Explainable Facial Emotion Recognition Through Hybrid Deep Learning and Attention-Based Feature Extraction

DUDEKULA SHARINTAZ[1], Mr. G VISWANATH[2]
[1]PG Scholar, Department of E.C.E., P.V.K.K. Institute of Technology, Rudrampeta, Alamuru, Anantapur, Andhra Pradesh – 515002, A.P.
[2]Associate Professor, Department of E.C.E., P.V.K.K. Institute of Technology, Rudrampeta, Alamuru, Anantapur, Andhra Pradesh – 515002, A.P.

Abstract: Facial Expression Recognition (FER) has emerged as an important application area of Human-Computer Interaction (HCI) which aims to automatically interpret the state of a person's face. In this paper, a deep learning-based framework which uses Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks is proposed to improve the classification of facial expression. In the first step, cascaded CNN are used to determine facial regions, after which the images are normalized and resized to have the same dimension. CNNs are used to extract spatial information from facial images, while BLSIMs are used to learn temporal information and the dynamic changes in facial expression. Extending the proposed framework, data sets with three basic emotional classes (happy, sad, and neutral) are used to evaluate it. Multiple epochs of training the model to optimize feature learning and classification performance. Experimental observations show the recognition accuracy improvement over the conventional approaches. The proposed approach is an effective method for tackling the task of facial expression analysis by integrating spatial feature extraction with temporal sequence modelling. Applications in healthcare, education, security and real-time affective computing systems.

Key Words: — Facial Expression Recognition, Convolutional Neural Network, Bidirectional Long Short-Term Memory, Deep Learning, Human-Computer Interaction, Emotion Classification.

1.Introduction

Human-Computer Interaction (HCI) is a field of study that uses system characteristics to enhance the effectiveness of humans and computers communicating with one another by responding and reacting to the user's emotions. Facial Expression Recognition (FER) plays an important role in applications such as healthcare, education, virtual assistants, and affective computing. However, it is difficult to accurately recognise emotions because of changes in illumination, head pose, facial occlusion and individual expression patterns. Traditional FER method uses features that were designed manually and uses classifiers based on traditional approach, which may not work well in real scenarios.

In this paper, we propose a hybrid deep learning system, the combination of Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (Bi-

LSTM) networks for facial emotion recognition system. CNNs extract spatial features from face images and Bi-LSTM networks learn temporal dependencies over the evolution of facial expressions. The framework is tested against fig1.1 data set: Japanese Female Facial Expression (JaFFE) and CK+. Normal performance assessment of classification algorithms. The purpose of the proposed approach is to enhance the emotion recognition and enable intelligent emotion-aware application in real world setting.



Fig 1.1: Sample images in JaFFE and CK+ dataset

Facial Expression Recognition (FER) plays an important role in Human-Computer Interaction (HCI) by identifying human emotions from facial images. This paper proposes a hybrid CNN-BiLSTM model, where CNN extracts spatial facial features and BiLSTM captures temporal dependencies. The combined approach improves emotion classification accuracy and supports real-time emotion-aware applications in education, healthcare, Human-Robot Interaction and personalized user interactions.

2.Literature survey

A literature survey of existing Facial Emotion Recognition (FER) techniques has been conducted with the emphasis on CNN, Bi-LSTM, attention, and Transformer approaches. These studies emphasize the significant role of spatial feature extraction and temporal modeling in enhancing the emotion classification accuracy and robustness.

Tian et al. (2023): Proposed a Fine-Tuned Channel–Spatial Attention Transformer for FER, which fused channel and spatial attention to recognize the important facial regions and features. Its objective is to enhance emotion classification performance in the presence of challenging conditions such as occlusion, head-pose variation, as well as illumination variation.

Xiao et al. (2024): Proposed Enhanced Hybrid Vision Transformer (EHVT) model to extract multi-scale features and applied patch-dropping to enhance local facial information and global context. The technique is to increase the accuracy of recognition and decrease unnecessary information in various illumination, pose and occlusion situations.

3.Existing System

Facial detection and tracking are important pre-processing steps in Facial Expression Recognition (FER). Face detection detects and locates human faces in images or video frames and face tracking constantly tracks the movement of faces from one frame to the next. Facial landmark localization can be performed on the face detected to extract key points such as eyes, eyebrows, nose, and mouth, which are crucial for emotion analysis followed by fig 3.1.

