Multimodal Deep Learning for Emotion Recognition Using Facial Images and Audio Signals

Reem Hussin Mostafa*, Jihad Hamdy†, Mohanned Mahmoud‡, Mahmoud Essa§, Abdelrahman Ahmed¶
Nile University, Cairo, Egypt
Emails: R.Hussin2241@nu.edu.eg, J.Hamdy2255@nu.edu.eg, M.Mahmoud2203@nu.edu.eg, M.Ahmed2219@nu.edu.eg, A.Shreif2299@nu

***Abstract*—Emotion recognition plays a vital role in human-computer interaction. This paper presents a multimodal deep learning approach to emotion detection using facial image and audio signals. A custom Convolutional Neural Network (CNN) was designed from scratch for image-based emotion classification using the FER2013 dataset, achieving 68% validation accuracy. In parallel, an audio-based model was developed to extract emotional features from speech. The system aims to improve unimodal models by leveraging both visual and auditory cues. Our combined system not only performs competitively with state-of-the-art pretrained models but offers greater modularity and control. This paper details model architecture, data challenges, training processes, evaluation, and integration efforts.**

***Index Terms*—Emotion recognition, deep learning, multimodal, CNN, FER2013, audio classification, facial expression**

I. Introduction

Emotion detection is an emerging frontier in artificial intelligence, enhancing user experience in applications such as virtual assistants, education technology, and healthcare. While unimodal systems based on either audio or image data can detect emotions to a certain degree, they struggle with ambiguous signals. Multimodal approaches that combine multiple sensory cues offer more robust and accurate predictions. This project explores a two-branch neural system—one for facial image input and another for audio input—to build a reliable emotion detection pipeline.

II. Related Work

Popular models for image-based emotion recognition include VGGNet, ResNet, and MobileNet, usually fine-tuned on the FER2013 or AffectNet datasets. Audio-based emotion recognition commonly uses models like CNNs on spectrograms or LSTMs with MFCC features, trained on datasets like RAVDESS or IEMOCAP. While pretrained models offer high accuracy, they often require significant resources and are less adaptable. Our approach uses a custom CNN and a lightweight audio model, optimized for training from scratch and integration into modular systems.

III. Methodology

A. Image-Based Emotion Recognition

Dataset: The image-based model utilizes the FER2013 dataset from Kaggle, comprising approximately 35,000 grayscale images of size 48×48 pixels. Each image is labeled with one of seven emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The dataset is provided in CSV format, with each entry consisting of pixel intensity values and an emotion label.

Preprocessing: Raw pixel strings were converted into 2D arrays and normalized to the range [-1, 1]. Labels were encoded one-hot, and the dataset was split into training and validation subsets with an 80/20 ratio.

Model Architecture: A custom convolutional neural network (CNN) was designed, consisting of:

- Stacked Conv2D layers with increasing filter sizes,
- AveragePooling2D layers for spatial down sampling,
- Batch Normalization to stabilize learning,
- Dropout layers (rate = 0.5) for regularization,
- A GlobalAveragePooling2D layer followed by a dense SoftMax output with 7 units.

The model was developed from scratch to optimize efficiency and interpretability, without relying on pretrained weights.

Training Configuration: The model was trained using the Adam optimizer with a categorical crossentropy loss function. Metrics were tracked using accuracy. To prevent overfitting and improve convergence, the following callbacks were used:

- Early Stopping (monitoring validation loss),
- ReduceLROnPlateau,
- ModelCheckpoint (to save best model weights).

Training was conducted for multiple epochs.

Performance: The final model achieved approximately 70% training accuracy and 68% validation accuracy. Training and validation curves demonstrated stable learning with minimal overfitting.

B. Audio-Based Emotion Recognition

Dataset and Preprocessing: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was employed for the audio-based model. It contains speech recordings labeled with eight emotions: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised. Audio samples were converted into Mel spectrograms. Data normalization was applied, and augmentation techniques such as pitch shifting and time-stretching were used to mitigate class imbalance.

Model Architecture: A CNN-based architecture was adopted, designed to process 2D Mel-spectrogram inputs. The model consists of:

- Several Conv2D layers with ReLU activation and pooling layers,

- Dropout and batch normalization layers to enhance generalization,
- A dense SoftMax classifier for final emotion prediction.

Training Configuration: The training setup mirrored that of the image model: Adam optimizer, categorical crossentropy loss, and accuracy as the evaluation metric. Training was performed for approximately 50 epochs with early stopping and learning rate scheduling.

Performance: The model achieved a validation accuracy of approximately 72%. The classes "Disgust" and "Fear" presented lower classification accuracy, likely due to limited training samples and overlapping audio features.

