Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 18, No. 2, 2025

Research on Facial Emotion Recognition Based on Deep Learning

Meilu Wang

Southwest Minzu University, Chengdu 610000, China

Abstract: With the continuous advancement of technology, the demand for intelligent devices is rapidly increasing. Emotion serves as a crucial channel for human information transmission, and facial expressions convey rich emotional cues. These cues not only possess significant research value but also have wide-ranging applications in intelligent systems. Consequently, facial emotion recognition has emerged as a prominent research focus in the field of computer vision. In this study, a facial expression classification model is developed based on the Mini-Xception neural network. The Mini-Xception architecture eliminates the large fully connected layers and adopts depthwise separable convolutions in place of standard convolutions. By reducing the network depth and structural complexity, it effectively lowers the overall computational cost while maintaining performance.

Keywords: Facial Emotion Recognition; Deep Learning; Convolutional Neural Network.

1. Introduction

Research on facial emotion recognition can be traced back to the 1950s and 1960s. With the development of computer technology, people began to attempt using computers to recognize the emotional information embedded in facial expressions. Early methods mainly relied on rule-based designs and manual feature extraction, such as facial geometric structure and color features. With the rise of deep learning, algorithms have been able to automatically extract deep features from facial images, reducing the reliance on handcrafted features. In particular, in 2012, AlexNet made a groundbreaking achievement in the ImageNet competition [1], which drove the widespread application of deep learning in computer vision and brought new opportunities for facial emotion recognition. Researchers gradually adopted Convolutional Neural Networks to process facial images, achieving more efficient and automated emotion classification, significantly improving recognition accuracy and robustness.

In recent years, emotion recognition based on facial expressions has become a highly focused research direction in fields such as psychology, psychiatry, and mental health [2]. Automatic facial emotion recognition plays a crucial role in areas like intelligent living, healthcare systems, emotion diagnosis for Autism Spectrum Disorder (ASD), schizophrenia assessment, Human-Computer Interaction (HCI) [3], Human-Robot Interaction (HRI) [4], and socially welfare systems based on HRI. By accurately recognizing facial emotions, intelligent systems can better understand human emotions, leading to more natural interaction experiences.

2. Facial Emotion Recognition Methods

Contemporary facial emotion recognition techniques can broadly be categorized into two major approaches: traditional feature-engineering-based methods and deep learning-based methods. While both aim to classify emotions from facial expressions, they exhibit substantial differences in terms of feature extraction strategies, model architectures, and overall performance.

Traditional methods predominantly rely on manually designed feature extraction techniques, such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Principal Component Analysis (PCA) [5]. Specifically, LBP captures local texture information by comparing the gray-level intensity of a central pixel with its neighboring pixels and encoding the relationships into binary patterns. PCA, in contrast, serves as a global feature extraction method, utilizing linear transformations to reduce dimensionality while preserving as much critical information from the original data as possible. HOG focuses on edge and structure representation by computing gradient orientation distributions within localized image regions, which are then aggregated into histograms. The extracted features are subsequently classified using machine learning algorithms such as Support Vector Machines (SVM)[6], K-Nearest Neighbors (KNN)[7], or decision trees. The advantages of such methods include relatively simple implementation and low computational cost, making them particularly attractive in resource-constrained environments. However, their performance is highly dependent on the quality of handcrafted feature design, and they tend to exhibit limited robustness to illumination changes, subtle facial variations, and occlusions, thereby constraining their applicability in complex real-world scenarios.

By contrast, deep learning approaches have fundamentally transformed the field by enabling the automatic extraction of hierarchical features directly from raw data, thus obviating the need for extensive manual feature engineering. These models excel at capturing nonlinear relationships between emotional states and are capable of recognizing subtle affective nuances, offering superior generalization and adaptability. Importantly, modern deep neural networks support pre-training on large-scale unlabeled datasets, followed by fine-tuning on task-specific labeled data, which significantly alleviates the dependency on costly human annotations. As a result, deep learning methods have consistently outperformed traditional techniques across a wide range of facial emotion recognition tasks.