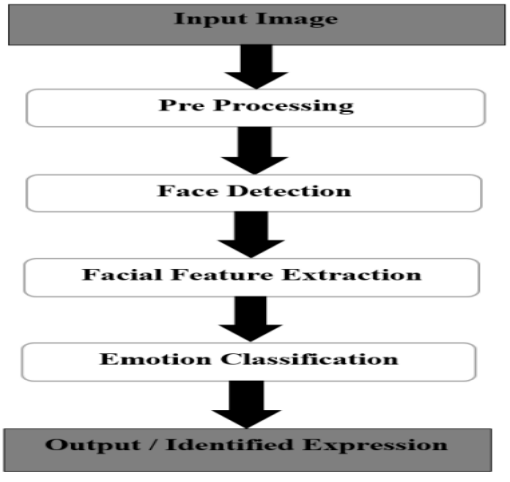


Fig 3.1: Flow Chart of the existing system

Facial shape and relationships between the various facial features can enhance the accuracy of facial landmark localization, as is

done in the case of Active Shape Models (ASM) and regression-based approaches. These isolated facial areas and landmarks aid deep learning in picking up the spatial and temporal characteristics of important areas that are important to emotion classification like happy, sad, and neutral expressions.

Global facial features (F1–F4) are extracted by measuring distances between key facial landmarks, including the eyes, nose, lips, and chin. The local feature (F5) is obtained through grid-based wrinkle analysis using grayscale conversion and Canny edge detection. Combining global and local features improves facial expression representation and emotion classification accuracy.

F1 = (distance from left to right eye ball) / (distance from eye to nose)

F2 = (distance from left to right eye ball) / (distance from eye to lip)

F3 = (distance from eye to nose) / (distance from eye to chin)

F4 = (distance from eye to nose) / (distance from eye to lip)

The existing face recognition and emotion detection system uses face detection, alignment, and landmark localization to extract facial features. For emotion classification, a supervised Support Vector Machine (SVM) classifier is employed to combine and classify geometric, texture, and appearance features. However, the system may face challenges in accurately recognizing emotions under varying illumination, head poses, and partial facial occlusions.

4.Proposed Methodology

The hybrid CNN-BiLSTM model is designed to classify facial expressions in Human-Computer Interaction (HCI) scenario. It combines together spatial and temporal feature learning for better emotion classification. In the beginning, images are pre-processed and then spatial features are extracted from the faces with the help of Convolutional Neural Networks (CNNs). Then, Bidirectional Long Short-Term Memory (Bi-LSTM) networks are used to understand the temporal changes in facial expressions. Finally, the extracted features are passed to a classification layer that classifies emotions, like happiness, sadness and neutral.

A. System Overview

The proposed Facial Emotion Recognition (FER) system is composed of sequential pipeline process. Images of the face are captured and pre-processed for uniformity of input. CNN layers are used to extract the spatial features of the human face that are distinguished by the human face, and Bi-LSTM networks are used to model the temporal dependence between expression sequences. A classification layer is used to classify the emotion in the input and assigns it to a pre-defined category follow block diagram show by fig 4.1.

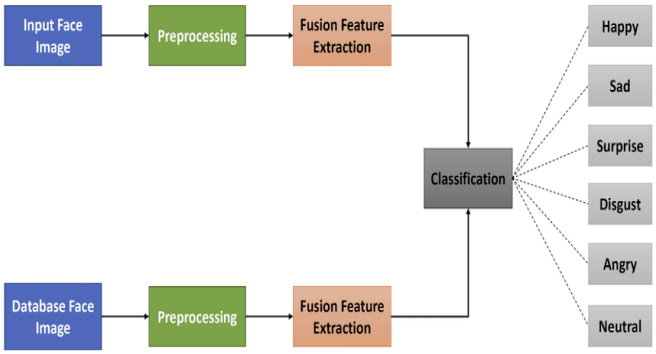


Fig 4.1. Block diagram of the proposed hybrid CNN–BiLSTM

Input Face Acquisition:

- Facial images are captured from the real-time input sources and from a stored facial image database.
- **Pre-processing:**
  The input and database images are pre-processed for face detection, alignment,

resizing, normalization and noise removal to improve image quality.

