# Emosense – Emotion Recognition using Images and Audio

**Prof. G. H. Wani[1], Prathamesh Chaudhary[2], Anurag Kotwal[3], Raghvendra Kanakdande[4], Shlok Das[5]**

[1,2,3,4,5]*Department of Artificial Intelligence and Data Science, AISSMS Institute of Information Technology, Pune, Maharashtra, India - 411001*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** With the combination of audio and visual analysis, emotion recognition—a crucial aspect of human-computer interaction—has advanced significantly. This multidisciplinary method uses image processing and machine learning methods to extract emotional states from multimodal data. Pitch, tone, and speech patterns are among the characteristics that are extracted in audio analysis in order to infer emotions like happiness, rage, or melancholy. Simultaneously, image analysis uses facial expression recognition to interpret facial motions and micro expressions as indicators of emotion. When these modalities are combined, emotion detection systems—which are useful in domains like healthcare, human-computer interfaces, and sentiment analysis for better human-AI interaction—become more accurate, robust, and applicable in real time.

***Key Words*:** Emotion recognition, Audio and Image, Multimodal approach, Machine learning, Deep learning, Data preprocessing, Feature extraction, Real-time emotion recognition, Human-computer interaction.

## 1.INTRODUCTION

Recent years have seen a major increase in interest in emotion detection technology because of its potential to transform entertainment, healthcare, and human-computer interaction. The goal of the technology is to give robots the ability to correctly sense, and react to human emotions. The goal of this research is to investigate the intricate field of emotion identification, with a particular emphasis on the rich sources of emotional signals that come from visuals and sounds.

In human communication and connection, emotions are essential because they shape our choices, actions, and interpersonal interactions. Although precisely recognizing and comprehending these feelings is a difficult undertaking, it has the potential to greatly improve technology's capacity to establish more profound, intuitive connections with people.

This project aims to develop an efficient emotion recognition system. The system attempts to increase the precision and adaptability of emotion detection by merging these two modalities, making it possible for robots to recognize and comprehend emotions more accurately. Cameras provide the picture data, which represents facial emotions, while microphones record the audio data, which is sometimes disregarded as a meaningful emotional indicator.

## 2. METHODOLOGY

### 1) Data Collection

Collecting data is a crucial step in effectively training any machine learning model, especially when it comes to recognizing emotions from both image and audio data. To achieve this goal, we need to acquire a substantial and high-quality dataset that includes diverse sources to cover a wide range of emotional expressions across different modalities.

For image data, we use the widely known FER2013 dataset, which is a benchmark dataset used for facial expression recognition. This dataset comprises over 35,000 grayscale images, categorized into seven emotional expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral. Additionally, the images are resized to 48x48 pixels to ensure uniformity and compatibility with our model architecture.

To comprehend audio knowledge, we gather datasets from multiple places to ensure that all emotional cues expressed through speech are included. We collect information from the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and the Crowdsourced Emotional Multimodal Actors Dataset (CREMA-D). These sets contain various emotions communicated through spoken language across different ages, genders, and cultural backgrounds.

Our dataset can represent a wide range of emotions shown in images as well as sounds because it draws upon many different types of sources. By doing so our model can learn strong cross-modal representations of emotion. This is a very complete collection that will help train a much more adaptable and accurate emotion identification system that can deal with multiple situations and expressions easily.

