ISSN:3007-9640 DOI: https://doi.org/10.70114/acmsr.2025.2.1.P332

Research on Facial Expression Recognition Based on The Combination of Improved LBP and CNN

Qingqing Wang [1, a *]

[1] Animation Art College, Jilin Animation Institute, Changchun, China.

[a] 286769533@qq.com

Abstract. This paper proposes an In view of the problem of insufficient effectiveness of facial expression feature extraction in complex environments, this paper uses CNN to extract global abstract features. Circular LBP is adopted as the basic operator, which fully considers the interaction between the central pixel and the surrounding pixels. Finally, the features extracted by LBP and the features extracted by the deep network are fused in a weighted manner to obtain the final result. The accuracy rate on the Fer2013 data set reaches 74.21%, which has a significant improvement in improving the recognition accuracy compared with other network models.

Keywords: Convolutional neural network; LBP feature; Feature fusion

1. Introduction

With the advancement of machine vision and artificial intelligence technologies, the development of facial expression recognition technology has been further propelled. Facial expressions encompass complex emotional information. Research indicates that human emotions are constituted by facial expressions, language, and the sounds emitted, among which facial expressions are the most prominent form of emotional expression. Nevertheless, due to multiple factors such as lighting, ethnicity, and angle, there are still considerable challenges in feature extraction in face expression recognition at present.

In 2021, Mohan et al. [1] proposed the FER-net network based on the convolutional neural network. This network consists of several relatively simple convolutional layers, mainly focusing on the local features of the eyes and corners of the mouth of the human face. In 2022, Filali et al. [2] proposed a hybrid architecture based on convolutional neural network and Stacked AutoEncoder, which combines the two feature vectors generated by the network and provides the generated vectors to a meaningful neural network. The latter assigns a group of neurons to each component of the vector to learn each feature, and combines them all together in the last layer for the research work of face expression recognition. In 2021, Wu et al. [3] proposed the FaceCaps network model on the basis of CNN, which uses the face expression features extracted from FaceNet to guide the 3D convolutional dynamic routing capsule network. In 2021, Sum et al. [4] proposed a multi-attention shallow and deep model for extracting shallow and deep features of the human face. Firstly, the network uses the attention mechanism model to extract the shallow features of the human face. Secondly, it uses the convolutional neural network to extract the deep features of the human face. Finally, the shallow and deep features are fused for face expression classification.

In this paper, aiming at the problem of insufficient effectiveness of facial expression feature extraction in complex environments, circular LBP is adopted as the basic operator, and equivalent patterns are used to improve the statistical nature of expression features. Then, the stable features of the image are extracted through rotation-invariant operations. Finally, the features extracted by LBP are weighted and fused with the features extracted by CNN to obtain the final result.

2. Research Algorithm of Facial Expression Recognition Based on Improved Convolutional Neural Network

ISSN:3007-9640 DOI: https://doi.org/10.70114/acmsr.2025.2.1.P332

2.1 Section Headings

Convolutional Neural Network (CNN) is an important deep learning model. It consists of an input layer, a convolutional layer, a pooling layer, and a fully connected layer, which can effectively process complex information [4]. CNN can segment images through convolutional layers and extract image features from multiple segmented images to complete the entire training. The classic convolutional neural network structure is shown in Fig. 1.

