Fuzzy Weighted Entropy Attention Deep Learning Method for Expression Recognition

Lisha Yao 1,2*, Haifeng Zhao 2

1 Institute of Information and Software, Anhui Xinhua University, Hefei, Anhui 230088, China
2 School of Computer Science and Technology, Anhui University, Hefei, Anhui 230601, China
*Corresponding author’s e-mail: yaolisha@axhu.edu.cn

Abstract. Deep learning network has been widely used in facial expression recognition. However, due to the complexity and variability of expression images and the influence of various factors such as illumination and individual differences, the recognition effect of existing methods needs to be improved. In order to improve the expressive ability of deep learning network, an attention mechanism based on fuzzy weighted entropy is introduced into deep learning network and applied to multi-channel facial expression recognition. Firstly, the convolutional neural network model is constructed to extract the convolutional layer multi-channel features of facial image. Then, the fuzzy weighted entropy of different channel features is calculated as the attention weight. Finally, the attention weight and multi-channel features were fused to extract the high-dimensional features of the facial expression image and sent to the full connection layer to complete the facial expression classification. Experiments in CK+ and JAFFE databases show that the average recognition rates of this method obtained 94.86% and 96.88%. It is better than the mainstream methods in recent years.

1. Introduction
Facial expressions are an important way to express emotions [1]. It has important application value in auxiliary medical treatment, auxiliary education, human-computer interaction and other fields. Expression recognition has become an important part of computer vision research and has attracted many researchers' attention. Traditional expression recognition methods mainly include three parts: image preprocessing, feature extraction and expression recognition and classification. The main research results are as follows: The key point comparison identification method proposed by Daniel et al. [2]. Liu et al. proposed a face expression recognition method based on face feature fusion extracted from salient regions of human faces [3]. Ye Zhen et al. proposed a hyperspectral image classification algorithm based on Gabor feature and local protection dimension reduction [4]. YuJun proposed the 3D modeling of dynamic facial expression recognition [5].

Since Hinton published an article in Science in 2006, deep learning has had a great impact on image classification and recognition. Deep learning network can build deeper and richer information of facial expressions, which has a good effect on image classification and recognition. Now the deep learning network can avoid the influence of artificial factors and extract more comprehensive and rich features. It can automatically extract and classify facial expression features and has a good performance in facial expression recognition. Literature [6] proposed a migrating convolutional neural network method for facial expression recognition, literature [7] proposed a facial expression...
recognition method based on cross-connection Lenet-5 network, and literature [8] proposed a convolutional neural network expression recognition method based on multi-resolution feature fusion. The above algorithm has achieved certain recognition effect, but the complex and changeable facial expression images are often easily affected by illumination, individual and ethnic differences. The representation ability of single channel convolutional neural network is not enough to represent complex and changeable expression images. In order to improve the representational ability of convolutional neural network and highlight the key target features of expression, this paper introduces the attention mechanism and proposes an attention mechanism based on fuzzy weighted entropy. In this paper, multi-channel low-level features of expression images are extracted by convolutional neural network, and the attention weight of multi-channel is calculated and the multi-channel features are fusion extracted by using the attention mechanism based on fuzzy weighted entropy, and the high-dimensional features are sent to the full connection layer to complete expression classification. Experimental results show that the proposed method optimizes the recognition effect.

2. Fuzzy Weighted Entropy Attention Mechanism

2.1. Fuzzy Weighted Entropy

Fuzzy theory can construct fuzzy sets for elements with uncertain or fuzzy boundaries, and use membership degree to judge the degree to which elements belong to a fuzzy set [9]. In literature [10], the author proposed the concept of exponential fuzzy weighted entropy and applied it to medical image fusion. Considering the uncertainty of the image, the matrix with the same size of the image is used as the fuzzy set, and the trigonometric function is used as the membership function. By calculating the membership function of each pixel as the weight of weighted entropy, the value of point \((i, j)\) in the regional window is \(x_{ij}\), and its corresponding trigonometric function membership is \(\mu_{ij}\), as shown in Equation (1):

\[
\mu_{ij} = \sin[\pi(1 - (x_{max} - x_{ij})/K)/2] + 1
\]