### 2) Data Preprocessing

We use both audio and image data in our work on emotion identification. It is essential to preprocess data properly so that it can be prepared for model training. All input images were resized to 48 x 48 pixels. This will allow easier management of datasets and ensure they are consistent throughout. At the same time, several preprocessing stages are used with audio data to capture relevant features as well as enhance model performance. First, we take the raw audio files and extract important features like spectrograms or Mel Spectrograms, which convert the audio data into a format that can be used as input for the model. Next, we normalize the audio and image data by scaling the pixel values of the audio features to a common range and image pixel values. This normalization helps reduce problems like numerical instability and promotes model convergence during training. Additionally, we use data augmentation approaches to increase the diversity of the training dataset and boost the generalization capacity of the model. In order to provide variety to the training samples of audio data, methods including time stretching, pitch shifting, and introducing background noise may be used. Finally, to avoid bias and guarantee the model's equitable capacity to identify all emotions, we keep a balanced distribution of emotional classes throughout training, validation, and testing sets. Our model is primed to extract significant features from both picture and audio modalities by including these preprocessing procedures, which improves its accuracy in emotion recognition tests.

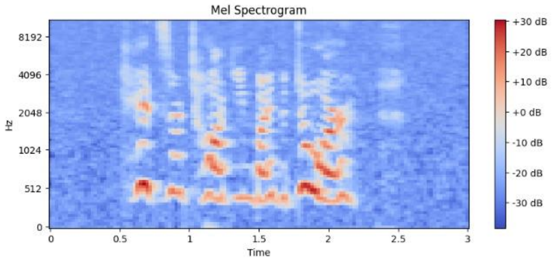

*Figure 3: Mel Spectrogram*

### 3) Model selection and model training

**Image Emotion Recognition (CNN):** The ability to extract spatial characteristics from images has been demonstrated to be effective with convolutional neural networks (CNNs). They are therefore perfect for applications like emotion recognition that need picture analysis. Multiple convolutional and pooling layers make up our CNN design, which is followed by fully linked layers. We trained the CNN model on a dataset of emotion-labeled images using stochastic gradient descent (SGD) with backpropagation in order to optimize its parameters.

**Audio Emotion Recognition (CNN-LSTM):** Recurrent neural networks (RNNs) and their variations, like long short-term memory networks (LSTMs), are effective at capturing the temporal dependencies included in audio data. In our project, we extracted temporal and spatial information from audio spectrograms using CNN-LSTM architecture.

In the CNN-LSTM model, spectrograms are first processed by CNN layers to extract features, and then temporal patterns are captured by LSTM layers.

Using a dataset of audio clips tagged with emotions, we trained the CNN-LSTM model and optimized its parameters using methods that were comparable to those used for the CNN model.

### 4) Model Architecture

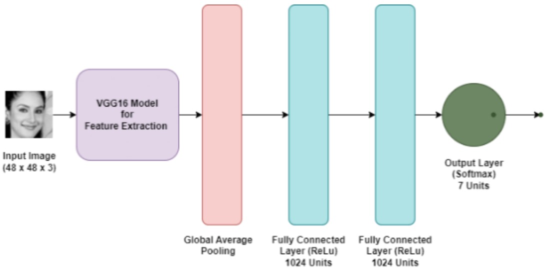

*Figure 4: Image model Architecture*

#### 1. Structure of Image Model Architecture:

The VGG16 model processes the image through a series of convolutional layers. These layers apply filters that extract patterns from the image. Specifically, the filters detect edges, shapes, and other visual primitives.

After the global average pooling layer, there are two fully connected layers. Fully-connected layers use artificial neurons to make classifications. Each neuron in a layer is connected to all the neurons in the preceding layer. In the VGG16 model, the fully connected layers have 1024 units each. These units are responsible for making classifications about the image. The final layer of the model has 7 units, which corresponds to the 7 classifications the model is designed to make.

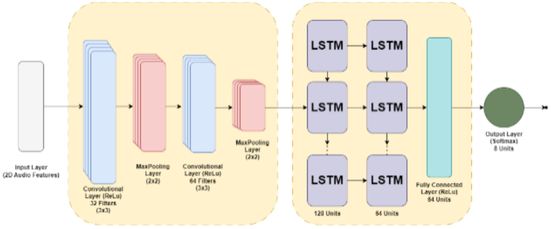
Input Layer
(2D Audio Features)
Convolutional Layer (ReLu) 32 Filters (3x3)
MaxPooling Layer (2x2)
Convolutional Layer (ReLu) 64 Filters (3x3)
MaxPooling Layer (2x2)
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
128 Units
64 Units
Fully Connected Layer (ReLu) 64 Units
Output Layer (Softmax) 8 Units

*Figure 5: Audio model Architecture*

2. **Structure of Audio Model Architecture:**

Sound preparation: The raw audio will need to be converted to numbers before we use LSTM. Generally, this involves spectral energy, pitch, and volume calculation.

