# Facial expression recognition based on ResNet and transfer learning

**Yixuan Chen**

Xi'an Jiaotong University, No.28, Xianning West Road, Xi'an, Shaanxi, 710049, P.R. China

zj1561894328@stu, xjtu.edu.cn

**Abstract.** With its potential to revolutionize a wide range of applications, including lie detection, social robotics, and driver fatigue detection, facial expression recognition is a field that is rapidly expanding. However, traditional machine learning methods have struggled with facial expression recognition due to limitations such as manual feature selection and limited representation capabilities. Additionally, these methods require large amounts of annotated data, which can be time-consuming and expensive to obtain. In order to overcome these difficulties, this paper suggests a novel method that builds recognition models using a multi-layer perceptron (MLP) and ResNet. This hybrid model offers improved performance over conventional CNN models, achieving an impressive accuracy rate of 85.71% on the FER_2013 dataset. Additionally, migration learning is used to increase the model's precision and avoid over-fitting. The FER_2013 dataset is used to train and test the model. The results of the trials show that the suggested model can recognize facial expressions while minimizing the overfitting problem typically associated with deep learning. The model will eventually include a self-attentive mechanism in the study in an effort to improve model resolution. By using it with color images, the team also hopes to increase the model's capacity for generalization.

**Keywords:** Facial expression recognition, ResNet, Deep learning, Transfer learning.

## 1. Introduction

A person's internal emotional states, intentions, or social interactions are visually expressed through changes in their facial features, which are referred to as facial expressions [1]. Facial expressions are among the most potent, common, and universal human signals for conveying emotional states and intentions, according to research [2,3]. As a result, the automatic recognition of facial expressions by computers has a wide range of applications, including the development of social robots and the detection of lies and driver fatigue [4]. The study of facial expression recognition has grown significantly over time and has been actively pursued.

By observing the movement of 20 distinct spots in an image sequence, Suwa et al. conducted a preliminary attempt to mechanically evaluate face expressions in 1978 [5]. Since then, a number of techniques for facial recognition utilizing conventional machine learning algorithms have been developed. These techniques, like the work of Suwa, require that the characteristics of images be manually defined and extracted beforehand. The features selected for facial classification were based solely on the expertise and perception of the researchers. The form of the eyebrows and the angle of the mouth's corners were among the characteristics they picked because they thought they were important. Following feature extraction, these characteristics are added to the KNN, SVM, and other conventional machine learning algorithm models as the foundation for model training and expression classification. For example, Bartlett et al. used PCA to extract facial expression features and SVM for classification in a study in 2003 [6]. Moghaddam et al. used LDA to extract facial expression features and K-nearest neighbor (KNN) for classification [7]. Lucey et al. used Gaussian mixture models (GMM) to model the probability distribution of facial expressions and used it for classification [8].

These traditional methods achieve good accuracy on a limited data set, they have some serious drawbacks. Traditional machine learning methods require manual selection of facial expression features, which is often subjective and may not be optimal, resulting in poor recognition performance [9]. This also makes it difficult to handle complex or subtle emotional expressions and to deal with variations in lighting, angle, and expression intensity. The subjectively selected features by researchers also lead to the poor generalization ability of the model. Secondly, traditional machine learning methods usually use linear or simple nonlinear models to represent facial expressions. The representation ability of such models is limited, which restricts the recognition and description ability of complex facial expressions. Traditional machine learning methods typically demand a significant amount of manually annotated data to train the model, thereby incurring both time and financial costs. Moreover, the quality of the annotated data can also impact the recognition performance [9].

In recent years, deep learning has been constantly developing, and with the rapid increase in computer computing power, it is no longer the main bottleneck for model training. Based on these two reasons, the application of deep learning in facial expression recognition has begun to rise [10]. Take the classic Convolutional Neural Network (CNN) structure in deep learning as an example. CNN can automatically learn features in images without the need for manually defining rules or extracting features. This enables CNN to more accurately capture subtle differences and variations in facial expressions. With proper structural design and training, CNN can be robust to lighting, posture, and other factors that may affect facial expression recognition. This can help improve the performance of the model in real-world scenarios, especially in less-than-ideal conditions. But CNN still has its flaws. The main disadvantage is that, in order to prevent overfitting, deep learning networks like CNN need a lot of training data. The training of traditional network architectures cannot be supported by the amount of the existing expression database [4]. To overcome the model limitations mentioned above in facial expression recognition, this paper introduces the deep residual network ResNet and Multi-layer Perceptron (MLP)to construct a recognition model. ResNet has a unique residual connection structure that solves the problem of gradient vanishing and explosion in deep neural networks. In this study, transfer learning is used to overcome the absence of a substantial facial expression dataset to assist training. On the FER_2013 [9] dataset, the model ultimately achieves an accuracy of over 85% while exhibiting little gradient vanishing. According to the experimental results, the suggested model can effectively recognize facial emotions by reducing the overfitting of the depth model.