In Equation (1), \(0.5(x_{max} - x_{min}) < K < x_{max} \cdot x_{min}\) represents the maximum range and \(x_{min}\) represents the minimum range. The weighted entropy weight \(p_{ij}\) corresponding to each feature point \((i, j)\) is calculated as shown in Equation (2):

\[
p_{ij} = \frac{\mu_{ij}}{\sum_{k=1}^{n} \mu_{ij}}
\]

Where, \(n\) is the feature number of each fuzzy subset. Assuming the size of the region window is \(M \times N\), then \(n = M \times N\). Finally, the fuzzy weighted entropy \(H\) of each fuzzy subset is calculated, as shown in Equation (3).

\[
H = \sum_{k=1}^{n} [p_{ij}e^{1-p_{ij}} + (1 - p_{ij})e^{p_{ij}}]
\]

2.2. Fuzzy Weighted Entropy Attention Mechanism

The mechanism of attention originates from the theory of attention mechanism [11]. Visual observation has certain characteristics of attention. Attention always focuses on something important or interesting. Attention is shown as the degree of importance of each dimension on the corresponding feature. In order to integrate resources reasonably and effectively, attention mechanism can be used to
measure the importance of different inputs to tasks and selectively integrate input information reasonably and effectively when deep learning network is used to process information. $X = [x_1, \cdots, x_N]$ is used to represent $N$ inputs, and $z$ is used to represent the positions of inputs. According to the theory of attention mechanism, the importance of the $i$-th input to the current task is represented in the form of probability, that is, $\alpha_i$ represents the weight of attention of the $i$-th input.

$$\alpha_i = p(z=i|X) = \text{soft max}(S(x_i))$$  (4)

Where, $S(x_i)$ represents the scoring function of attention importance, and the corresponding fuzzy weighted entropy value of the $i$-th input is taken here, namely:

$$S(x_i) = H(x_i)$$  (5)

3. Multi-Channel Fuzzy Weighted Entropy Convolutional Neural Networks for Attention

3.1. General Design of Convolutional Neural Network

Convolutional neural network has a strong ability to express data, but in some practical tasks, convolutional neural network is difficult to achieve the desired effect. Because facial expression images are complex and changeable and susceptible to illumination, individual and ethnic differences, convolutional neural network (CNN) has encountered a bottleneck in the task of expression classification. Therefore, a fuzzy weighted entropy attention mechanism is proposed to improve the representational ability of convolutional neural network (CNN) model. In this paper, a convolutional neural network model of multi-channel fuzzy weighted entropy attention is designed. The model is divided into four parts: convolutional neural network feature extraction, fuzzy weighted entropy attention weight generation, attention weight and low-level feature fusion, and classifier. In the convolutional neural network feature extraction part, the convolutional neural network model structure is designed, and different convolutional kernel is used to generate multi-channel low-level features of expression images. The fuzzy-weighted entropy attention weight generation part is used to calculate the attention weight of the fuzzy-weighted entropy corresponding to the low-level features of each channel expression. The fusion part of attention weight and low level features is to fuse attention weight and multi-channel low level features to get the high level feature map of expression. In the part of classifier, the probability of expression high-level feature map corresponding to different categories is calculated. The category with the highest probability is the classification result. The overall architecture design of the above model is shown in Figure 1:

![Figure 1. The overall architecture of a multichannel fuzzy weighted entropy attention convolutional neural network model.](image)
one fully connected layer and one Softmax layer. The size of the input image is 96 × 96. Convolution layer of C1, C2 and C3 convolution kernels are 5 × 5 size, has 32, 64, 128 a convolution kernels, respectively. The pooling layer S1 adopts random pooling, and its pooling window size is 2 × 2. There are 300 neurons in the full connection layer. In order to prevent overfitting, Dropout operation is used in front of the full connection layer, and its value is set to 0.4. The Softmax layer contains seven neurons that divide expressions into seven categories. Specific parameter Settings of the model are shown in Table 1.

| Structure design of model |  
|---------------------------|
| input 96×96 |  
| C1: 5×5 conv, 32 |  
| C2: 5×5 conv, 64 |  
| S1: 2×2, random-pooling |  
| C3: 5×5 conv, 128 |  
| Dropout: 0.4 |  
| FC: 300 |  
| Softmax: 7 |  

3.2. Feature Extraction by CNN

Through the convolution and pooling operation of the designed CNN model, the features of the input expression images are extracted. Assume that the size of the feature graph output by the final convolutional layer is \( H \times W \times D \), then \( D \) represents the number of feature channels, and the size of the feature graph of each channel is \( N = H \times W \). Then the extracted low-level features can be expressed as \( X \in R^{(N \times D)} = \{ x_{i,d} \}, i \in \{1, \ldots, N \}, d \in \{1, \ldots, D \} \).