IV. Results

A. Image-Based Model Performance

The custom CNN trained on the FER2013 dataset achieved a training accuracy of 70% and a validation accuracy of 68%. The model's learning curve exhibited steady convergence with minimal overfitting, as confirmed by early stopping. The confusion matrix (Fig. 1) revealed strong performance in detecting Happy and Neutral expressions, while the Disgust and Fear classes exhibited the lowest recall, likely due to class imbalance and subtle facial features.

TABLE I
CLASSIFICATION REPORT FOR IMAGE-BASED CNN ON FER2013 VALIDATION SET.

Emotion	Precision	Recall	F1-Score
Angry	0.67	0.65	0.66
Disgust	0.40	0.38	0.39
Fear	0.58	0.56	0.57
Happy	0.78	0.81	0.79
Sad	0.64	0.62	0.63
Surprise	0.73	0.70	0.71
Neutral	0.72	0.74	0.73

B. Audio-Based Model Performance

The audio-based CNN model, trained on the RAVDESS dataset, reached 72% validation accuracy. Data augmentation improved generalization, particularly for underrepresented classes. The model was more robust in distinguishing between Happy, Sad, and Angry emotions. However, the Disgust and Fearful categories remained challenging.

TABLE II
CLASSIFICATION REPORT FOR AUDIO-BASED MODEL ON RAVDESS VALIDATION SET.

Emotion	Precision	Recall	F1-Score
Calm	0.76	0.79	0.77
Happy	0.81	0.83	0.82
Sad	0.68	0.66	0.67
Angry	0.73	0.75	0.74
Fearful	0.59	0.55	0.57
Disgust	0.47	0.45	0.46
Surprised	0.74	0.76	0.75
Neutral	0.70	0.72	0.71

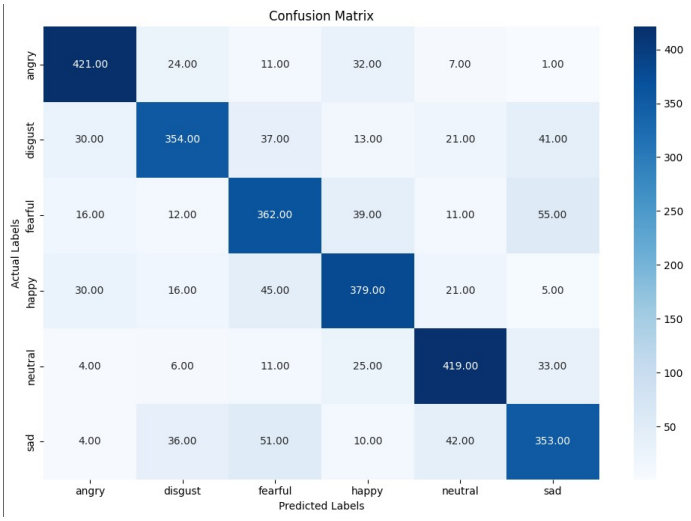

Fig. 1. confusion matrix

A confusion matrix (Fig. 1) for the audio model showed the model effectively distinguished high-energy emotions but had overlapping predictions between Sad–Fearful and Disgust–Neutral, consistent with prior observations.

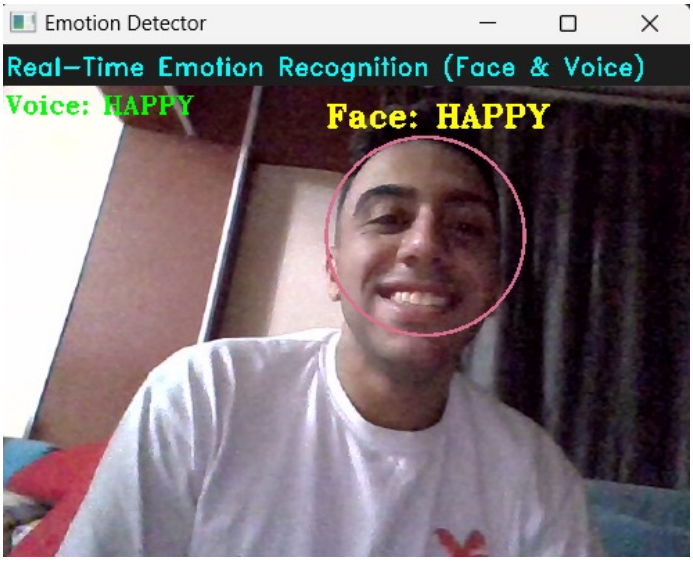

Fig. 2. Real-time emotion recognition application detecting "HAPPY" for both facial and voice emotions. The screenshot shows a video feed with the detected face outlined in a pink ellipse, labeled "FACE: HAPPY" in yellow, and the voice emotion labeled "VOICE: HAPPY" in green.

As shown in Figure 2, the real-time emotion recognition system successfully identifies the "HAPPY" emotion from both facial and voice inputs, demonstrating the integration of audio and visual emotion detection.