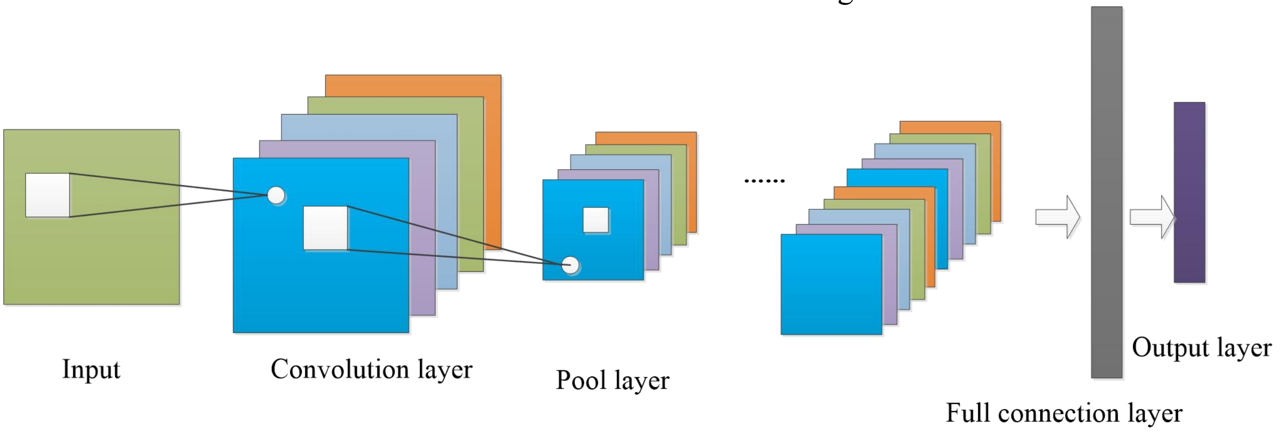

Fig. 1 Schematic diagram of convolutional neural network

(1) Input layer

The main task of the input layer is to input information. In order to make CNN training more effective, data cleaning is generally done in advance, such as null value filtering, normalization processing, etc.

(2) Convolutional Layer

Convolutional layer is the core structure of convolutional neural networks. Perform convolution operation on a fixed size moving window and enclosed area to extract the features of each area. The convolution operation is shown in equation (1):

$$y^j = f(W^j \otimes X + c^j) \tag{1}$$

Among them, X represents the input of the convolutional layer. W^j represents the convolution kernel, with a commonly used size of 3×3, 4×4, 5×5. $\otimes$ represents convolution operation, which is to perform product addition and sum operation between each position of the convolution kernel and the corresponding position of the input to obtain the final convolution value. $f(\bullet)$ is the activation function. After the operation of the activation function, the feature matrix under the action of the convolution kernel is obtained. c^j is the bias constant. The operation process of the convolutional layer is shown in Fig. 2.

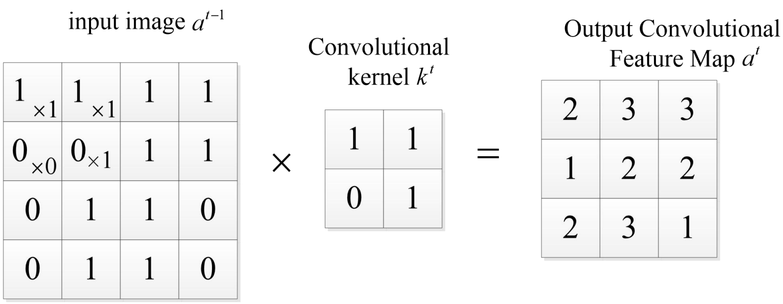

Fig. 2 Convolutional layer operation process

(3) Pooling layer

Usually, convolutional layers use multiple convolutional kernels for feature extraction, obtaining a large number of feature matrices. If fully connected layers are directly connected later, it will result in a large computational load. To avoid this situation, pooling layers are usually added after

ISSN:3007-9640 DOI: https://doi.org/10.70114/acmsr.2025.2.1.P332
the convolutional layer. Pooling processing can have a dimensionality reduction effect and effectively prevent overfitting caused by too many convolutional kernels

(4) Fully connected layer

The fully connected layer is equivalent to a classifier in a convolutional neural network. The original data has undergone convolution and pooling in order to amplify and extract effective features. The feature vectors that enter the fully connected state are subjected to non-linear output after being fully connected, thus completing the recognition task.

(5) Output layer

The output layer is the final layer of a convolutional neural network, which outputs the predicted values in the input image or data. If the model predicts a classification problem, *softmax* can be used for logistic regression for classification prediction. If it is an image recognition problem, the result can be the specific coordinate value of the recognized object in the image.

2.2 LBP Features

The Local Binary Pattern (LBP) is an algorithm for extracting local texture features, which has a relatively good improvement effect in the case of weak light or uneven illumination. In the traditional LBP algorithm, a square is used as the neighborhood, with the central point pixel as the standard. After sliding and comparing with the surrounding pixels, a binary processing is performed, and the LBP code is generated clockwise as the output value of this point pixel.