3.3. Fuzzy Weighted Entropy Attentional Weight

In order to improve the ability of the model to extract complex features, this paper proposes a fuzzy weighted entropy attention mechanism to generate attention weights. For the extracted low-level feature \( X \), let the attention feature graph be \( A \) and the dimension of attention be \( D \), corresponding to the selection process of \( D \) input features. \( A \in R^{(D)} = \{ \alpha_{d} \}, d \in \{1, \ldots, D \} \).

\[
\alpha_d = p(z = d \mid X) = \text{soft max}(S(x_d)) \tag{6}
\]

In Equation (6), \( S(x_d) \) represents the scoring function of the \( d \)-dimensional attention. For input \( S \) of each dimension of attention, the Softmax function is calculated as follows:

\[
\text{Soft max}(S) = \sum_{d=1}^{D} e^{S_d} \tag{7}
\]

The scoring function \( S(x_d) \) used in this paper is the fuzzy weighted entropy value \( H(x_d) \) corresponding to the \( d \)th input \( x_d \), i.e.

\[
\alpha_d = \text{Soft max} \sum_{d=1}^{D} e^{H(x_d)} \tag{8}
\]
Softmax function is used to amplify the key areas of attention and suppress the attention of other channels.

3.4. Multi-Channel Attention Feature Fusion

In order to extract the high-level feature $\Psi$, it is necessary to fuse the multi-channel low-level feature $X$ with the fuzzy weighted entropy attention weight feature graph. The dimension of the obtained high-level feature $\Psi$ is $N \times D$, and its fusion mode is shown in Equation (9). The fusion diagram is shown in Fig. 2.

$$\Psi(x_d) = \frac{1}{N} \sum_{i=1}^{N} \alpha_d \cdot x_{id}$$

Figure 2. Schematic diagram of multi-channel attention feature fusion.

3.5. Classifier

High-level feature $\Psi$ has a strong ability of image representation, so Softmax layer is used to get the probabilities corresponding to the seven categories through the full connection layer. The calculation method is as follows:

$$p_c(\Psi) = \text{Softmax}(W_c \Psi + \beta_c), c = 1, \cdots, C$$

In Equation (10), $W_c$ and $\beta_c$ are the connection weights and bias parameters of the full connection layer. Here $C = 7$ represents the seven expression categories. The dimension of the high level feature $\Psi$ is $N \times D$. When the number of channels $D$ is large, if the initialization parameter method of general network is adopted, the classifier in the full connection layer will reduce the learning ability and cannot express rich information by using high-level features. In order to fully express the rich category information of the high level features and improve the efficiency of network learning, the SVM model is introduced to initialize the parameters of the full connection layer. The dichotomous SVM models of each category are trained by high-level features, and the minimum value of each SVM model is obtained by Equation (11):

$$f_c(W_c, h_c) = \left[\frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i^c (w_c \cdot \varphi_i - h_c))\right] + \lambda \|w_c\|^2$$

$N_{\text{data}}$ is the amount of data used for initialization, and $y_i^c$ is the $i$th sample's information for each category $c$. $l_i$ is the correct category label of the $i$th sample, then:

$$y_i^c = \begin{cases} 1, & l_i = c \\ 0, & l_i \neq c \end{cases}$$

(12)
$W_c$ and $b_c$ are parameters of SVM that can be used in parameters $W_c$ and $\beta_c$ of the full connection layer. Therefore, parameters of the full connection layer can be initialized by training the SVM model to improve the classification accuracy of the network and save training time.

4. Results and Discussion

4.1. Experimental Environment and Data Preparation

This paper uses Python language to implement the experiment based on TensorFlow platform. CPU is Intel(R) Core(TM)i5, independent graphics card NVIDIA GeForce GTX 1050, 3G video memory, 4G memory. In order to illustrate the effectiveness of the expression recognition method in this paper, the CK+ database established by Carnegie Mellon University and the JAFFE database in Japan were used as data sets. CK+ database contains 7 types of facial expression images of 123 subjects, and 2940 facial expression images are collected. Figure 3 is an example of 7 types of facial expression images in CK+ database. JAFFE database contains 7 types of facial expression images of 10 Japanese women, with a total of 213 facial expression images. Figure 4 is an example of 7 types of facial expression images in JAFFE database. The experiment adopted 10-fold cross validation, which divided the data set into 10 parts on average. Each time, 9 of them were taken as the training set and the remaining one was taken as the test set. Finally, the average value of 10 tests was taken as the recognition result.