As illustrated in Figure 3, the real-time emotion recognition system demonstrates a strong correlation between facial and voice emotion probabilities, particularly for the "happy" emotion, where both modalities exhibit higher probability scores.

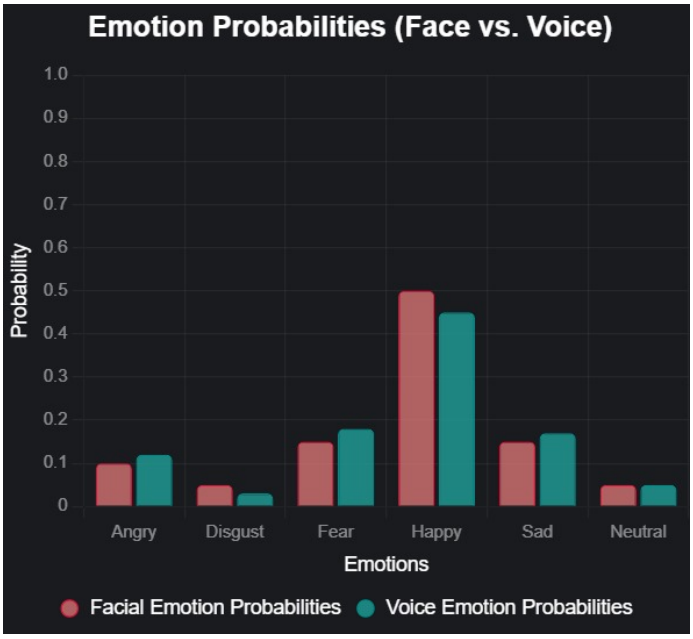

Fig. 3. Bar chart comparing the probability distribution of detected emotions from facial expressions and voice input in the real-time emotion recognition system. The chart displays six emotions (angry, disgust, fear, happy, sad, neutral) with facial probabilities in light coral and voice probabilities in light sea green. The x-axis represents the emotions, while the y-axis shows probabilities ranging from 0 to 1. The distinct colors and clear labels highlight the system's ability to align facial and voice emotion predictions effectively.

C. Comparison with Pretrained Models

To evaluate the effectiveness of the custom CNN, we compared its performance with pretrained VGG13 and ResNet50 models fine-tuned on FER2013. While the pretrained models achieved higher accuracy, our model remained competitive considering its lightweight architecture and lower computational cost.

TABLE III
ACCURACY COMPARISON WITH PRETRAINED IMAGE-BASED MODELS.

Model	Accuracy	Parameters	Notes
Custom CNN	68%	∼1.2M	Trained from scratch
VGG13	71%	∼9M	Pretrained, deeper network
ResNet50	72%	∼23M	Transfer learning applied

V. Challenges Faced

- **Data Quality & Imbalance:** Disgust and Fear had very few examples in both image and audio datasets, causing bias.
- **Hardware Limitations:** GPU and RAM constraints limited training larger models or ensembles.
- **Low Image Resolution:** The 48×48 size limited fine-grained facial feature recognition.
- **Audio Ambiguity:** Subtlety and speaker variability complicated emotional cue extraction.
- **Multimodal Synchronization:** Aligning model predictions and latency was non-trivial; fusion methods had to be simple and robust.

Despite these, we developed competitive standalone models and successfully integrated them.

VI. Conclusion and Future Work

We developed a multimodal emotion detection system combining a custom CNN for facial image analysis and a CNN-based audio classifier. Both models were trained from scratch and designed for lightweight deployment. Although pretrained models outperform ours slightly, our solution provides flexibility, modularity, and clearer interpretability. Future work includes real-time fusion, temporal modeling with RNNs, and expanding datasets to naturalistic emotions.

References

References

[1] I. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *arXiv preprint arXiv:1307.0414*, 2013.
[2] A. Molla Hosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affective Comput.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019, doi: 10.1109/TAFFC.2017.2740923.
[3] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, p. e0196391, 2018, doi: 10.1371/journal.pone.0196391.
[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[6] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022, doi: 10.1109/TAFFC.2020.2981446.
[7] M. Sajjad *et al.*, "A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines," *Alexandria Eng. J.*, vol. 72, pp. 817–840, Jun. 2023, doi: 10.1016/j.aej.2023.01.017.
[8] J. de Lope Asiaín and M. Graña Romay, "An ongoing review of speech emotion recognition," *Neurocomputing*, vol. 528, pp. 1–11, Apr. 2023, doi: 10.1016/j.neucom.2023.01.002.
[9] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
[10] A. V. Geetha, M. Mala, R. Priyadharsini, and N. Nadeem, "Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions," *Inf. Fusion*, vol. 105, p. 102218, May 2024, doi: 10.1016/j.inffus.2023.102218.