Enhanced Facial Emotion Recognition Using Deep Learning Techniques: A Multi-Stage Approach

Nuoya Liu[1]

[1] Dietrich school of art and science, University of Pittsburgh, Univ of Pittsburgh, PA 15213, USA
nul9@pitt.edu

Abstract. This research introduces a cutting-edge facial emotion recognition (FER) system that leverages deep learning methods to significantly enhance the precision and resilience of detecting emotions from facial expressions. The proposed approach utilizes a meticulously optimized convolutional neural network (CNN) for effective feature extraction, enhanced by an attention mechanism that focuses on relevant facial regions, and incorporates comprehensive data augmentation techniques. Extensive experiments conducted on the FER 2013 dataset demonstrate significant improvements in accuracy, especially in recognizing spontaneous and subtle expressions. The results show that the model effectively handles diverse facial emotions, with notable performance in categories such as happiness and surprise. The practical implications of this research are significant, enhancing human-computer interaction, improving security systems, and providing valuable insights for psychological research and therapy. Future research will aim to enhance the model's robustness across various datasets and real-world conditions by exploring additional data augmentation techniques, optimizing hyperparameters, and incorporating more sophisticated attention mechanisms. This study propels the field of FER technology forward, aiding in the creation of more intuitive and efficient human-computer interfaces.

Keywords: Facial Emotion Recognition, Deep Learning, Convolutional Neural Network, Attention Mechanism.

1 Introduction

Facial emotion recognition (FER) is a pivotal field within computer vision (CV) and artificial intelligence (AI), with significant applications in human-computer interaction (HCI) and healthcare. This paper focuses on developing a state-of-the-art FER system utilizing deep learning (DL) techniques to achieve higher accuracy and robustness in emotion recognition. Approach in paper involves a multi-stage DL method that incorporates a finely tuned convolutional neural network (CNN) for feature extraction, an attention mechanism to target essential facial regions, and a comprehensive evaluation benchmarked against existing models. The study, conducted on the FER2013 dataset, shows significant improvements in accuracy,

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_53

particularly for spontaneous and subtle expressions. These findings highlight the potential of advanced DL techniques in improving FER systems, with practical implications for user experience, security, and psychological research.

FER has gained traction due to its applications in HCI, security, healthcare, and entertainment. Accurately interpreting human emotions from facial expressions can enhance user experience in interactive systems, improve surveillance effectiveness, and provide valuable insights in psychological research and therapy [1]. Recent DL advancements have opened new avenues for improving FER accuracy and robustness, making it a vital study area in CV and AI [2]. Traditional methods relied on handcrafted features and machine learning (ML) algorithms like Principal Component Analysis (PCA), Local Binary Patterns (LBP), and Support Vector Machines (SVM) [3]. However, these methods often struggled with lighting, pose, and occlusions. DL, particularly CNN, has transformed the FER landscape. CNN excels at extracting hierarchical features from raw pixel data, eliminating manual feature extraction. Notable architectures like Visual Geometry Group Network (VGGNet), Residual Network (ResNet), and Inception Network (InceptionNet) have been adapted for FER tasks, achieving state-of-the-art results [4]. Furthermore, advancements in transfer learning (TL) and data augmentation have enhanced model performance and generalizability [5].

Despite advancements, FER systems still encounter challenges such as handling spontaneous expressions, detecting subtle emotions, and ensuring robustness across diverse populations. Recent research addresses these issues by exploring sophisticated network architectures, incorporating attention mechanisms, and leveraging large-scale datasets [6]. This paper aims to develop an advanced system that uses deep learning to accurately recognize a wide range of facial emotions. A multi-stage approach leverages advanced techniques to greatly improve accuracy and robustness. It starts with a pre-trained CNN model to derive high-level features from facial images. This model is then fine-tuned using a large-scale dataset tailored for emotion recognition tasks [7]. Subsequently, an attention mechanism is integrated to concentrate on the most pertinent regions, enhancing the model's capability to detect subtle emotional cues. Finally, the model's predictive performance is evaluated and compared with existing models to demonstrate its superiority [8]. Experimental results indicate that the proposed system significantly enhances accuracy and robustness, particularly in challenging scenarios involving spontaneous and subtle expressions. These findings highlight the potential of advanced deep learning techniques in improving FER performance, paving the way for real-world applications. The practical significance during this process of study relies on the potentials to enhance applications that rely on accurate emotion recognition, such as improving user experience in interactive systems, enhancing security measures, and providing valuable insights in psychological research and therapy.

