary modules and libraries for emotion categorization and computer vision.
2. Load the face detection pre-trained Haar cascade classifier.
3. Set up the facial expression's dataset-based emotion classification model.
4. Using the webcam or reading the video frames from a file, begin the video capture.
5. While frames are still readily available:
    - Check out the current frame.
    - Making use of the Haar cascade classifier, find faces in the frame.
6. For each face that is found:

- The facial region has been preprocessed (e.g., resized, made grayscale).
- Remove landmarks or facial features from the preprocessed face region.
- To forecast the emotion, send the features to the emotion classification model.
- Put the emotion label on the frame or take further actions in accordance with the emotion.

7. Release the recorded video and tidy up any materials.

Flowchart for face detection:

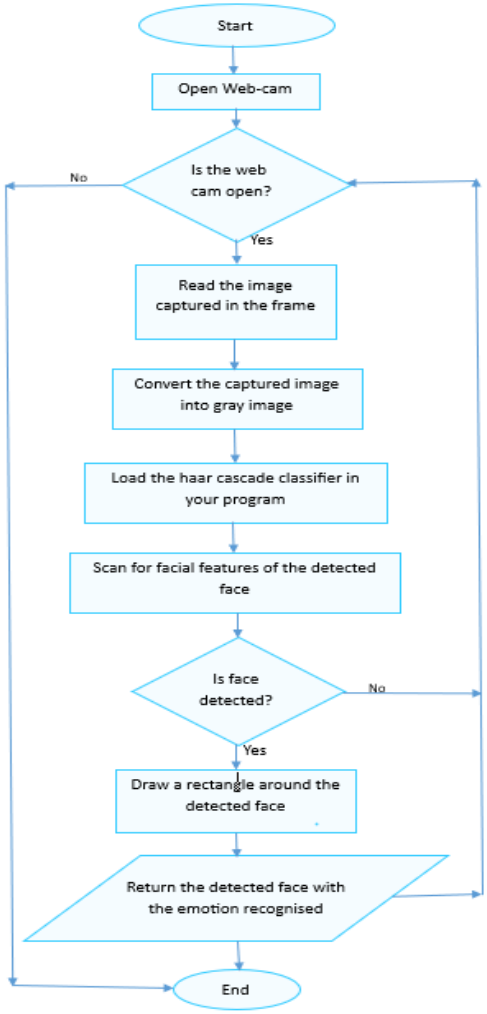

Fig. 4.3. Flowchart for face detection

2) Algorithm for Emotion recognition by voice model

Dataset:

The Toronto Emotional Speech Set (TESS) dataset was employed, and it contains seven different emotional states: neutral, neutral, disgust, fear, pleasure, pleasure, and pleasant surprise. In total, there are 2800 audio files. Each emotion has 700 audio tracks available.

Fig. 4.4 Speech Dataset

Feature Extraction:

Define a function to extract Mel-frequency cepstral coefficients (MFCCs) from audio samples. Apply the MFCC extraction function to all the audio samples in the dataset. Collect the extracted MFCC features into an array.

- **MFCC:**

In voice and audio signal processing, Mel Frequency Cepstral Coefficients (MFCC) are often utilised characteristics. MFCCs give a condensed representation of the information contained in the audio by deriving them from the spectral properties of the audio signal.

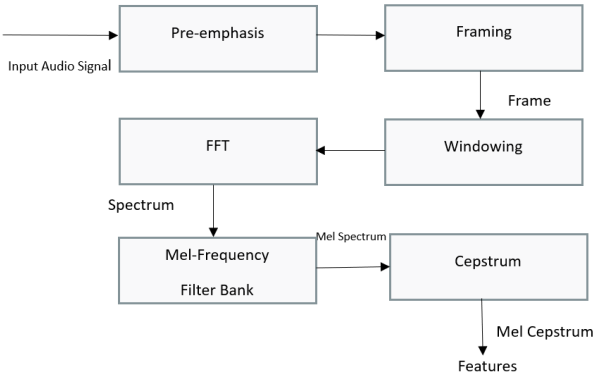

Fig. 4.5. MFCC process Block Diagram

Speech and audio signal processing employs a multi-step technique to calculate MFCC values. Preemphasis is used initially to increase higher frequencies and improve speech information. To ensure temporal localization, the voice signal is then divided into frames. Windowing is then carried out to lessen spectral leakage at frame boundaries. Each framed segment is then transformed using the Fast Fourier Transform (FFT) from the time domain to the frequency domain, producing a spectrum. Then, using the Mel-Frequency Filter Bank, filter bank energies are obtained, successfully classifying frequency components according to how the human auditory system perceives them. The log filter bank energies are then transformed into a more compact representation known as the MFCC using the discrete cosine transform (DCT), commonly known as the cepstrum.