- **Fusion Feature Extraction:** Modality-aggressive feature representations based on multiple use discriminative feature descriptors (shape, texture, deep features for instance) have been used to create a strong fused feature representation to be discriminative to faces.
- **Classification:** The features obtained through fusion of input image is matched with the database features with the help of a trained classifier to capture the patterns related to different emotions.
- **Emotion Recognition Output:** The system detects the expressions and assigns them to a predefined emotion category (Happy, Sad, Surprise, or Disgust), Angry or Neutral.

B. Dataset Description: The FER database consists of images of faces or video clips which are categorized by seven emotions: happy, sad, angry, fearful, surprised, disgusted and neutral. It contains expression variation, head pose variation and illumination variation.

C. Pre-processing: Facial images are scaled, then grayscale and cropped to make focus the face. Normalization helps to minimize the difference in illumination and pixel-intensity, which can enhance feature extraction and training.

D.CNN-Based Spatial Feature Extraction: CNN layers classify faces based on low level edges & texture to high level emotion related patterns. The spatial features are then fed into the Bi-LSTM network to be further processed and classified into emotions.

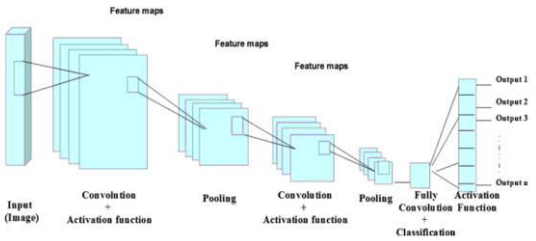


Fig. 4.2: CNN-Based Facial Feature Extraction.

"Let's define the convolution as follows:

$$y(i,j) = \Sigma\,\Sigma\, x(i+m, j+n) \cdot w(m,n) + b \quad \ldots.(1)$$

Here, x is the image that is being convolved, w denotes the kernel, and b denotes the bias. A Rectified Linear Unit activation function is added to give the network nonlinearity:

$$f(x) = \max(0, x) \ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2)$$

A pooling layer is added to decrease the spatial size" of feature maps for fine variations in facial appearance and for robustness. The last convolutional feature map is then flattened into compact feature vectors representing the spatial parameter of the facial expression.

E. BiLSTM-Based Temporal Modelling: CNN is used to extract spatial features from facial images and these features are then passed into a Bidirectional Long short term memory network (BiLSTM) which is used to learn temporal dependencies of facial expression. Shown in fig 4.3 LSTM gates – forget, input and output – manage information flow and are able to remember features. While the emotion recognition task is difficult, BiLSTM can be used to process sequences in both forward and backward directions.

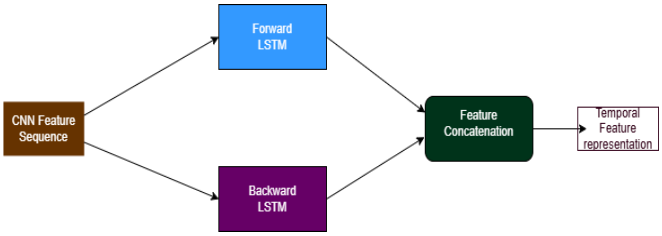


Fig 4.3: Bidirectional LSTM based temporal modelling

Bi-LSTM Output: The forward and backward hidden states are linked together for preserving the contextual information from the sequences of facial expressions.

Training Configuration: The CNN-BiLSTM model is trained to categorize the facial expressions into the pre-defined classes of emotions, and tested with the standard classification metrics.

Evaluation Metrics: The accuracy is used to measure the overall correctness of the prediction, precision is used to measure reliability of the prediction, recall is used to identify actual emotion classes and F1 score is used to balance precision and recall.

5.Results and Discussion

CNN-BiLSTM model evaluation measures are accuracy, precision, recall, and f1 score. The results evaluate the effectiveness of CNN based spatial feature extraction and BiLSTM based temporal modelling technique for facial emotion classification for future use in real time Human Computer Interaction (HCI) systems.

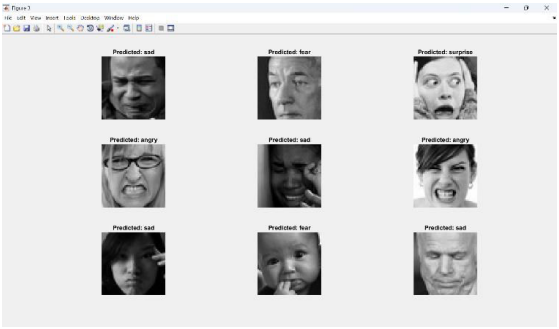


Fig 5.1: Qualitative Performance of Facial Emotion Classification

Fig 5.1 illustrates the facial emotion predictions of the hybrid CNN-BiLSTM model, classifying grayscale images into emotions such as sadness, fear, surprise, and anger. The results demonstrate effective emotion recognition through spatial feature extraction and sequential learning.