Recent advancements have further demonstrated the effectiveness of end-to-end models, particularly

Convolutional Neural Networks (CNNs), for extracting high-level emotional features from facial images [8]. Prominent CNN architectures, such as VGGNet, GoogleNet, and ResNet[9–11], have been widely employed for static facial expression recognition. In the domain of dynamic facial expression analysis, sequential models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks [12], and Transformer-based architectures have been introduced to capture the temporal evolution of expressions. Collectively, these deep learning approaches provide a powerful foundation for advancing the field of facial emotion recognition.

Nevertheless, several open challenges remain. Issues such as imbalanced data distribution, occlusions, illumination variability, and cross-cultural differences in emotional expression continue to hinder robust performance in real-world applications. Addressing these challenges calls for further optimization of model architectures, training paradigms, and data acquisition strategies, with the ultimate goal of achieving more reliable and context-aware emotion recognition systems.

3. Facial Expression Recognition Based on Mini-Xception

The facial expression classification network employed in this study is the Mini-Xception convolutional neural network, whose architecture is shown in Fig. 1. The network is composed of four residual depthwise separable convolution modules, with each convolutional layer followed by Batch Normalization and a ReLU activation function, thereby enhancing both training stability and nonlinear representational capacity. After the feature extraction stage, a Global Average Pooling (GAP) layer is applied to perform dimensionality reduction, and a softmax activation function is subsequently used to generate the final emotion category predictions. This architecture achieves a favorable balance between recognition accuracy and computational efficiency, significantly reducing the number of parameters while remaining suitable for lightweight deployment and real-time facial emotion recognition tasks.

The constructed Mini-Xception network is a simplified and optimized version of the original Xception architecture. By reducing the depth and structural complexity of the network, the overall computational cost is effectively lowered. Compared with traditional convolutional neural networks, Mini-Xception removes the parameter-heavy fully connected layers and replaces standard convolution operations with Depthwise Separable Convolutions. This modification reduces both the number of trainable parameters and computational overhead, thereby accelerating the training process and enhancing the model's generalization ability, particularly in small-sample scenarios and real-world applications.

(1) Data Preprocessing Procedures

The training dataset used in this study is the FER2013 dataset [13], which was originally released by Kaggle in 2013 as part of a facial expression recognition challenge. FER2013 consists of facial images collected from individuals of different age groups, captured under varying viewpoints, and in some cases with partial occlusions, thereby providing substantial diversity and complexity. The dataset contains a total of 35,887 grayscale facial images with a resolution of 48×48, all of which have been uniformly cropped and aligned. The expression categories are divided into seven basic emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The distribution of samples across these categories is summarized in Table 1.

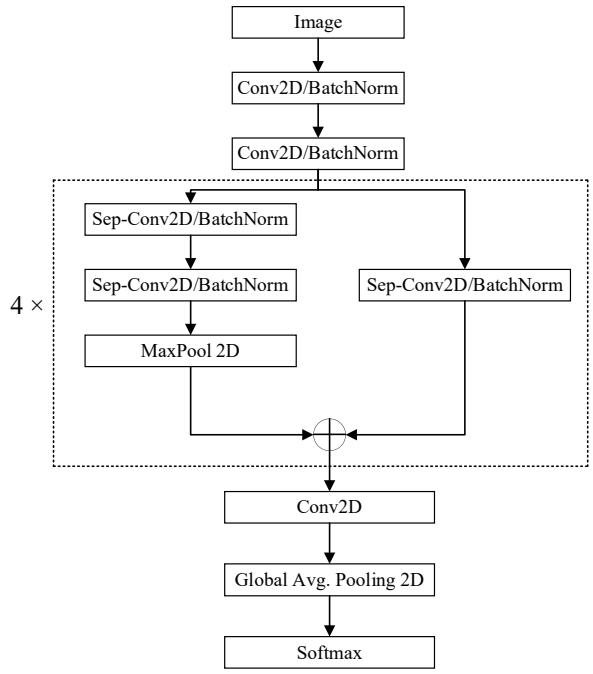

Fig. 1 Schematic diagram of the Mini-Xception network architecture

Table 1. Dataset Distribution by Quantity

Expression	Quantity	Proportion (%)	Numeric Label
angry	4953	14%	0
disgust	547	2%	1
scared	5121	14%	2
happy	8989	25%	3
sad	6077	17%	4
surprised	4002	11%	5
neutral	6198	17%	6