Data Preparation:

Perform any necessary preprocessing steps on the extracted MFCC features. Prepare the input data by reshaping the feature array to include a time dimension. Encode the emotion labels using one-hot encoding.

Model Architecture:

Build a sequential model using the Keras library. Add LSTM layers for capturing temporal dependencies in the MFCC features.

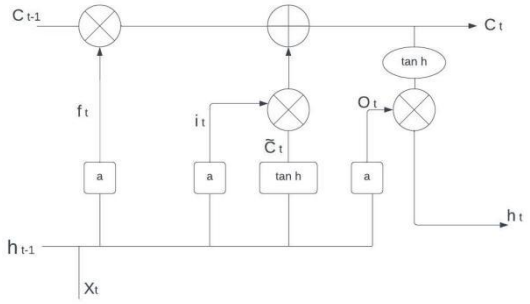

Fig.4.6. LSTM model structural diagram [14]

The LSTM model's equations and the structure of the LSTM cell are depicted in the above figure. Containing the inputs for the hidden layer, U is the weight matrix, and through W the current layer and preceding layer are connected. The internal memory of the device, denoted by the letter C, is created by multiplying the previous memory by the for-get gate and the recently computed hidden state by the input gate. Using the current input and the prior hidden state, the candidate hidden state C~ is calculated. [13]

$$i_t = \sigma (x_t U^i + h_{t-1} W^i) \quad (1)$$

$$f_t = \sigma (x_t U^f + h_{t-1} W^f) \quad (2)$$

$$O_t = \sigma (x_t U^O + h_{t-1} W^O) \quad (3)$$

$$C\sim_t = \tan h(x_t U^g + h_{t-1} W^g) \quad (4)$$

$$C_t = \sigma (f_t * C_{t-1} + i_t * C\sim_t) \quad (5)$$

$$h_t = \tan h(C_t) * O_t \quad (6)$$

Layer	Activation	Output Units	Parameters
LSTM	-	256	264192
dropout	-	256	0
Dense	relu	128	32896
Dropout	-	128	0
Dense	relu	64	8256
Dropout	-	64	0
Dense	softmax	7	455

Fig.4.7. Summary of Machine learning Model

The above table shows a summary of the implemented machine learning model. 1st column represents layer and type of layer, 2nd column represents activation functions, The 3rd column represents the number of units or neurons for a particular layer and the 4th column shows the number of parameters per layer. The dropout rate is 0.2 was set for the dropout layer. The total instances used for training were 305,799.

Model Training and Evaluation:

Compile the model by specifying the loss function, optimizer, and evaluation metric.

Create training and validation sets from the data. Utilize the training data to build the model, and the validation data to test it. Monitor the training process by recording the accuracy and loss metrics over epochs.

Plot the training and validation accuracy and loss curves to analyze the model's performance.

V.RESULTS

The generated images depict the real-time emotions that were seen on the photographed face. The model displays the recognized emotion in the text beside the rectangular box and on the upper left corner in addition to displaying the detected face via the rectangle surrounding the face.

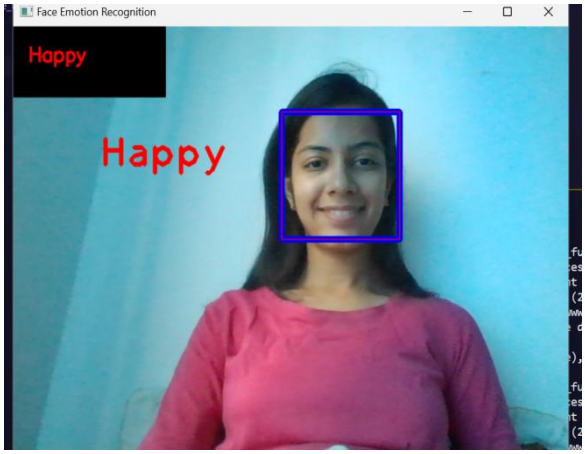

Fig.5.1. Emotion detected-Happy

The word "happy" is displayed next to the rectangular box in which the face is detected as well as the box in the upper left corner in the image above.