Although the traditional LBP algorithm has light invariance and rotational invariance, and has shown excellent performance in solving the problem of facial feature extraction under complex lighting conditions, there are still many problems. Firstly, the traditional LBP adopts the comparison of a single pixel, and the description of the gray-scale change in this direction is not comprehensive, and the extracted features are extremely susceptible to interference; secondly, it does not consider the interaction between the gray-scale values of other pixels in the neighborhood and the gray-scale value of the pixel at the center of the neighborhood, and it is very likely to ignore some important local feature information, thereby leading to errors in recognition accuracy and final classification judgment.

This paper introduces a threshold *M* to replace the initial value of the neighborhood center pixel. By calculating the difference between the center pixel and the gray values of its surrounding pixels, and performing a weighted summation of the absolute values, and then taking the average as the threshold *M*, not only the role of the center pixel is considered, but also the relationship between the center pixel and the surrounding pixels is taken into account. The calculation formula of the *M* threshold is as follows:

$$M = \frac{1}{P}\sum_{i=0}^{P-1}\left|g_c - g_i\right| \tag{2}$$

Where *P* is the number of pixel points in the image neighborhood.

In order to enhance the connection between adjacent pixels and the center pixel, it is necessary to recalculate the pixel values in the neighborhood. The calculation formula of the neighborhood pixel points is as follows:

$$N_i = \left|g_c - g_i\right| \quad (i = 0,1...,P-1) \tag{3}$$

$$K_i = \frac{N_{i-1} + N_i + N_{i+1} + M}{4} \quad (i = 0,1...,P-1) \tag{4}$$

In the formula, N_i is the absolute value of the difference between the gray value of each neighboring pixel and the center point pixel, and then the calculated N_i value is summed with the gray value before and after it and the gray value of the center point; K_i is the new gray value of this point pixel.

The calculated neighboring pixels are compared with the threshold *M*. If the value is greater than *M*, it is recorded as *1*; otherwise, it is recorded as *0*. Then, a specific binary code is generated in

ISSN:3007-9640 DOI: https://doi.org/10.70114/acmsr.2025.2.1.P332

sequence, and the value obtained after converting it to decimal is the LBP value of this point pixel. The calculation formula is as shown in Formula (5):

$$S(x)=\begin{cases}1 & K_i \geq M \\ 0 & K_i < M\end{cases}$$

$$LBP=\sum_{i=0}^{P-1} S(K_i - M)\times 2^P \tag{5}$$

The equivalent pattern is adopted to reduce the dimension of the local binary pattern operator, and solve the problem that there are too many binary patterns in face feature extraction. The proposal of the equivalent pattern effectively improves the system statistics.

In most cases, most of the images collected in an unrestricted situation are not frontal images, so the original LBP is unstable in feature extraction. The circular LBP basic operator adopted in this paper can perform rotation-invariant operations to obtain robust features. The calculation formula is as shown in Formula (6):

$$LBP_{P,R}^{rot}=\min\{ROR(LBP_{P,R,i} \mid i=0,1...,P-1)\} \tag{6}$$

3. Analysis of Simulation Results

3.1 Experiment on the Effect of Feature Fusion

The Fer2013 dataset consists of 35,887 images with a size of 48x48. Among them, the training set contains 28,710 images, the test set contains 3,562 images, and the validation set contains 3,615 images. In order to enhance the data training ability, the number of images in the dataset is increased by horizontal flipping and mirroring.

The Jaffe dataset is a facial expression dataset of Japanese women, which contains 213 grayscale images of seven basic facial expressions (anger, disgust, fear, happiness, sadness, surprise, and neutral) performed by 10 Japanese female actors. Each expression has three different degrees. This dataset is small, but it has a high annotation accuracy.

The CK+ dataset contains 593 expression sequences of 123 subjects, and all sequences are from a neutral expression to a peak expression. Only 327 of the 593 sequences have expression codes (0 = neutral, 1 = anger, 2 = contempt, 3 = disgust, 4 = fear, 5 = happiness, 6 = sadness, 7 = surprise).