Before model training, the images in the FER2013 dataset are subjected to normalization processing. Normalization is an important preprocessing technique for reducing intra-class feature mismatches. In the context of emotion recognition tasks, normalization helps alleviate intra-class inconsistencies and enhances the model's ability to capture image details. The computation of linear normalization is defined as shown in Equation (1).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Image augmentation is a technique that generates diverse samples by applying a series of random transformations to training images, with the aim of expanding the original dataset and enhancing the generalization ability of the model. By performing random transformations on images during training, the model's over-reliance on specific image attributes (such as position, orientation, and brightness) can be effectively reduced, thereby improving its robustness across different scenarios. For the augmentation of the FER2013 dataset, multiple operations were applied, including random cropping, rotation, scaling, brightness adjustment, and color transformations. Random cropping diversifies the spatial position of faces within the images, which helps reduce the model's sensitivity to the location of the target. In

addition, brightness adjustment and color perturbations were employed to simulate varying environmental conditions, thereby improving the model's adaptability to illumination changes. Since FER2013 images are originally grayscale, the augmentation process preserved their grayscale characteristics, which reduces the model's reliance on color information and directs its focus toward facial texture and structural features.

(2) Construction of the Network Model

Emotion recognition is typically formulated as a multi-class classification problem. The input facial images first pass through convolutional layers to extract local features, followed by pooling layers for dimensionality reduction, which preserves essential information while reducing computational cost. Finally, in the classification stage, the extracted features are integrated by fully connected layers to output the corresponding emotion categories.Previous studies have demonstrated that increasing the number of hidden layers in neural networks can improve classification accuracy; however, this also leads to significantly longer training times and reduced deployment efficiency. To achieve more efficient emotion recognition, the Mini-Xception model constructed in this study introduces structural optimizations and simplifications over the original architecture. The model is trained on preprocessed facial expression images and their corresponding labels to obtain the final expression recognition model.During the prediction phase, preprocessed images (without data augmentation) are fed into the trained model for forward inference. The model output is processed by a softmax activation function, producing a probability vector with seven elements, each corresponding to one of the seven basic facial expression categories. The class with the highest confidence score is ultimately selected as the model's prediction.

4. Experimental Results and Analysis

(1) Training of the Model

In this study, a Mini-Xception convolutional neural network model was constructed using the Keras deep learning framework, with the preprocessed FER2013 facial expression dataset employed as the training data. During the data partitioning process, 20% of the samples from the test set were randomly selected as the validation set, while the remaining samples were used for model training, ensuring no overlap between the training and validation sets to prevent data leakage. The training parameters of the model are summarized in Table 2.

During the model training process, data augmentation was applied to the FER2013 dataset, and the parameter settings of the augmentation operations are summarized in Table 3.

During the model training process, the Adam optimizer [14] was employed to optimize the network parameters. The Adam algorithm combines the ideas of momentum and adaptive learning rates, and is invariant to gradient diagonal rescaling, which makes it particularly suitable for deep neural networks with a large number of parameters and high-dimensional data. Moreover, it requires little manual adjustment of hyperparameters during training, thereby offering strong stability and efficient convergence. The parameter update equations are formulated as follows:

$$\theta_{t+1}=\theta_t-\alpha\cdot\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\varepsilon} \quad (2)$$

Here, θ_t denotes the model parameters at the current time step, α is the learning rate, and $\hat{m}_t$ and $\hat{v}_t$ represent the bias-corrected estimates.$\varepsilon$is a small constant introduced to prevent division by zero. The model adopts the cross-entropy loss function as the objective function for optimization, which is defined as follows:

$$J(\theta)=-\frac{1}{m}\sum_{i=1}^{m} y^i\log\left[\left(h_\theta(x^i)\right)+(1-y^i)\log\left(1-h_\theta(x^i)\right)\right] \quad (3)$$

In this study, the model training was implemented under the Keras framework. To enhance the generalization ability of the model, the ImageDataGenerator() was employed to perform real-time augmentation on each batch of data. By applying random transformations to image samples during training—such as rotation, translation, scaling, and flipping—the diversity of the training data was expanded, thereby alleviating overfitting and effectively improving the model's recognition performance on unseen data.

