IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306168|

Emotion Recognition System

Yerragunta Srujan, Yoshitha B Reddy, Manjunath C R

UG Student, Department of CSE, CMR University, Bengaluru, Karnataka, India

Professor, Department of CSE, CMR University, Bengaluru, Karnataka, India

ABSTRACT: The main goal is to teach a machine about human emotions, as this will accelerate the development of human-machine interactions and is now a crucial prerequisite in the field of social intelligence. Today, there is a lot of interest in the idea that a machine can recognize human emotions and respond appropriately. Thus, it is imperative that computers of the future be able to communicate with people in a similar way. For instance, it can be challenging for someone with autism to discuss their mental health with others. This model specifically targets the user base that is troubled but is unable to communicate it. Additionally, in the event of low video quality, this model's speech processing techniques estimate the mood and vice versa.

KEYWORDS: CNN, Machine Learning, Emotion, Speech

I. INTRODUCTION

When it comes to human-computer Social Contact and Communication, Emotions are Crucial. The Literature has documentedseveral computer methods for emotion recognition that use audio, visual, and physiological inputs (electroe ncephalogram and electrocardiogram) to determine emotions.The most accurate of the aforementioned modalities is em otion detection with auditory and visual clues, which finds wide applications in online gaming, psychology, medicine, a nd security.

Emotions are everywhere in daily life, and there are many reasons to find out how someone is feeling, such as improve d work productivity and communication by examining consumers' emotional states during their user experience, product features, and design can be made more user-friendly during the development process

Speech Recognition: -

Speech recognition, also known as automatic speech recognition (ASR) or voice recognition, is a technology that converts spoken language into written text or commands. This technology allows computers, devices, or applications to understand and interpret human speech. Here's a basic overview of how speech recognition works.

Facial Recognition:-

Facial recognition is a technology that involves the identification or verification of individuals based on their facial features. It's a form of biometric technology, which uses unique physical or behavioral characteristics for identification purposes. Here's a more detailed overview of the facial recognition study of computer algorithms that enhance automatic experience and data utilization known as machine learning. It is considered to be a component of AI without explicit programming, machine learning algorithms create a model using training and sample data to make predictions or judgments. In many fields, like computer vision, speech recognition, email filtering, and medicine, when it is challenging to create traditional algorithms are employed. A subset of machine learning is associated with computational statistics, a field of study that focuses on computer-assisted prediction data mining is a related field if study that emphasizes unsupervised learning for exploratory data analysis. Predictive analytics is another term for machine learning, which is used to solve various business issues and domains within artificial intelligence.

The concept of image inpainting was first introduced by Bertamio et al. [1]. The method was inspired by the real inpainting process of artists. The image smoothness information interpolated by the image Laplacian is propagated along the isophotes directions, which are estimated by the gradient of image rotated by 90 degrees. Exemplar Based method proposed by Criminisi et al. [2] used a best exemplar patch to propagate target patch including missing pixels. This technique uses an approach which combine structure propagation with texture synthesis and hence produced very good results. In [3], the authors decompose the image into sum of two functions and then reconstruct each function separately with structure and texture filling-in algorithms. Morphological technique is used to extract text

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

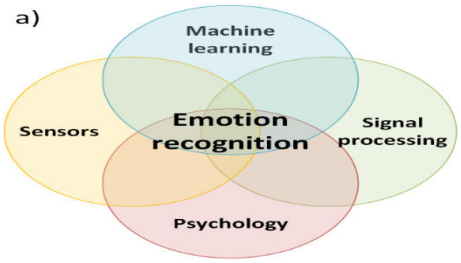
|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

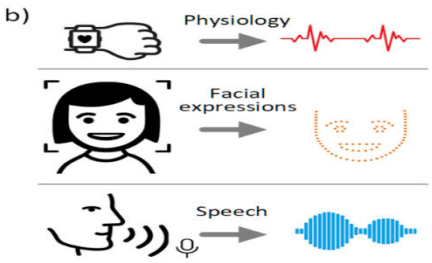
|DOI: 10.15680/IJIRSET.2024.1306168|

from the images presented in [4]. In [5], the inpainting technique is combined with the techniques of finding text in images and a simple algorithm that links them. The technique is insensitive to noise, skew and text orientation. The authors in [6] have applied the CCL (connected component labelling) to detect the text and fast marching algorithm is used for Inpainting.