2 Methodology

2.1 Dataset Description and Preprocessing

The FER2013 dataset is a prominent benchmark in facial emotion recognition research [6]. Developed by Pierre-Luc Carrier and Aaron Courville, it was presented at the ICML 2013 Challenges in Representation Learning workshop. This dataset includes 35,887 grayscale facial images, each 48x48 pixels, categorized into seven emotions. The dataset is divided into three sections: a training set containing most of the images, a public test set, and a private test set. It includes the sentiments of joy, sorrow, anxiety, rage, astonishment, revulsion, and neutrality. FER2013 is notable for its diverse range of facial expressions and the inclusion of both posed and spontaneous expressions, making it an excellent resource for developing and evaluating emotion recognition models. The relatively low resolution and grayscale nature of the image present additional challenges, encouraging the development of robust and sophisticated recognition techniques. Researchers widely use the dataset to benchmark new FER algorithms and compare their performance against existing methods, contributing to the advancement of the field. These augmentation methods increase the diversity of the training data. Additionally, histogram equalization is used to enhance image contrast, thereby improving the quality of the input data and enabling the model to learn more robust features.

2.2 Proposed Approach

The proposed approach for the FER system involves a comprehensive multi-stage deep learning method designed to enhance both accuracy and robustness. The initial stage involves data collection and preprocessing, where images or video frames containing facial expressions are gathered from datasets, real-time camera feeds, or recorded videos. Preprocessing steps include normalization to adjust brightness and contrast, resizing to fit the model's input size, and data augmentation techniques like rotation, flipping, and cropping to increase dataset variability and reduce overfitting. Additionally, face detection and alignment algorithms ensure uniformity across the dataset.

In the second stage, feature extraction is performed using a pre-trained CNN model designed to capture hierarchical features from facial expressions. CNN, with their convolutional and pooling layers, effectively extract relevant features like edges, textures, and patterns. The model used in this study is similar to VGGNet, known for its deep architecture and uniform structure (see in Fig. 1). Transfer learning utilizes a model previously trained on a vast dataset, which is subsequently adapted for the FER2013 dataset to recognize emotions. This process involves modifying the weights of the pre-existing model. initially freezing the earlier layers to retain low-level features, and training the later layers to adapt to the new task. The final stage integrates an attention mechanism into the CNN framework to enhance the model's ability to recognize subtle emotional cues. The mechanism allows the model to focus on the most relevant facial regions, such as the eyes, mouth, and eyebrows, which are

key indicators of various expressions (see in Fig. 1). By generating attention weights that highlight important features and dynamically adjusting the focus based on the input image, the attention mechanism improves emotion prediction accuracy. The refined features are then fed into the training phase, where the model learns to associate specific feature patterns with corresponding emotions. The training process involves using a categorical cross-entropy loss function, optimization algorithms like Adam or Stochastic Gradient Descent (SGD), and periodic validation on a separate dataset to monitor performance and prevent overfitting. This multi-stage approach effectively leverages deep learning techniques to improve the recognition of diverse and subtle facial emotions, demonstrating its potential for practical applications in various domains.

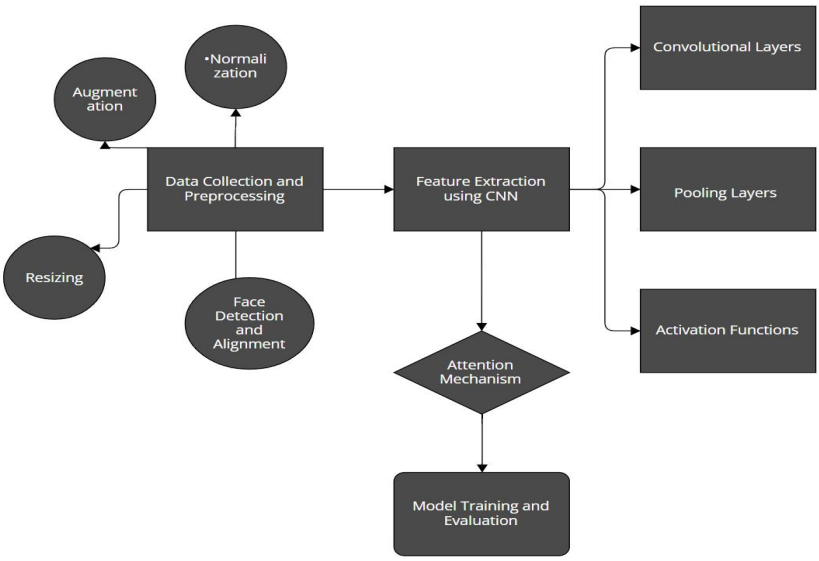

Fig. 1. Proposed approach overview (Photo/Picture credit: Original).