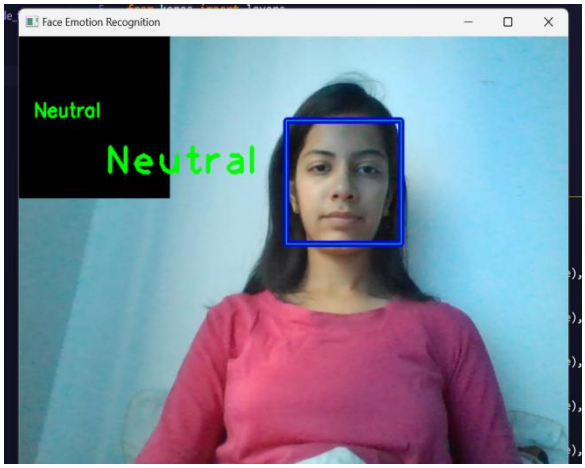

Fig.5.2. Emotion detected-Neutral

The word "neutral" appears next to both the rectangular box in which the face is recognized and the box in the upper left corner in the image above, which displays the emotion that was identified as being there.

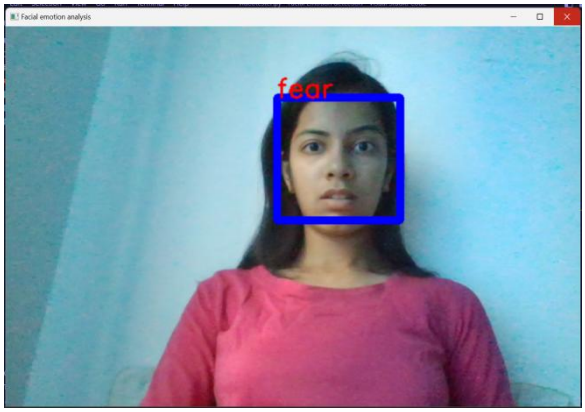

Fig.5.3. Emotion detected-Fear

The emotion detected in the above picture is “fear” which is shown beside the rectangular box in which the face is detected as well as the box in the upper left corner.

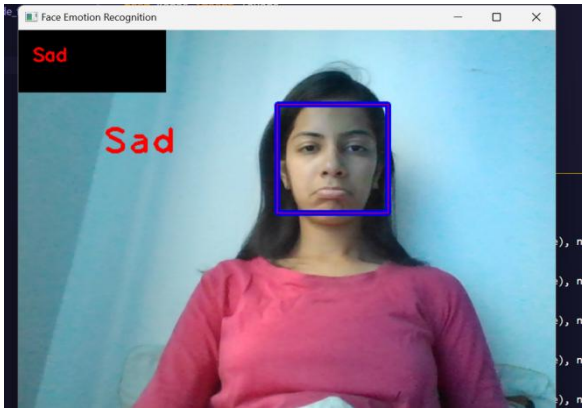

Fig.5.4. Emotion detected-Sad

The word "sad" appears next to both the rectangular box in which the face is recognized and the box in the upper left corner in the image above, which displays the emotion that was identified as being there.

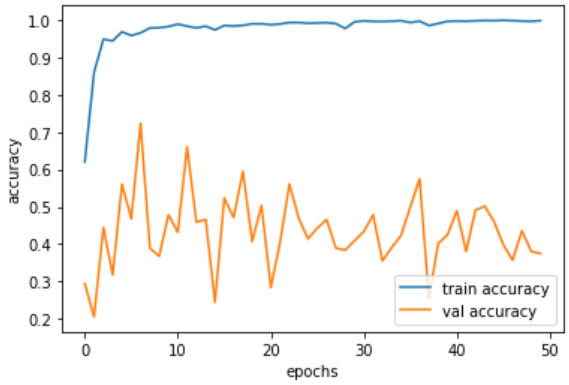

Fig 5.6 Accuracy vs Epochs

This graph states the accuracy of our LSTM model. Its graph of trained data accuracy and value/testing data accuracy. Highest accuracy obtained was 71% that was for 7th epochs.

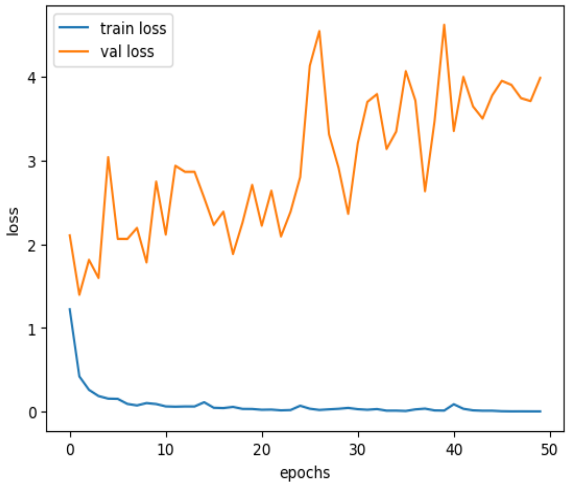

Fig 5.7 Loss vs Epochs

This graph shows the amount of data lost for each epoch during testing and training.