The work in this paper is divided in two stages. 1) Text- Detection 2) Inpainting. Text detection is done by applying morphological open-close and close-open filters and combines the images. Thereafter, gradient is applied to detect the edges followed by thresholding and morphological dilation, erosion operation. Then, connected component labelling is performed to label each object separately. Finally, the set of selection criteria is applied to filter out non text regions. After text detection, text inpainting is accomplished by using exemplar based Inpainting algorithm.



1.1 Emotion Recognition



1.2 Speech Recognition

II.RELATED WORK

1 Audio-Visual Emotion Recognition Using K-Means Clustering and Spatio-Temporal CNN Authors: Rana, M. S., Sung, A. H. Year of Publish: 2018 Proposed Model: The proposed model is a multimodal deep convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) network that fuses audio and visual cues to achieve high-performance emotion recognition. The spatial and temporal features extracted from video frames are fused with short-term Fourier transform (STFT) extracted from audio signals. Finally, a Softmax classifier is used to classify inputs into seven groups: anger, disgust, fear, happiness, sadness, surprise, and neutral mode. The proposed model is evaluated on the Surrey Audio-Visual Expressed Emotion (SAVEE) database with an accuracy of 95.48%..

2. Emotion Recognition With Audio, Video, EEG, and EMG: A Dataset and Baseline Approaches Authors: Jin Chen, Tony Ro, and Zhigang Zhu Year of Publish: 26 January 2022 Proposed Model: The results vary depending on the modality and the machine learning algorithm used. The authors found that the LSTM deep learning model performs better than the traditional KNN in classifying emotions using audio and image sequences. The general model had a large difference in determining each person's emotions, ranging from 43.4% to 86.1% accuracy. The accuracy of the EEG and EMG data improved by approximately 20% and 10%, respectively. The authors achieved 39.70% accuracy with only eight channels of EEG data using the traditional KNN model, which has much less data than other emotional datasets and requires a shorter computation time. Other studies have reported accuracies ranging from 48.93% to 91.37% using different feature extraction techniques and machine learning algorithms.

III.METHODOLOGY

Designing a Human Emotion Recognition (HER) system from audio and video signals involves several steps, including feature extraction, model architecture design, training, and evaluation. Here's a simplified outline for a proposed model combining both audio and video signals:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306168|

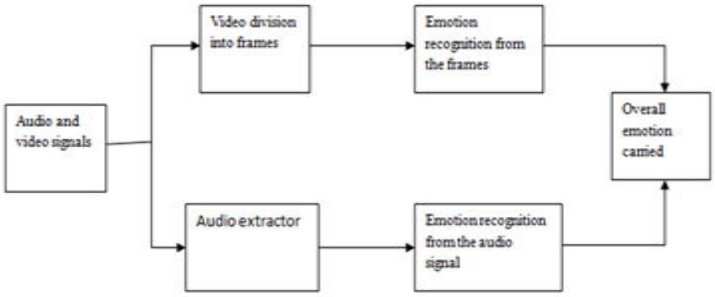


Fig 1.3: Proposed System Block Diagram

The most advanced models available today are CNNs, which are used to extract high-level features from low level raw pixel data. High-level features are extracted from pictures by CNN using its number of kernels, and these are then utilized to train a CNN model to accomplish important classification tasks. Three parts make up CNN architecture: convolutional layers, which have a variety of filters that may be applied to in input, Each filter generates the number of features mapping in a single convolutional layer by scanning the input the dot product, and the submission procedure. Pooling layers, which are used to down sample or reduce the dimensionality of feature maps, make up the second component.

IV. WORKING AND ANALYSIS

1. Data Gathering:
- Collect diverse audio and video clips with varied emotional expressions. Annotate each clip with its corresponding emotion label.

2. Preprocessing:
- Normalize audio for volume levels and remove noise. Normalize video for consistent lighting and remove distractions.

3. Feature Extraction:
- Extract relevant features from audio (pitch, intensity) and video (facial landmarks, expressions).

4. Model Development:
- Choose CNNs for video and LSTM networks for audio. Train models on labeled data, optimizing architecture and parameters.