## 2. Methodology

### *2.1. Dataset description and preprocessing*

FER_2013 is a facial expression dataset with grayscale facial images and is used to train and test the model [9]. 48x48 pixels are the resolution of each image. Automatic alignment has been used to position the faces such that they are nearly centered and take up a similar amount of space in each image. Using the emotional expressions on each face, one of seven categories is to be assigned to each image. The seven categories are surprise, disgust, anger, sadness, happiness, fear, and nature. The dataset has two parts, the dataset of training and the dataset of testing. The dataset of training consists of 28,709 samples. The dataset of testing contains 3,589 examples [9]. Figure 1 shows a portion of the Fer_2013 dataset.

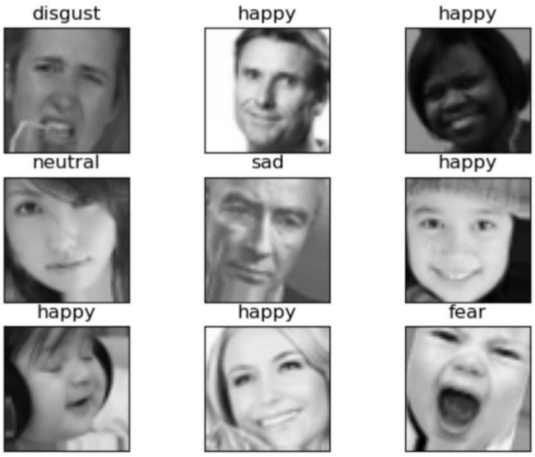

**Figure 1.** The example of FER_2013 dataset.

A series of preprocessing methods are used to enhance the dataset in the project. An 8:2 ratio is used to divide the dataset of training for FER_2013 into 2 parts. One part is used to train the model and the other part is used to verify the model during training. The project applies pixel normalization, random image translation, and random image rotation to the samples in the training set. Pixel normalization is used to reduce the amount of computation required to make the training process stable and efficient. Random image rotation and translation can increase the diversity and richness of the dataset. On the validation set, the project just applies pixel normalization and uses raw images on the testing set.

### 2.2. *Proposed approach*

In this paper, ResNet50 is introduced to construct a facial expression classification model. Transfer learning is introduced to solve the data scarcity problem to accelerate model training. Based on research and experience, the ResNet structure has high flexibility in facial expression identification, hence this study provides a facial expression classification model based on ResNet50. Next, this section will explain the structure of the model used in the paper. As shown in Figure 2, the input image is first fed into the ResNet50 network for feature extraction. The ResNet50 network's input size is 48*48 with 3 channels, matching the size of the images in the dataset. After the image undergoes feature extraction in ResNet50, it becomes a stacked feature with a size of 2*2 and 2048 channels. After using a dropout layer to randomly invalidate some features, the stacked features are input into a flattened layer to be unfolded, resulting in a feature vector with a length of 8192. This feature vector is then batch normalized and inputs into a fully connected layer for further feature extraction, resulting in a feature with a length of 32. The extracted features are then input into a multi-layer perceptron, which outputs a vector of length 7. The position of the maximum value in the vector represents the classification result of the expression.