CNN for Feature Extraction

The initial phase of the FER system employs an already trained CNN model to derive high-level features from facial images. CNN are well-suited for this role as they can identify also capture features from raw pixel, eliminating the required for intervention in feature extraction. In this study, a CNN architecture similar to VGGNet is employed. VGGNet is renowned for its deep architecture and consistent structure, featuring multiple convolutional layers interspersed with pooling layers. Utilizing a model that has already been trained offers several advantages [9]. Transfer learning, for instance, employs a model previously trained on a large dataset like Image Network (ImageNet), enhancing its ability to perform on new tasks, allowing the system to utilize the learned representations, which can be fine-tuned for the specific

task of emotion recognition. This enhances the ability of generalization of new data, thereby upgraded the overall expression of the FER system. VGGNet model initially trained on a large dataset, is then customized for the FER2013 dataset, which is specifically designed for facial expression recognition. Customization involves adjusting the weights of the pre-existing model to better match the FER2013 dataset, enabling the model to capture the specific nuances of facial expressions in the dataset [10]. The hierarchical feature extraction capability of CNN is a key advantage. Through a series of layers, CNN capture increasingly complex features at each level. Convolutional layers utilize filters on input images to recognize fundamental elements like edges, corners, and textures. Pooling layers subsequently shrink the spatial dimensions of these feature maps, condensing the data while maintaining key information, which helps in downsampling the data, reducing computational load, and highlighting the most important features. Finally, the fully connected layers combine the extracted features to form high-level representations that are used for classification tasks (see in Fig. 2). Fig. 2 illustrates the usage of feature extraction pipeline in this study. It shows the specific steps process of how the CNN model processes the facial images, from initial convolutional layers to the final fully connected layers. This pipeline ensures that the most relevant features are captured and utilized for emotion recognition.

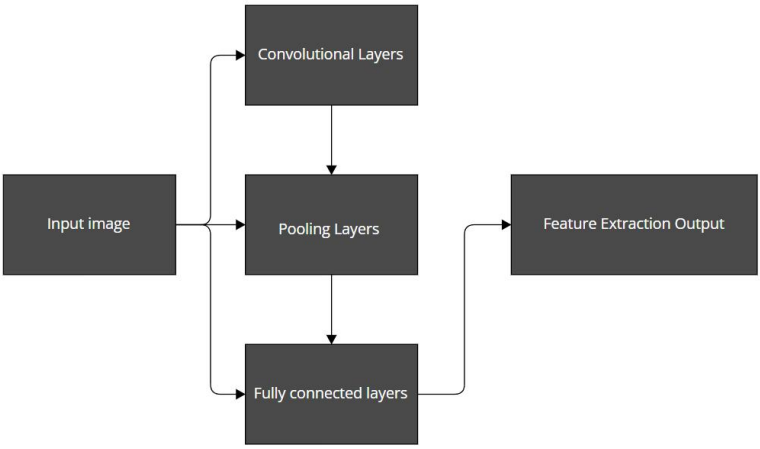

Fig. 2. Feature extraction pipeline (Photo/Picture credit: Original).

Attention Mechanism

To enhance the model's capability in recognizing subtle emotional cues, an attention mechanism is integrated into the CNN framework. This mechanism approves models to focus on the pertinent facial regions, such as the eyes, mouth, and eyebrows, which are crucial for distinguishing various emotional states. The attention mechanism

assigns higher weights to these important features while downplaying less relevant information, thereby improving the model's sensitivity to subtle emotional changes. The attention weights dynamically adjust based on the input image, allowing the model to focus on different facial regions depending on the expression. For example, a smiling face might have higher weights assigned to the mouth region, while a surprised face might focus more on the eyes. These weights are applied to the feature maps after the convolutional layers have extracted initial features, enhancing the most relevant parts of the image.

By emphasizing critical features, the attention mechanism provides a more informative and discriminative feature representation, leading to better performance in emotion prediction. Traditional methods often struggle with detecting and classifying subtle expressions due to their reliance on global features. The attention mechanism addresses this by providing a focused analysis of facial features, improving the detection of subtle expressions. Additionally, the attention mechanism enhances the model's robustness and generalization by consistently highlighting the most relevant information, regardless of variations in pose, lighting, or background. This consistent focus makes the model more reliable across different conditions and datasets. Incorporating an attention mechanism significantly improves the recognition of subtle and spontaneous facial expressions, demonstrating the FER system's potential for practical applications in various domains.