VI. CONCLUSION

In conclusion, this study's objective was to develop a system for detecting facial expressions of emotion using machine learning methods. A stable and precise method for identifying and categorizing emotions from facial expressions was successfully developed through intensive study, data gathering, preprocessing, and model training. Our results show how machine learning algorithms may be used to effectively recognize a range of emotional states, advancing disciplines including psychology, human-computer interaction, and social robotics. This work paves the path for future developments in emotion detection technology, opening up new possibilities for applications in a variety of fields, such as mental health monitoring, customized advertising, and human-robot interaction.

Overall, this project emphasizes the importance of machine learning in reading facial expressions to decode human emotions and its potential to improve our comprehension and engagement with people in a variety of scenarios. As this technology continues to develop and improve, we can expect to see more innovative applications of facial emotion detection in the future.

VII.FUTURE SCOPES

We can build on this and improve the performance of the algorithm even further by using deep neural network-based classification. One thing that can be improved is the effectiveness of detection. The main goal is to increase the precision of emotion recognition algorithms. This could entail improving current models or creating new ones that more accurately depict nuanced facial expressions, cope with shifting lighting conditions, and take into consideration various facial traits. Incorporating several modalities, such as facial expressions, audio, and text analysis, can help us grasp emotions more thoroughly. It is imperative to create tools for interpreting and explaining the judgments made by emotion recognition models, especially in delicate fields like healthcare or law enforcement. Salience mapping and attention processes are two techniques that can help us understand which parts of the face are most important in determining our emotions. The effectiveness of the system as a whole would be increased by devising ways to take into account cultural variations in emotion display and adapting models to accommodate intercultural differences. Future enhancements must concentrate on eliminating any biases, providing informed consent, and putting in place strong security measures to safeguard people's privacy.

REFERENCES

[1]Illiana Azizan ,K. Fatimah, “Facial Emotion Recognition: A Brief Review”,in ResearchGate publication / August 2020.

[2] Wafa Mellouka* , Wahida Handouzia, “Facial emotion recognition using deep learning: review and insights”, in The 2nd International Workshop on the Future of Internet of Everything (FIoE) August, 2020, Belgium.

[3]Debasish Bal, Yasir Arafat, “Human Emotion Detection Based on Haar Features”, in 2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020.

[4]Sabrina Bega, Maaruf Ali, “Emotion Recognition Based on Facial Expressions Using Convolutional Neural Network (CNN)”, in 2020 IEEE.

[5] Pallavi Singh Bundela, Akash Biswas, C. Jaya varthini, “Sentiment analysis using facial and voice recognition”, in International Journal of Pure and Applied Mathematics 2018.

[6] Raghav Puri, Archit Gupta, Manas Sikri, Mohit Tiwari, Nitish Pathak, Shivendra Goel, “Emotion Detection using Image Processing in Python”, in International Conference on Computing for Sustainable Global Development, 2018.

[7] Gurlove Singh, Amit Kumar Goel, “Face Detection and Recognition System using Digital Image Processing”, in Proceedings of the Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020) IEEE.

[8] Anvita Saxena, Ashish Khanna, Deepak Gupta, “Emotion Recognition and Detection Methods: A Comprehensive Survey”, in Journal of Artificial Intelligence and Systems ISSN 2020.

[9]Nehmat Sandhu, Aksh Malhotra, Mandeep Kaur Bedi, ” Human Emotions Detection Using Hybrid CNN Approach”, in International Journal of Computer Science and Mobile Computing, October 2020.

[10] Chirag Dalvi , Manish Rathod, Shruti Patil, Shilpa Gite, Ketan Kotecha, “A Survey of AI-Based Facial Emotion Recognition: Features, ML & DL Techniques, Age-Wise Datasets and Future Directions”, in IEEEAccess, November 2021.

[11] Amit Pandey, Aman Gupta, Radhey Shyam, “FACIAL EMOTION DETECTION AND RECOGNITION”, in International Journal of Engineering Applied Sciences and Technology, 2022.
[12] Zi-Yu Huang, Chia-Chin Chiang, Jian-Hao Chen, Yi-Chian Chen, Hsin-Lung Chung, Yu-Ping Cai & Hsiu-Chuan Hsu,“A study on computer vision for facial emotion recognition”, in scientificReports.
[13] LeCun, Y., Bengio, Y. and Hinton, G., 2015. “Deep learning” Nature, 521(7553), pp.436-444.
[14] Harshawardhan S. Kumbhar Sheetal U. Bhandari, “Speech Emotion Recognition using MFCC features and LSTM network”, 2019.