Table 2. Training Parameter Settings

Parameter	Value
Batch Size	32
Epoch	150
Patience	50
Num Classes	7

Table 3. Parameter Settings for Data Augmentation

Augmentation Type	Parameter
Random Rotation/ °	[-10, 10]
Horizontal Shift	0.1
Vertical Shift	0.1
Horizontal Flip	yes
Scaling / Zoom	0.1

(2) Analysis of Results

Fig.2 illustrates the variation trends of training and validation accuracy for the Mini-Xception model. As the training process progresses, although the training curve exhibits certain fluctuations, the overall trend is upward, with both training and validation accuracy gradually improving. Eventually, both curves stabilize, indicating that the model has converged and achieved satisfactory learning performance.

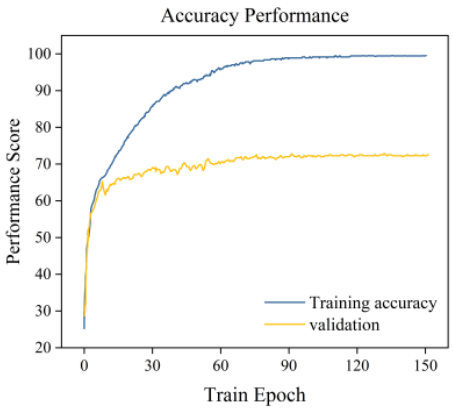

Fig. 2 Training Curves

After the model training was completed, the set of parameters achieving the highest validation accuracy was selected and further evaluated on the validation set. To analyze the classification performance of the model during the testing phase, a confusion matrix was employed, and the results are presented in Fig. 3.

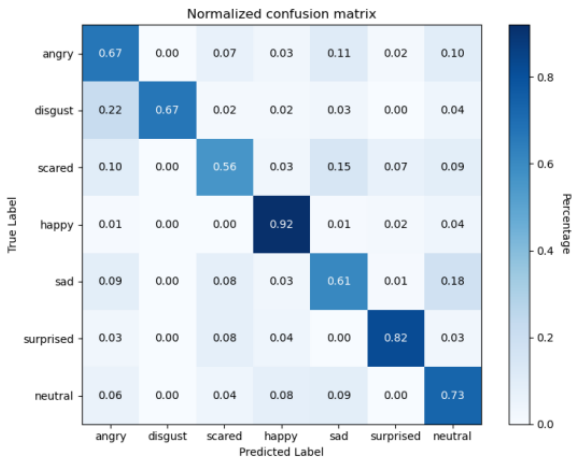

Fig. 3 Confusion Matrix on the Test Set

From the confusion matrix, it can be observed that the Mini-Xception model performs best in recognizing the "Happy" expression, achieving an accuracy of 92%. In contrast, its performance on the "Scared" expression is relatively weaker, with an accuracy of only 56%. This performance gap may be attributed to the imbalanced distribution of samples across different categories in the training set. Furthermore, the confusion matrix reveals a relatively high misclassification rate between "Sad" and "Scared", as well as between "Angry" and "Disgust", indicating that these categories share overlapping visual features, which makes them more challenging to distinguish. Overall, the Mini-Xception model attains an accuracy of 71.12% on the entire test set, demonstrating a reasonably satisfactory recognition performance.

Comparison with Other Algorithms

To validate the effectiveness of the proposed network model, this paper conducts comparative experiments on the FER2013 dataset against several mainstream facial expression recognition networks. The classification accuracies of different methods on this dataset are shown in Table 4. Analysis of the experimental results indicates that the network model constructed in this paper achieves superior recognition accuracy, further demonstrating its effectiveness and advantages in facial expression recognition tasks.