![Figure 3. 7 kinds of facial expression image in JAFFE expression dataset.](image1)

![Figure 4. 7 kinds of facial expression image in the CK+ dataset.](image2)

Because the data volumes of CK+ database and JAFFE database are different, the experiment sets different parameters for network training of the two data sets, and the specific parameter Settings are shown in Table 2.

| Data set | CK+ | JAFFE |
|----------|-----|-------|
| Learning rate | 0.00001 | 0.000001 |
| Number of test iterations | 200 | 2000 |
| Weight decay | 0.0005 | 0.0005 |
| Batch quantity | 20 | 10 |
| Maximum number of iterations | 100000 | 100000 |

4.2. Performance Analysis of the Method in This Paper

Based on the constructed CNN model, the low-level features of the multi-channel expression images were trained and the high-level features of the expression images were obtained through fuzzy weighted entropy attention fusion to complete the expression classification. After 100,000 iterations of training, the average recognition rate of 10 times of 10-fold cross validation statistical experiment was calculated as the recognition effect.

The confusion matrices of experimental 7 expressions in CK+ database and JAFFE database are shown in Table 3 and Table 4.
Table 3. CK+ confusion matrix %.

| Expression | anger | disgust | fear | happy | neutral | sad | surprise |
|------------|-------|---------|------|-------|---------|-----|----------|
| anger      | 95.72 | 1.56    | 0.57 | 0.36  | 0.28    | 1.02| 0.49     |
| disgust    | 2.27  | 94.24   | 0.98 | 0.21  | 0.84    | 0.92| 0.54     |
| fear       | 1.21  | 0.73    | 95.15| 0.18  | 0.62    | 1.81| 0.30     |
| happy      | 0.12  | 0.23    | 0.27 | 97.66 | 0.81    | 0.23| 0.68     |
| neutral    | 1.16  | 0.88    | 0.61 | 0.79  | 92.72   | 2.82| 1.02     |
| sad        | 1.27  | 1.29    | 2.11 | 0.28  | 2.27    | 92.36| 0.42     |
| surprise   | 0.45  | 0.67    | 1.11 | 0.33  | 0.82    | 0.44| 96.18    |

Table 4. JAFFE confusion matrix %.

| Expression | anger | disgust | fear | happy | neutral | sad | surprise |
|------------|-------|---------|------|-------|---------|-----|----------|
| anger      | 100   | 0       | 0    | 0     | 0       | 0   | 0        |
| disgust    | 0.96  | 96.77   | 0.73 | 0.12  | 0.52    | 0.72| 0.18     |
| fear       | 0.78  | 0.49    | 96.78| 0.26  | 0.73    | 0.63| 0.33     |
| happy      | 1.62  | 1.56    | 1.55 | 90.69 | 2.11    | 1.09| 1.38     |
| neutral    | 0     | 0       | 0    | 0     | 100     | 0   | 0        |
| sad        | 0.35  | 0.57    | 0.97 | 0.32  | 0.46    | 97.01| 0.32     |
| surprise   | 0.48  | 0.51    | 0.78 | 0.37  | 0.73    | 0.25| 96.88    |

As can be seen from Table 3 and Table 4, the identification effect of the proposed method in Jaffe database is better than that in CK+ database. The reason is that the sample of Jaffe database comes from 10 Japanese women, which is relatively simple, while the sample of CK+ database comes from 123 different countries and different genders, which is relatively complex. It can be seen from Table 3 and Table 4 that the error rate is high between the expressions of anger and boredom, sadness and calm, fear and sadness. The reason is that the facial states of these expressions are relatively similar, and the differences are not obvious and it is not easy to distinguish, which also shows that facial expression recognition is a fuzzy and complex task.