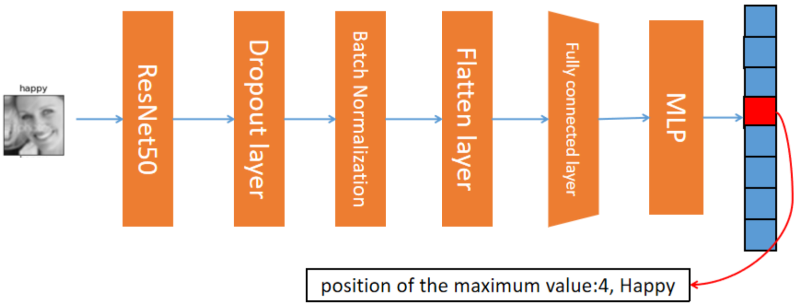

**Figure 2.** The structure of the mode.

*2.2.1. Resnet and transfer learning.* The deep neural network ResNet's introduction of the residual block is its most significant contribution. The residual block allows information to be directly networked by connecting across layers. The network depth is increased while preventing overfitting.ResNet is a deep neural network model. It is most important contribution is the introduction of the concept of residual connections, which allows information to be directly passed through the network by connecting across layers, thereby enabling the possibility of training very deep networks. The fundamental component of ResNet, as depicted in Figure 3, is the residual block, which has two or three convolutional layers and a residual connection that crosses these layers. The input and output of the residual block maintain the same dimension, allowing for cross-layer connections. During training, the residual connection allows the gradient to be passed directly from later layers to earlier layers, making training easier. The model in the paper uses the ResNet50 network. It is a kind of ResNet model containing 50 layers of convolutional networks. The number of parameters in this model is 23587712.

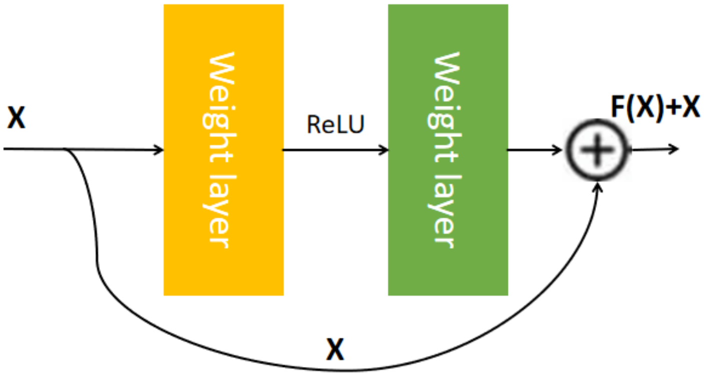

**Figure 3.** The schematic diagram of residual block.

Transfer learning is a machine learning method that uses a previously learned model to train a new model. Transfer learning can produce good results even with little data since the pre-trained model has a strong generalization ability. Because the facial expression classification dataset is too small and easy to cause overfitting, this study uses the ResNet50 model that was previously trained on the ImageNet dataset as the transfer learning model. ResNet50 can only have the parameters of last four layers trained. This gets around the issue of tiny datasets, significantly raises the model's classification accuracy, and prevents overfitting.

*2.2.2. Multi-layer perceptron.* A feedforward neural network made up of numerous completely connected layers with nonlinear activation functions is called a multi-layer perceptron (MLP). MLP stacks multiple fully connected layers to learn higher-order features. The flexibility of MLP is that it can be used to process different types of data and solve various tasks. The output layer of the MLP commonly employs the softmax function in a classification job to transform the output into a probability distribution for the class. The identity function or the sigmoid function is often used by the output layer in a regression job to convert the output to a real value or range of real values. Additionally, MLP may include sigmoid, ReLU, tanh, and other activation functions to provide nonlinearity and enhance the model's capacity for feature representation. The MLP model employed in this study has three completely linked layers, as seen in Figure 4. MLP first performs batch normalization of 32-dimensional feature vectors (obtained by image feature extraction). The normalized data is then fed into the ReLU activation layer. Next, the result inputs the dropout layer and the fully connected layer. The size of the feature vector is unaffected by the completely linked layer. After a series of operations, the length of the eigenvector is still 32. Then iterate this series of operations again and get a 32-dimensional vector. The resulting vector is batch normalized, then fed into the ReLU activation layer, then the results are fed into the Dropout layer, and finally into the full connection layer, resulting in a vector of length 7. This 7-dimensional vector is the result of the MLP, and the magnitude of each element of the vector is the probability of its corresponding class.

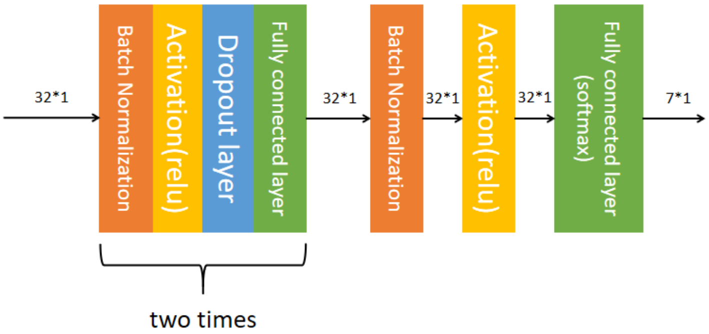

**Figure 4.** The multi-layer perceptron structure.

*2.2.3. Loss function.* Cross-entropy is used as the loss function in this model. In multi-classification situations, cross-entropy is frequently employed to quantify the difference between two probability distributions. Cross-entropy calculates the gap between the actual and anticipated probability distributions and returns a metric value as follows:

$$L = \frac{1}{N}\sum_i \sum_{c=1}^{M} y_{ic} \log p_{ic} \tag{1}$$

where N means sample size, M means Number of categories, $y_{ic}$ means types of samples(if the true class of sample i equals c, the sign function $y_{ic}$ returns 1; otherwise, it returns 0), $p_{ic}$ means the likelihood that sample i from the observed data belongs to class c.

### *2.3. Implementation details*

The model is compiled using Python 3.10. The model is trained on a GPU with 8GB of VRAM. The model is tested on the testing set after 50 training epochs. Other hyperparameter configurations are presented in Table 1.

**Table 1.** Hyperparameter setting listtitle.

| Hyperparameter | Value |
|---|---|
| *width_shift_range(image preprocessing)* | 0.1 |
| *height_shift_range(image preprocessing)* | 0.1 |
| *rotation_range(image preprocessing)* | 15 |
| *batch size* | 64 |
| *dropout rate* | 0.5 |
| *optimizer* | Adam |

## 3. Result and discussion

The results of model testing and training are presented in this chapter along with an analysis. In this study, the model goes through 50 iterations of training using the training and validation sets. Figure 5 shows that as training progresses, the loss of the model on the training set rapidly decreases. Although the model's loss on the validation set varies greatly, it typically shows a downward trend.

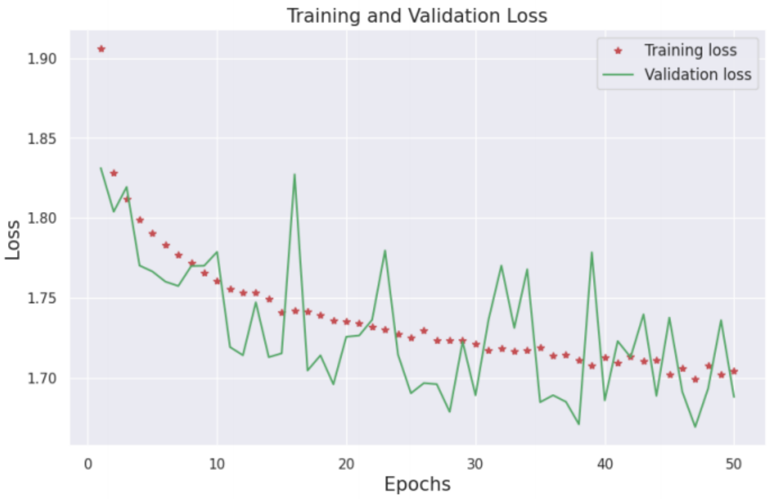

**Figure 5.** The loss of the model during training and validation.

Figure 6 illustrates that the model's training accuracy quickly rose to a high level in the early training stages, indicating the model is learning the patterns in the data well. The validation accuracy remains above 85%, signaling good generalization performance. The small gap between the training and validation accuracies suggests the model did not significantly overfit the training data. In summary, the accuracy results demonstrate the model attained good performance while avoiding overfitting.