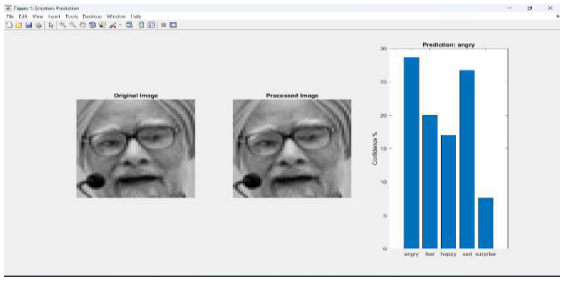


Fig 5.2: Emotion Probability Distribution for Prediction

The fig 5.2 CNN-BiLSTM model successfully detects the extreme emotion "angry", which is displayed in the processed facial image with about 28% probability based on CNN-BiLSTM. The probability distribution further shows that there are similarities with the sad category, which makes it difficult to distinguish visually similar emotions.

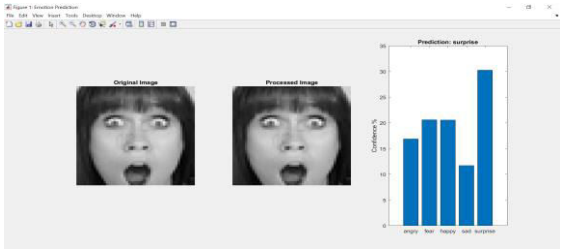


Fig 5.3: Emotion Probability Distribution Prediction

As can be seen in Fig 5.3, CNN-BiLSTM has a higher confidence score, tending to predict the emotion of surprise as the main emotion. The confusion matrix presents the actual performance of a classification model, showing the correctness of the classifications and the errors made in the classification of similar emotions, e.g., neutral.

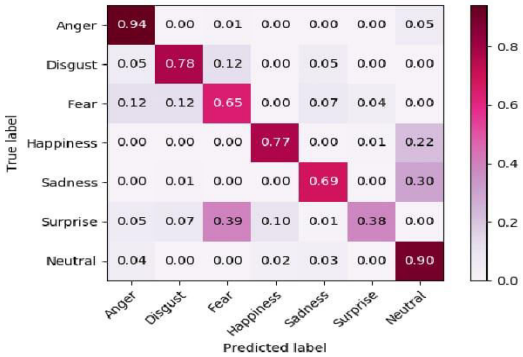


Fig 5.4: Confusion Matrix Facial Emotion Recognition

The CNN-BiLSTM model demonstrates good performance in recognizing facial emotions, achieving its highest accuracy for Label 4. Most misclassifications occur between visually similar emotions, such as neutral and sadness. The hybrid model effectively captures spatial and temporal facial features.

i) Training Performance Analysis

The fig 5.5 training performance shows steady improvement in accuracy and a continuous reduction in loss across epochs, indicating effective model learning. CNN extracts spatial features from facial images, while BiLSTM captures sequential patterns to improve emotion classification.

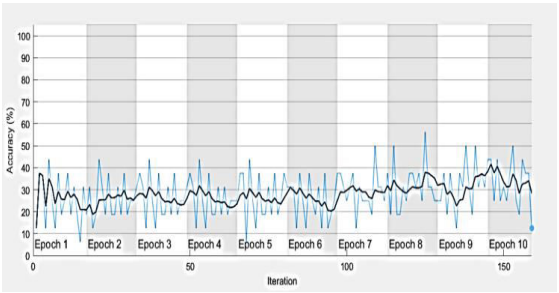


Fig 5.5: Training accuracy variations over epochs

The fig 5.6 training loss shows minor fluctuations and gradually converges over 10 epochs, indicating stable model learning. The CNN extracts spatial facial features, while BiLSTM captures temporal dependencies, improving facial emotion recognition.