4.3. Performance Analysis of Different Recognition Methods

In order to further illustrate the performance of the method presented in this paper, the experimental results of facial expression recognition were calculated based on CK+ database by 10 times 10 fold cross validation. In the experiment, the traditional Gabor feature based machine learning method (reference [4]) and the convolutional neural network based expression recognition method (reference [7] and reference [8]) were respectively used to compare with the proposed method. Fig. 5 shows the average recognition rate of seven types of expressions in different methods.

![Figure 5. Average recognition rate of 7 kinds of facial expressions in different algorithms.](image)
The comprehensive performance data of different algorithms are shown in Table 5. As can be seen from Table 5, compared with the traditional machine learning recognition method and the recognition method based on convolutional neural network, the recognition effect of the proposed method has been improved to a certain extent on the premise of ensuring the recognition efficiency, but the recognition efficiency is still lower than the single feature recognition method in literature [4].

Table 5. Comprehensive performance comparison of different algorithms.

| Algorithm      | Recognition rate | Recognition time (ms) |
|----------------|------------------|-----------------------|
| Ref.[4]        | 92.10%           | 1690                  |
| Ref. [7]       | 89.01%           | 2246                  |
| Ref. [8]       | 92.06%           | 2655                  |
| Proposed method| 94.86%           | 1856                  |

5. Conclusions

In this paper, a fuzzy weighted entropy attention convolutional neural network algorithm for expression recognition is proposed. The fuzzy weighted entropy attention model is used to fuse the low-level features of multiple channels of expression images into higher-dimensional features to complete expression classification. The average recognition rate in CK+ database and JAFFE database is 94.86% and 96.88%, respectively. Compared with other recognition methods, it can be seen that the proposed method integrates multi-channel low-level features through the fuzzy weighted entropy attention model, which improves the model's robustness and representational ability, and improves the recognition rate while ensuring the recognition efficiency. Although the method in this paper has achieved some results, further exploration and research on different databases and refining facial expression features are needed in the future work.

Acknowledgments

Special thanks to the following funds for their support: Key Research Project of Natural Science in Universities of Anhui Province (No.KJ2020A0782); Provincial quality engineering in Anhui Province Grass-roots teaching and research office demonstration project (No.2018jyssf111); Data Science and Big Data Technology University First-Class Undergraduate Program Construction Center (No. 2020ylyzx02); University-level Quality Engineering Demonstration Experiment and Training Center "Big Data Comprehensive Experiment and Training Center" (No. 2020 sysxx01); 2020 Anhui Provincial College Student Innovation Plan Project (No. 202012216083).

References

[1] Wang X Y, Fan J Z, Liu H M, et al.(2019) Multi-expression sequence fusion recognition based on probabilistic cooperative representation. Laser and Optoelectronics Progress, 56: 79-85.
[2] Acevedo D, Negri P, Buemi M E, et al.(2017) A simple geometric-based descriptor for facial expression recognition. In: IEEE international Conference on Automatic Face&Gesture Recognition. Washington, DC. pp. 802-808.
[3] Liu Y P, Li Y B, Ma X, et al.(2017) Facial expression recognition with fusion features extracted from salient facial areas. Sensors, 17:712.
[4] Ye Z, Bai L, Nian Y J.(2016) Hyperspectral image classification algorithm based on Gabor feature and locality-preserving dimensionality reduction. Acta Optica Sinica, 36: 504-513.
[5] Yu J, Wang Z.(2017) A Video-Based Facial motion tracking and expression recognition system. Multimedia Tools and Applications, 76: 14653-14672.
[6] Zhai Y K, Liu J.(2018) Facial expression recognition based on transferring convolutional neural network. Journal of Signal Processing, 34: 729-738.
[7] Li Y, Lin X Z, Jiang M Y.(2018) Facial expression recognition with Cross-connect LeNet-5 network. Acta Automatic Sinica, 44: 176-182.
[8] He Z C, Zhao L Z, Chen C.(2018) Convolution neural network with multi-resolution feature fusion for facial expression recognition. Laser and Optoelectronics Progress, 55: 370-375.
[9] Zhao H F, Chen S H.(2018) KFCM algorithm with weighted membership for brain tissue segmentation of MRI image. Journal of Computer-Aided Design and Computer Graphics, 30:2055-2062.
[10] Yao L S, Zhao H F.(2011) Medical image fusion based on PCNN and exponentially weighted fuzzy entropy. Computer Engineering and Application, 47:198-201.
[11] Rensink R A.(2000) The Dynamic Representation of Scenes. Visual Cognition, 7:17-42