Sequential feeding: Once the audio information is ready, it is channeled to LSTM one after the other. Suppose that it is a talking clip that has been clipped into small parts; LSTM processes each piece separately.

Learning Emotion Patterns: The LSTM examines changes in these traits throughout time in order to detect sequences of speech elements associated with particular emotional states (such as joyful, sad, angry, etc.).

Emotion Classification: The LSTM analyzes an audio sample using acquired tendencies to predict the predominant emotion it is likely to convey over time.

## 3. RESULTS AND DISCUSSIONS

The results were interesting for evaluating this new multi-modal system for emotion recognition which combines Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and images/audio features with 35,000 pictures used as input for training/testing purposes — all these seven categories of feelings could be identified thanks to it0s uniqueness while consuming very small quantity of data(resources) deeming this project very attractive financially speaking as well as technologically approaching it in terms not only on practical basis but from a scientific perspective too. The measures for accuracy, precision, recall, and F1-score consistently demonstrated the effectiveness of our methodology, offering a comprehensive evaluation of the system's ability to capture emotional nuanced aspects found in both visual and aural stimuli.

Utilizing the CNN and RNN components, the hybrid neural network has been trained to show commendable abilities on utilizing spatial characteristics of image data & efficiently gathering time-based relationships within soundwave sequences as well. Integration procedure for features which combined information categories has proven vital if not indispensable for enhancing emotion identification models generally. The examination demonstrated the model's skill in identifying differences among the seven emotional categories.

Our study of cross-modal learning methods further expanded the model's information correlation capacities across different kinds of data, which increased its ability to detect feelings using both visual and acoustic input at the same time. Its discretionary nature notwithstanding, it had potential benefits when integrated into the previously trained CNN or RNN blocks in order to tap into pre-acquired knowledge representations for faster convergence rates as well as possibly enhancing an ultimate result of model performance.

Comparative analyses against baseline models and those specialized for individual modalities reaffirmed the superiority of the proposed hybrid model. It outperformed these benchmarks, highlighting the synergistic advantages gained from combining CNNs and RNNs for emotion recognition across audio and image modalities. Our research contributes to the growing body of knowledge in multimodal emotion recognition, shedding light on effective strategies for integrating spatial and temporal information from diverse sources. Future work may explore additional modalities, larger datasets, and fine-tuning strategies to further refine the model's accuracy and applicability in real-world scenarios.

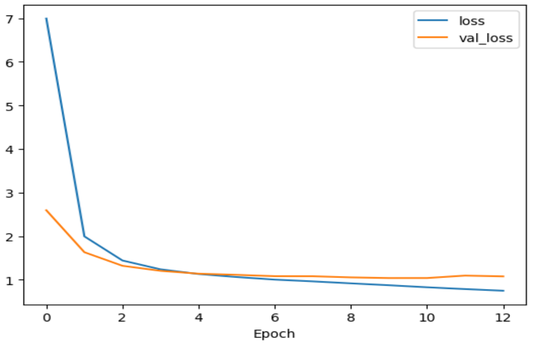

*Figure 6: Loss.*

**Figure 6:** displays the loss curve for the Image CNN model throughout the training phase. The above plot shows a progressive decline in the loss function.