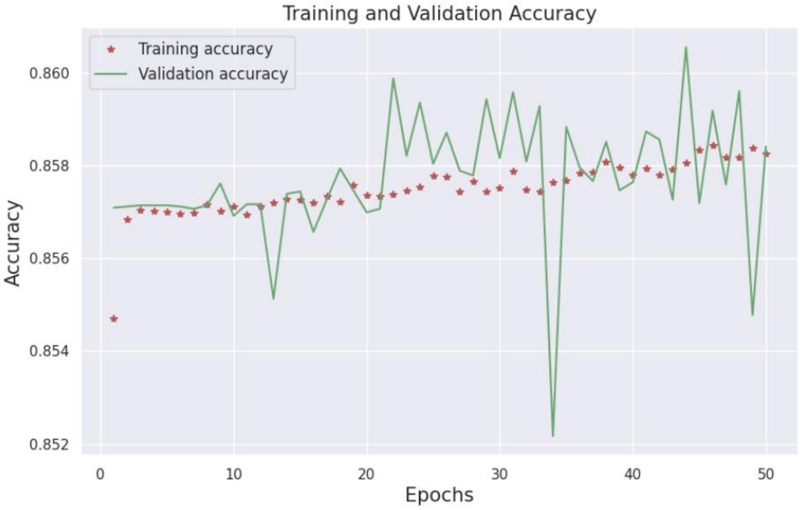

**Figure 6.** The accuracy of model during training and validation.

As shown in Table 2, the conventional CNN model shows overfitting. Not only is the traditional model more accurate than the evaluation and test on the training set, but the loss is also lower. In contrast, the model constructed in this paper avoids overfitting to achieve higher accuracy. The conventional CNN network model has clearly overfitted itself when the outcomes of the model are contrasted with those of the latter. The accuracy of the traditional CNN model on the training set is significantly higher than that on the other datasets, and the loss is significantly lower than that on the other datasets. In contrast, the model created in this study not only avoided overfitting but also attained high accuracy.

**Table 2.** The contrast between traditional CNN model and the model of this paper.

| Model | Training Loss | Training Accuracy | Validation Loss | Validation Accuracy | Testing Loss | Testing Accuracy |
|---|---|---|---|---|---|---|
| Traditional CNN model | 0.13 | 0.9613 | 2.11 | 0.6314 | 1.98 | 0.6420 |
| The model of this paper | 1.71 | 0.8580 | 1.68 | 0.8581 | 1.78 | 0.8571 |

Numerous factors contribute to the great accuracy of the face expression recognition model in this study. First and foremost, ResNet is a good architecture for this job since it can classify facial expressions accurately by adjusting to the underlying characteristics. Second, applying transfer learning to a model that has already been trained on a sizable dataset improves model performance and minimizes overfitting. This strategy enables the model to rapidly increase accuracy in the early phases of training, suggesting that fewer training epochs are required to get effective outcomes. The appropriateness of ResNet, transfer learning, and faster convergence brought on by transfer learning all work together to make the model successful at identifying facial expressions.

## 4. Conclusion

This paper has presented a novel method that combines a multi-layer perceptron with ResNet to create a recognition model in order to address the issues that are frequently associated with having a small dataset for facial expression detection. The purpose of this study is to enhance the accuracy of the model as well as reduce overfitting during the training process. To achieve these objectives, transfer learning is used, which enables the model to learn more effectively from the available data. The experiments are conducted on the Fer_2013 dataset. It achieves an accuracy rate of 85.71% over the traditional CNN model. Overfitting is avoided entirely during the training process. In future research, the study aims to further enhance the model's performance by incorporating a self-attentive mechanism into the model. This is expected to improve the model's ability to recognize facial expressions accurately and efficiently. Additionally, the researchers aim to apply the model to color images to improve its generalization abilities and ensure that it can work effectively across different types of images. Overall, the proposed approach represents a significant contribution to the field of facial expression recognition. These findings have important implications for the development of more effective and efficient recognition models in the future.

## References

[1] Tian Y Kanade T 2011 Facial Expression Recognition Handbook of Face Recognition Springer London 0487-0519.
[2] Darwin C Prodger P 1998 The Expression of the Emotions in Man and Anzimals Oxford University Press.
[3] Tian Y I Kanade T Cohn J F 2001 Recognizing action units for facial expression analysis IEEE Transactions on pattern analysis and machine intelligence 23(2): 0097-0115.
[4] Li S Deng W 2022 Deep Facial Expression Recognition: A Survey IEEE Transactions on Affective Computing 13(3): 1195-1215.
[5] Suwa M Sugie N Fujimora K 1978 A preliminary note on pattern recognition of human emotional expression International Joint Conference on Pattern Recognition: 0408–0410.
[6] Bartlett M S Littlewort G Fasel I Movellan J R 2003 Real time face detection and facial expression recognition: Development and applications to human computer interaction 2003 Conference on computer vision and pattern recognition workshop IEEE 5(1): 0053-0053.
[7] Moghaddam B Pentland A 1995 Probabilistic visual learning for object detection Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition:0786-0793.
[8] Lucey P Cohn J F Kanade T Saragih J Ambadar Z Matthews I 2010 The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops:0094-0101.

[9] Goodfellow I J Erhan D Carrier P L Courville A Mirza M Hamner B Bengio Y 2013 Challenges in representation learning: A report on three machine learning contests Neural Information Processing: 20th International Conference (ICONIP) Springer berlin heidelberg 0117-0124.

[10] Ge H Zhu Z Dai Y 2022 Facial expression recognition based on deep learning Computer Methods and Programs in Biomedicine 0215-0215.