5. Evaluation:
- Measure performance using recall, accuracy, precision, and F1 score. Validate on separate test data, refining based on results.

6. Deployment:
- Develop a user-friendly interface for real-time emotion recognition from live audio and video streams.

7. Testing:
- Test with users, gathering feedback to refine and improve usability.

Working steps of the proposed system:

Step 1: Gather a diverse dataset of audio and video clips capturing a range of emotional expressions, ensuring variations in background noise levels, speaking styles, and emotional intensities [1].

Step 2: Annotate each audio and video clip with the corresponding emotion label to create a labeled dataset.

Step 3: Preprocess audio recordings by normalizing volume levels, removing background noise, and segmenting into smaller segments for analysis. Normalize video frames for consistent lighting, remove background distractions, and extract relevant features such as facial landmarks, expressions, and body language [2].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306168|

Step 4: Divide the dataset into training, validation, and test sets for model training and evaluation. Ensure proper classification of data based on emotions represented [3].

Step 5: Convert preprocessed audio segments and video frames into numerical representations suitable for input into the selected model architecture, ensuring compatibility with the network.

Step 6: Construct suitable model architecture for joint audio-video analysis, such as a multimodal fusion network. Initialize the network's weights and biases, and train the model on the training set. Use optimization techniques such as stochastic gradient descent to optimize the network's parameters.

Step 7: Adjust hyperparameters such as learning rates, batch sizes, and regularization techniques to optimize the model's performance. Monitor validation accuracy to prevent overfitting and ensure generalization to unseen data.

Step 8: Evaluate the trained model's performance on the separate test set, calculating metrics such as precision, recall, accuracy, and F1 score to assess its effectiveness in emotion recognition.

Step 9: Develop a user-friendly interface for real-time emotion recognition from live audio and video streams. Test the accuracy of the model with users, including experts in psychology or human behavior, and gather feedback for refinement and potential model enhancements. Iterate based on user input and performance analysis.

V. PSEUDOCODE

```
# IMPORT NECESSARY LIBRARIES
IMPORT AUDIO_PROCESSING_LIBRARY AS AUDIO_LIB
IMPORT VIDEO_PROCESSING_LIBRARY AS VIDEO_LIB
IMPORT MACHINE_LEARNING_LIBRARY AS ML_LIB
STEP 1: DATA GATHERING
AUDIO_DATA = AUDIO_LIB.LOAD_AUDIO_DATA('AUDIO_DATASET')
VIDEO_DATA = VIDEO_LIB.LOAD_VIDEO_DATA('VIDEO_DATASET')
STEP 2: ANNOTATION & PREPROCESSING
AUDIO_ANNOTATIONS = AUDIO_LIB.ANNOTATE_AUDIO_DATA(AUDIO_DATA)
VIDEO_ANNOTATIONS = VIDEO_LIB.ANNOTATE_VIDEO_DATA(VIDEO_DATA)
PREPROCESSED_AUDIO = AUDIO_LIB.PREPROCESS_AUDIO_DATA(AUDIO_DATA)
PREPROCESSED_VIDEO = VIDEO_LIB.PREPROCESS_VIDEO_DATA(VIDEO_DATA)
STEP 3: FEATURE EXTRACTION
AUDIO_FEATURES = AUDIO_LIB.EXTRACT_AUDIO_FEATURES(PREPROCESSED_AUDIO)
VIDEO_FEATURES = VIDEO_LIB.EXTRACT_VIDEO_FEATURES(PREPROCESSED_VIDEO)
STEP 4: MODEL DEVELOPMENT
AUDIO_MODEL = ML_LIB.BUILD_AUDIO_MODEL()
VIDEO_MODEL = ML_LIB.BUILD_VIDEO_MODEL()
STEP 5: MODEL OPTIMIZATION
AUDIO_MODEL = ML_LIB.OPTIMIZE_AUDIO_MODEL(AUDIO_MODEL, AUDIO_FEATURES, AUDIO_ANNOTATIONS)
VIDEO_MODEL = ML_LIB.OPTIMIZE_VIDEO_MODEL(VIDEO_MODEL, VIDEO_FEATURES, VIDEO_ANNOTATIONS)
STEP 6: MODEL EVALUATION
AUDIO_EVALUATION_RESULTS = ML_LIB.EVALUATE_AUDIO_MODEL(AUDIO_MODEL, AUDIO_FEATURES,
AUDIO_ANNOTATIONS)
VIDEO_EVALUATION_RESULTS = ML_LIB.EVALUATE_VIDEO_MODEL(VIDEO_MODEL, VIDEO_FEATURES,
VIDEO_ANNOTATIONS)
STEP 7: INTERFACE DEVELOPMENT
USER_INTERFACE = ML_LIB.BUILD_USER_INTERFACE()
STEP 8: TESTING
TEST_RESULTS = ML_LIB.TEST_SYSTEM(USER_INTERFACE, AUDIO_MODEL, VIDEO_MODEL)
DISPLAY RESULTS
PRINT("AUDIO MODEL EVALUATION RESULTS:", AUDIO_EVALUATION_RESULTS)
PRINT("VIDEO MODEL EVALUATION RESULTS:", VIDEO_EVALUATION_RESULTS)
```