Table 4. Recognition Accuracy of Different Algorithms on the FER2013 Dataset

Method	Accuracy (%)
Zhang Yuqing et al. [15]	68.10
Yan Chunman et al. [16]	68.90
Khemakhemet al. [17]	70.59
Zhouet al. [18]	70.91
Proposed Method	71.12

5. Summary

This paper focuses on facial expression recognition based on deep learning. A Mini-Xception network architecture is constructed, and the model is trained using preprocessed facial expression images and their corresponding labels to obtain the final expression recognition model. Multiple experiments conducted on the publicly available FER-2013 dataset demonstrate that the improved network achieves a relatively high recognition rate. Overall, although convolutional neural networks have become relatively mature in the field of facial expression recognition, challenges remain in capturing subtle variations of expressions and efficiently extracting and leveraging discriminative features. Moreover, the effective integration of key facial region features with both global and local representations is also a crucial factor for improving recognition performance. Future research may further explore directions such as multimodal emotion analysis, micro-expression recognition, and practical applications of the technology, in order to enable expression recognition systems to better adapt to individual differences and diverse population needs.

Acknowledgements

This project is supported by the Innovative Research Project for Graduate Students of Southwest Minzu University (Project No. YCYB2024070).

References

[1] Krizhevsky A, Sutskever I, Hinton G.ImageNet Classification withDeep Convolutional Neural Networks. Advances in neural informationprocessing systems, 2012, 25(2): 84-90

[2] Suchitra, Suja, P. and Tripathi, S., Real-time emotion recognition from facial imagesusing Raspberry Pi II, 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2016:666-670.

[3] Pantic M, Valstar M, Rademaker R and Maat L, Web- Based Database for Facial Expression Analysis, IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands, 2005:317-321.

[4] Gross R,Matthews I,Cohn J,Kanade T and Baker S, Multi-PIE, Proc Int Conf AutomFace Gesture Recognit, 2010., vol. 28, no. 5:807-813.

[5] Zhang T, Liu Y,Ren S, et al, Steganography algorithm of differential histogram shiftbased on LBP facial texture features, Computer Application Research, 2020, vol. 37,no. 6: 1774-1778.

[6] Yu J, Bhanu B. Evolutionary feature synthesis for facial expression recognition[J]. Pattern Recognition Letters, 2006, 27(11): 1289-1298.

[7] Li M, Xu H, Liu X, et al. Emotion recognition from multichannel EEG signals using K-nearestneighbor classification[J]. Technology and health care, 2018, 26(S1): 509-519.

[8] Bhasin A,Mistry A. Convolutional neural networks for problems in transportphenomena: A theoretical minimum. Journal of Flow Visualization and ImageProcessing,2023, 30(3): 1-38.

[9] SIMONYAN K, ZISSERMAN A.Very deep convolution-alnetworks for large - scale image recognition[J].arXiv preprint arXiv,1409. 1556, 2015.

[10] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper withconvolutions[J].arXiv preprint arXiv,1409. 4842, 2014.

[11] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deepresidual learning for image recognition [C] / / Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Piscataway,NJ:IEEE,2016: 770-778

[12] Kim D H, Baddar W J, Jang J, et al. Multi-objective based spatio-temporal featurerepresentation learning robust to expression intensity variations for facial expression recognition[J].IEEE Transactions on Affective Computing, 2017, 10(2): 223-236.

[13] Goodfellow I J, Erhan D, Luc Carrier P, et al. Challenges in representation learning: a report on three machine learning contests[J]. Neural Networks, 2015(64): 59-63.

[14] Kingma D P, Ba J L. Adam: A method for stochastic optimization[J]. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015: 1-15.

[15] Y. Zhang, N. He, and R. Wei, "Facial expression recognition based on convolutional neural network fused with SIFT features," Computer Applications and Software, vol. 36, no. 11, pp. 161–167, 2019.

[16] C. Yan, X. Zhang, and Q. Wang, "Facial expression recognition based on improved MobileNetV2," Computer Engineering and Science, vol. 45, no. 6, pp. 1071–1078, 2023.

[17] KHEMAKHEM F,LTIFI H.Facial expression recognition using convolution neural network enhancing with pre-processing stages[C]//Proceedings of 2019 IEEE/ ACS16th International Conference on Computer Systems and Applications (AICCSA). Piscataway, NJ: IEEE,2019: 1-7.

[18] ZHOU Jiancan,JIA Xi,SHEN Linlin,et al.Improved softmax loss for deep learning based face and expression recognition[J].Cognitive Computation and Systems, 2019, 1(4):97-102.