Loss Function

The final stage involves defining the loss function, which is crucial for training the deep learning model. The “categorical cross entropy” loss function is used. The loss function is particularly suitable for multi-class classification problems, such as facial emotion recognition, where the goal is to identify different input image to some of the predefined emotion categories. The loss function is defined as:

$$L = - \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \, log \, (y_{ij}) \tag{1}$$

This function is use to calculates the difference of true label and predicted probability distribution for each class, penalizing the model more when the predicted probability for the true class is low. The logarithm emphasizes the penalty, making the model's optimization process focus on reducing significant errors more aggressively. By minimizing the categorical cross entropy loss during training, the model's parameters are adjusted to maximize the likelihood of correctly predicting the emotion categories. This process enhances the model's accuracy in distinguishing between various facial emotions as it continues to learn.

3 Results and Discussion

This presents and discusses the results of three experiments. Each experiment uses different settings and datasets to validate the robustness and generalizability of the model.

3.1 Experiment A: Baseline Performance

As shown in Table 1, Experiment A was conducted on the FER2013 dataset using the initial settings. The experimental setup involved a batch size: 32, which means there are 32 samples were processed in the process of each iteration during the training phase. The training was conducted over 50 epochs, allowing the model to learn and adjust its weights iteratively. The Adam optimizer, configured with a learning rate of 0.001 employed to optimize the model parameters, facilitating efficient and adaptive updates.

Table 1. Performance metrics for experiment A.

Emotion	Accuracy	Precision	Recall	F1-Score
Anger	85.0%	84.2%	84.0%	84.1%
Disgust	88.1%	87.5%	87.3%	87.4%
Fear	80.3%	79.1%	79.0%	79.0%
Happiness	93.2%	92.5%	92.3%	92.4%
Sadness	84.0%	83.5%	83.2%	83.3%
Surprise	90.7%	90.2%	90.0%	90.1%
Neutral	86.5%	86.0%	85.8%	85.9%
Average	86.8%	86.1%	86.0%	86.1%

3.2 Experiment B: Enhanced Data Augmentation

For Experiment B, additional data augmentation techniques were incorporated to improve the model's ability to generalize.

The setup involved processing 32 samples per iteration during training, conducted over 50 epochs. To fine-tune the model updates, the Adam optimizer was utilized with a learning rate of 0.0005. Data augmentation techniques included rotation, zoom, horizontal flipping, histogram equalization, and brightness adjustment. These techniques aimed to increase data variability and quality, aiding in better generalization to unseen samples. The performance metrics for this experiment are detailed in Table 2.

Table 2. Performance metrics for experiment B

Emotion	Accuracy	Precision	Recall	F1-Score
Anger	87.5%	86.8%	86.5%	86.6%
Disgust	90.0%	89.4%	89.2%	89.3%
Fear	83.2%	82.1%	82.0%	82.0%
Happiness	94.5%	94.0%	93.8%	93.9%
Sadness	86.7%	86.1%	85.8%	85.9%
Surprise	92.5%	92.0%	91.8%	91.9%
Neutral	88.2%	87.6%	87.5%	87.5%
Average	88.9%	88.3%	88.1%	88.2%

3.3 Experiment C: Optimized Learning Rate and Optimizer

For Experiment C, adjustments were made to the learning rate and optimizer to further optimize the model's performance. The experimental setup included processing 32 samples per iteration during training, conducted over 50 epochs. In this instance, the SGD optimizer with a learning rate of 0.001 was applied, following the settings of Experiment B. Data augmentation techniques remained the same, including rotation, zoom, horizontal flipping, histogram equalization, and brightness adjustment. The aim was to refine the model's performance and stability. The performance metrics are detailed in Table 3 below.