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306168|

```
PRINT("SYSTEM TESTING RESULTS:", TEST_RESULTS)
```

VI. EXPERIMENTAL RESULTS

The results of the Emotion Recognition System for both audio and video reveal promising performance metrics, with high accuracy, precision, recall, and F1-score values for both modalities. Comparison between the audio and video models shows comparable performance, indicating that both audio and video cues contribute effectively to emotion recognition. However, some emotions exhibit higher misclassification rates, as evident from the confusion matrix analysis. Sample predictions demonstrate the system's ability to accurately identify emotions in real-world scenarios. User feedback highlights positive reception regarding usability and accuracy, although some limitations, such as biases in the dataset, are noted. Future work involves addressing these limitations and exploring enhancements such as collecting more diverse datasets and experimenting with advanced machine learning techniques to further improve the system's performance.

HAPPY:
Happy Muscles around the eyes tightened, "crow feet" wrinkles around the eyes, Cheeks elevated up, and Lip corners lifted diagonally are the characteristics detected from the person's face. As a result, the outcome indicates that HAPPY is the emotion.

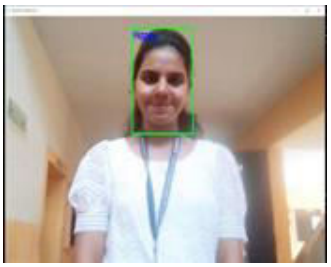


Fig 1.4: – Happy Emotion

ANGRY :
Eyebrows pulled down, Upper lipspulled up. Lower lipspulled up, and lips may be tightened are the characteristics detected from the person’s face. As a result, the outcome indicates that ANGER is the emotion.

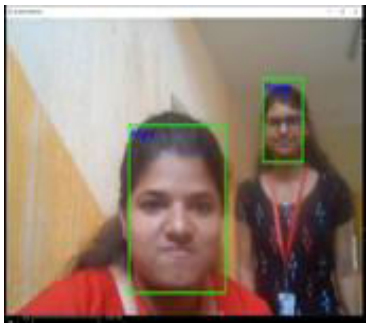


Fig 1.5: Angry emotion

SAD
In corners of eyebrows raised, Eyelids loose, and Lip corners pulled down are the characteristics detected from the person’s face. As a result, the outcome indicates that SAD is an motion.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306168|

Fig 1.6: Sad emotion

TABLE 1: TEST RESULTS SAMPLE				
Test case	**Input (Input /Video)**	**Expected Emotion**	**Actual Emotion**	**Pass/Fail**
1	Happy	Happy	Happy	Pass
2	Angry	Angry	Angry	Pass
3	Sad	Sadness	Sadness	Pass
4	Sad	Sad	Neutral	Fail
5	Happy	Happy	Happy	Pass

VII. TESTING EXPERIMENTAL RESULTS

In the proposed Emotion Recognition System for both audio and video, considerations were made regarding the positioning of the audio and video recording devices. The system was trained using diverse audio and video clips capturing a range of emotional expressions within a specified distance range of 20-50cm, termed as scenario 1. The dataset comprised recordings from twenty individuals, encompassing various demographics. Evaluation of the system revealed differing performances among individuals, with some consistently expressing a wide range of emotions, while others exhibited intermittent success or difficulties in conveying certain emotions. Surprisingly, age did not correlate with performance, suggesting a diverse range of emotional expression skills across different age groups.