In order to explore the role of α in the fused features, detailed feature fusion experiment analyses were carried out on the Fer2013 dataset, the Jaffe dataset, and the CK+ dataset respectively. The value of α was gradually increased from 0 to 1. When $\alpha = 0$, all were CNN features, and when $\alpha = 1$, all were LBP features. The experimental results are shown in Fig. 3.

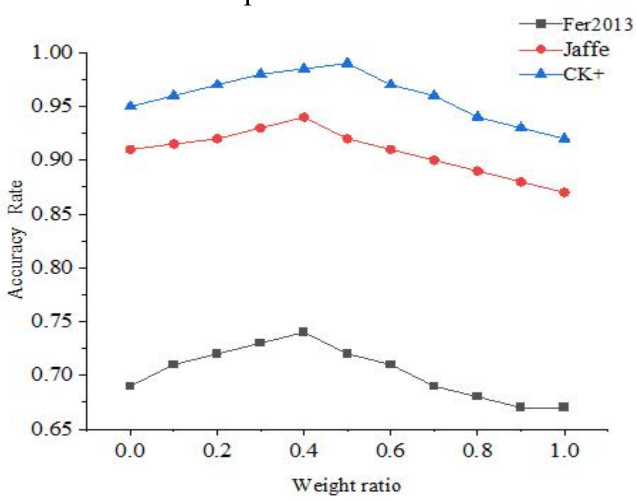

Fig. 3 The relationship diagram between the combination weight ratio and the accuracy rate

ISSN:3007-9640 DOI: https://doi.org/10.70114/acmsr.2025.2.1.P332

It can be known from the experimental results that as the value of α increases, the accuracy rate of the model begins to increase within a certain range. For the Jaffe and Fer2013 datasets, when $\alpha = 0.4$, the accuracy rate reaches the maximum, while for the CK+ dataset, when $\alpha = 0.5$, the accuracy rate reaches the maximum. Subsequently, when α continues to increase, the accuracy instead begins to decrease, because as the value of α increases, the fused features tend to be more inclined towards the LBP features, and the global features extracted by CNN gradually begin to lose their role, having an impact on the analysis of global features. Considering comprehensively, the value of α is fixed at 0.4.

3.2 Experiment of model performance

In order to verify the feasibility of the algorithm in this paper and measure the quality of the recognition results generated by the model, the experiments separately tested the average accuracy of the Attention-CNN model, the LSTM model, and the convolutional matrix decomposition model under the condition that the number of training epochs is 150, as shown in Table 1.

Table 1. Average accuracy rate of recognition under different algorithms

Algorithm	Average accuracy rate(%)
CNN model	69.03
Attention-CNN model	71.27
Convolutional matrix decomposition model	66.89
improved LBP + CNN	74.21

Based on the data in the table, it can be known that by integrating the local binary pattern algorithm on the basis of the CNN network model, the accuracy on the Fer2013 data set has reached 74.21%, and there is a significant improvement in the recognition rate compared with other algorithms.

Acknowledgements

This work was supported by the research projects of Jilin Province's educational science planning in 2024. (NO. GH24060).

References

[1] Mohan, K., Seal, A., Krejcar, O. et al. FER-net: facial expression recognition using deep neural net. Neural Computer & Application[J]33,9125-9136 (2021)

[2] Hajar Filali, Jamal Riffi, Ilyasse Aboussaleh, Mohamed Adnane Mahraz, Hamid Tairi. Meaningful Learning for Deep Facial Emotional Features[J]. Neural Processing Letters,. 2022,54(1): 387-404.

[3] WuF, Pang C, Zhang B. FaceCaps for facial expression recognition[J]. Computer Animation and Virtual Worlds,202l,32(3-4):e2021

[4] Sun X, Xia P, Ren F. Multi-attention based deep neural network with hybrid eatures for dynamic sequential facial expression recognition[J]. Neurocomputing2021,444:378-389

[5] Qingqing W . Micro-expression recognition method based on CNN-LSTM hybrid network[J]. International Journal of Wireless and Mobile Computing,2022,23(1):67-77.