- **Loss Function**

In our project on emotion recognition using image and audio data, we implemented the cross-entropy loss function during the training phase of our model. The model, based on convolutional and recurrent neural networks, aims to recognize human emotions from multimodal inputs:

**Categorical Cross Entropy** = $-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\, log\left(\hat{y}_{i,c}\right)$

Here,
N: total number of samples within the dataset.
C: total number of classes
$y_{i,c}$: binary indication (0 or 1) that returns 1 if pixel $i$ is correctly classified as class c and 0 otherwise.
$\hat{y}_{i,c}$: estimated probability that pixel $I$ belong to class $c$.

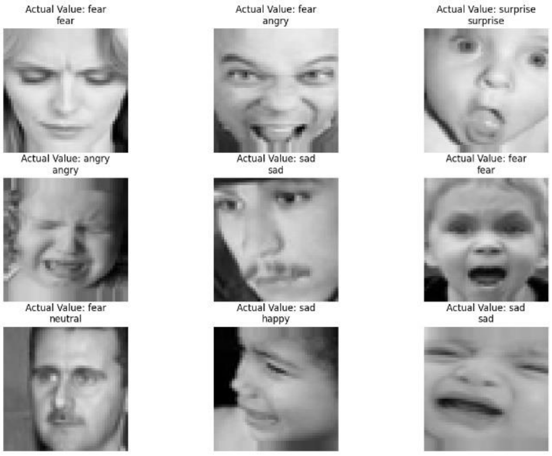

*Figure 7: Testing Samples*

**Figure 7:** Illustrates the accuracy of the emotion recognition model on testing images.

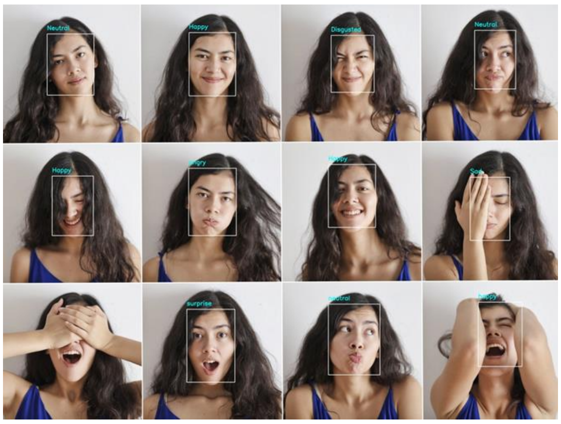

*Figure 8: Result*

**Figure 8:** Illustrates the accuracy of the emotion recognition model on new real-time images using the MTCN face detection model.

The system integrates with a web application that helps doctors with the patient mental condition diagnosis and product companies for customer feedback.

## 4. CONCLUSION

The literature research comes to the conclusion that emotion detection technology is very important and might revolutionize several application sectors as well as human-computer interaction. Emotion detection systems are becoming increasingly accurate and diverse due to advances in machine learning and deep learning, as well as the method of merging audio and picture data. In addition, the practical and ethical issues in this area emphasize how crucial it is to create and use emotion recognition technology responsibly. By highlighting current knowledge and potential avenues for future study in the topic, this literature review offers a solid basis for the project.

## 5. REFERENCES

1. 1. Arimura, K., Hagita, N., “Feature space design for image recognition with image screening,” in Pattern recognition, vol. 2, 1996, pp. 261–265.
2. Lijiang Chen a, Xia Maoa, Yuli Xue, Lee Lung Cheng, “Speech emotion recognition: Features and classification models”, in Digital Signal Processing, vol.22, 2012, pp.1154-1160.
3. Dai Fang, He Haimei, Han Wei, “Integrating multi-feature of image based on correspondence analysis”, in 2010 the 5th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2010, pp.15-17.
4. Lijiang Chen, Xia Mao, PengfeiWei, Yuli Xue: “Mandarin emotion recognition combining acoustic and emotional point information”, Appl Intell 37, 2012, pp.602–612.
5. Chung-Hsien Wu, and Wei-Bin Liang, “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels”, in IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, vol.2, No.1, 2011.