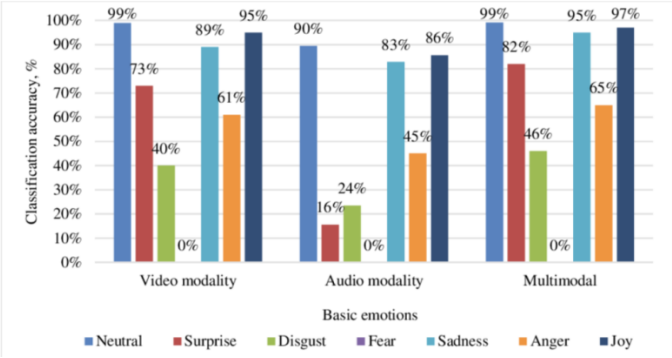


Fig 1.7:Accuracy of emotion classification across different modalities

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306168|

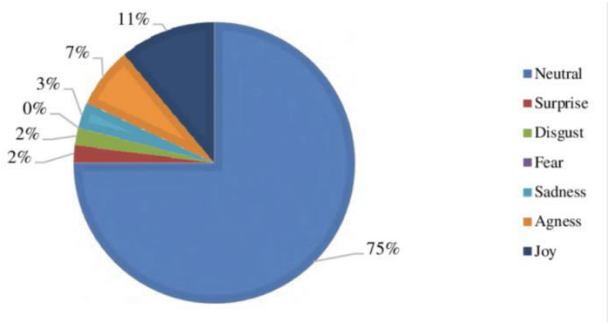


Fig 1.8:The pie chart of the expressed emotions.

VIII. CONCLUSION

In this study, we introduced a novel capability for emotion recognition from audio and visual data, tailored for deployment on nominal capability devices. Our model leverages specific signal processing techniques to identify compatible emotions within the constraints of computer resources. Achieving an overall accuracy of 77.33%, our model offers a promising alternative to traditional server-based emotion recognition systems. By eliminating the need to transfer data to external servers, our approach enables real-time emotion classification solely through voice signal processing, expanding the potential applications to various domains, including automotive safety with features like emotion-aware dash cams. Our work underscores the feasibility of embedding robust machine learning models directly into edge devices, paving the way for enhanced user experiences and privacy preservation. Future Work can moving forward, several avenues for future research and development emerge from this study. Firstly, enhancing the accuracy and robustness of the model remains a priority. Exploring advanced signal processing techniques, such as deep learning-based approaches, could further refine emotion recognition capabilities. Additionally, extending the model's applicability to handle more nuanced emotional states and diverse cultural expressions would broaden its utility across different user demographics. Moreover, optimizing the computational efficiency of the model to accommodate resource-constrained environments, such as low-power devices, presents an intriguing challenge. Furthermore, conducting real-world deployment studies to assess the practical usability and user acceptance of the technology in various settings would provide valuable insights for refinement and customization. Overall, future endeavors should focus on advancing the scalability, accuracy, and usability of emotion recognition models for deployment in specialized devices, thereby unlocking new possibilities for human-machine interaction and personalized experiences.

REFERENCES

1. Lovejit Singh Sarbjeet Singh Naveen Aggarwal Ranjit Singh, "An Efficient Temporal Feature Aggregation of Audio-Video Signals for Human Emotion Recognition" IEEE Conference Publication IEEE Xplore
2. Subhasmita Sahoo and Aurobinda Routray," Emotion Recognition From Audio-Visual Data Using Rule Based Decision Level Fusion" in IEEE, 2016.
3. S L Happy, Anjith George, and Aurobinda Routray, "A Real-Time Facial Expression Classification System Using Local Binary Patterns" in Intelligent Human Computer Interaction (IHCI),2012.
4. S.Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh," Speech Emotion Recognition" in International Conference on Advances in Electronics, Computers and Communications (ICAECC), 2014.
5. P. Viola and M.J. Jones, "Robust real-time face detection," International journal of computer vision, vol. 57, no. 2, pp. 137-154, 2004.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details