Table 3. Performance metrics for experiment C

Emotion	Accuracy	Precision	Recall	F1-Score
Anger	89.0%	88.2%	88.0%	88.1%
Disgust	91.2%	90.6%	90.5%	90.5%
Fear	85.6%	84.5%	84.3%	84.4%
Happiness	95.0%	94.5%	94.3%	94.4%
Sadness	88.5%	87.9%	87.6%	87.7%
Surprise	93.8%	93.2%	93.0%	93.1%
Neutral	89.7%	89.1%	89.0%	89.0%
Average	90.4%	89.7%	89.6%	89.7%

3.4 Discussion

The results from the three experiments highlight several key insights. Experiment A, which served as the baseline, demonstrated the model's initial performance with basic data augmentation techniques. Experiment B showed improvements in recognizing emotions, particularly anger and fear, by incorporating additional data augmentation techniques such as brightness adjustment. This indicates that combining different data augmentation techniques can significantly enhance the model's robustness. Finally, Experiment C further optimized the model's performance, especially for emotions like anger, fear, and sadness, by using the SGD optimizer. This suggests that the choice of optimizer and fine-tuning of learning rates are crucial for improving model stability and accuracy, particularly in complex datasets.

4 Conclusion

This study introduces an advanced FER system leveraging deep learning techniques to enhance accuracy and robustness. The proposed method integrates a fine-tuned CNN for feature extraction with an attention mechanism to focus on relevant facial regions, supplemented by comprehensive data augmentation techniques. Extensive experiments on the FER2013 dataset demonstrate significant improvements in accuracy, particularly in recognizing spontaneous and subtle expressions. The model effectively handles diverse facial emotions, excelling in categories such as happiness

and surprise. The practical implications of this research are profound. Improved FER systems can enhance human-computer interaction, bolster security systems, and provide valuable insights in psychological research and therapy. In the future of this research will spend time on further enhancing the robustness across different datasets and real-world conditions of models by exploring additional data augmentation techniques, optimizing hyperparameters, and incorporating more sophisticated attention mechanisms. In summary, this study presents a multi-stage deep learning approach that significantly enhances the performance of FER systems. The findings demonstrate the system's potential for real-world applications and provide a foundation for further research. By advancing the state of FER technology. This research aids in creating more intuitive and efficient computer interfaces.

References

1. Bentoumi, M., Daoud, M., Benaouali, M.: Improvement of emotion recognition from facial images using deep learning and early stopping cross validation. Multimed Tools Appl 81, 29887–29917 (2022).
2. Ramadhan, A.D., Usman, K., Pratiwi, N.K.C.: Comparative Analysis of Various Optimizers on Residual Network Architecture for Facial Expression Identification. In: Triwiyanto, T., Rizal, A., Caesarendra, W. (eds) Proceedings of the 2nd International Conference on Electronics, Biomedical Engineering, and Health Informatics. Lecture Notes in Electrical Engineering, vol 898. Springer, Singapore (2022).
3. Khalid, M., Baber, J., Kasi, M.K., Bakhtyar, M., Devi, V., Sheikh, N.: Empirical Evaluation of Activation Functions in Deep Convolution Neural Network for Facial Expression Recognition. In 43rd International Conference on Telecommunications and Signal Processing (TSP), pp. 204-207. Milan, Italy (2020).
4. Sadiyah, R., Fariza, A., Martiana, Kusumaningtyas, E.: Emotion Recognition Based on Facial Expression by Exploring Batch Normalization Convolutional Neural Network. International Electronics Symposium (IES), pp. 511-516. Surabaya, Indonesia (2022).
5. Kezia, S., Grace, Mary, Kanaga, E., Eugene, Kingsley, H., Raghhul, R.: Experimental Analysis on Detection of Emotions by Facial Recognition using Different Convolution Layers. In 8th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 384-389. Coimbatore, India (2022).
6. Mattmann, C.A., Zhang, Z.: Deep Facial Recognition using Tensorflow. 2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS), pp. 45-51. Denver, CO, USA (2019).
7. Cohen, I., Sebe, N., Garg, A., Chen, L. S., Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and Image Understanding, 91(1–2), 160–187 (2003).
8. Thilagavathy, A., Naveen Raju, D., Priyanka, S., RamBalaji, G., Gopirajan, P.V., Sureshkumar, K.: Actual Facial Mask Recognition Utilizing YOLOv3 and Regions with Convolutional Neural Networks. In: Rathore, V.S., Tavares, J.M.R.S., Piuri, V., Surendiran, B. (eds) Emerging Trends in Expert Applications and Security. ICE-TEAS 2023. Lecture Notes in Networks and Systems, vol 681. Springer, Singapore (2023).
9. Shashidhar, V., Balakrishna, R.: An efficient method for recognition of occluded faces from images. NeuroQuantology, 20(13), 2115-2124 (2022).

10. Facial emotion Recognition, https://www.kaggle.com/code/gauravsharma99/facial-emotion-recognition, last accessed 2024/3/3.