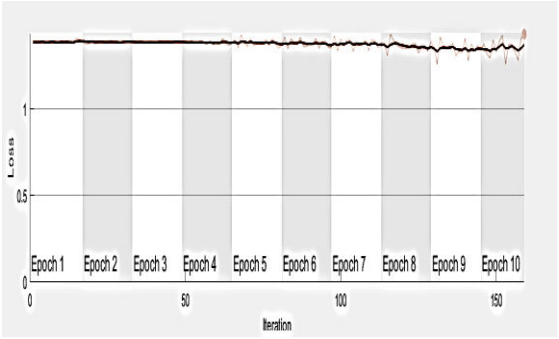


Fig 5.6: Training loss variations over epochs

After 10 epochs, the training loss is relatively stable, showing that the model is generally converging and that there are not significant variations across the epochs.

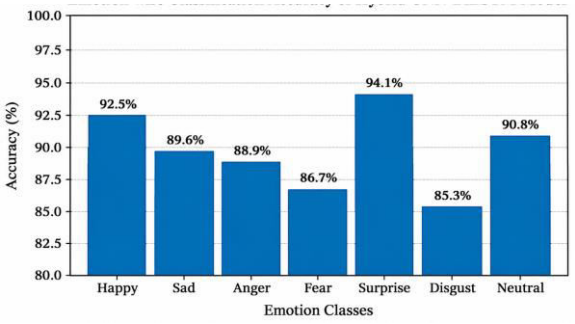


Fig 4.3: Emotion-wise accuracy comparison of CNN-BiLSTM

The CNN-BiLSTM model accurately identifies emotions such as happiness and surprise, while BiLSTM improves the recognition of subtle emotions like fear and disgust.

6.Conclusion and Future Work

This proposed hybrid CNN-BiLSTM model is an effective method to Facial Emotion Recognition (FER) in Human-Computer Interaction (HCI) applications. CNN module learns discriminative features in the spatial domain of facial images and Bi-LSTM module learns the temporal domain of facial expression frames. The framework uses a combination of spatial feature extraction and the temporal sequence modelling to enhance emotion classification of emotions like happiness, sadness, and neutral. The system can enable emotion-aware applications in healthcare, education, virtual assistants and intelligent user interfaces. Despite this, it requires good quality of data, different lighting shades, facial occlusion and computation needs in order to be used in practice. More experimental testing is required to prove the performance results in a wider variety of real-world situations.

Future Work: To gain the global features of the face and boost emotion recognition accuracy, a CNN-BiLSTM model can be combined with Vision Transformers (ViT). Future research may be directed toward real-time recognition, increased tolerance to pose

and lighting changes, and recognition of small-scale emotions.

References

1. A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going Deeper in Facial Expression Recognition Using Deep Neural Networks," IEEE Winter Conference on Applications of Computer Vision (WACV), 2016.
2. Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-Temporal Convolutional Features with Nested LSTM for Facial Expression Recognition," Neurocomputing, vol. 317, pp. 50–57, 2018.
3. "Three Convolutional Neural Network Models for Facial Expression Recognition in the Wild," Neurocomputing, vol. 355, pp. 82–92, 2019.
4. S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195–1215, 2022.
5. Y. Khaireddin and Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013," arXiv preprint, 2021.
6. Y. Ming, H. Qian, and G. Liu, "CNN-LSTM Facial Expression Recognition Method Fused with Two-Layer Attention Mechanism," 2022.
7. J. Zhong, T. Chen, and L. Yi, "Face Expression Recognition Based on NGO-BILSTM Model," Frontiers in Neurorobotics, vol. 17, 2023.
8. C. Tang, D. Zhang, and Q. Tian, "Convolutional Neural Network–Bidirectional Gated Recurrent Unit Facial Expression Recognition Method Fused with Attention Mechanism," Applied Sciences, vol. 13, no. 22, 2023.
9. T. Kopalidis, V. Solachidis, N. Vretos, and P. Daras, "Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets," Information, vol. 15, no. 3, 2024.
10. "Facial Emotion Recognition Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," 2024.
11. "Convolutional Neural Networks-Based Evaluation for the FER-2013 and Revised CK+ Datasets," Proceedings of the International Conference on Electronics and Devices, Computational Science, 2024.
12. "Cross-Database Facial Expression Recognition Based on Multi-feature Representation Multi-layer Domain Adaptive Fusion Network," Engineering Applications of Artificial Intelligence, 2025.
13. "An Interactive Information-Based DCNN-BiLSTM Model with Dual Attention Mechanism for Facial Expression Recognition," 2025.