Facial Expression Recognition Algorithm Based on Convolution Neural Network and Multi-Feature Fusion

Xiaoping Tang, Tianping zhang*

Computer Department, Wuhan Donghu University, Wuhan, Hubei, 430212, China
email: tangxiaoping@wdu.edu.cn
*Corresponding author’s e-mail: zhangtianping@wdu.edu.cn

Abstract: Facial expression recognition is a hot topic in the field of computer vision. The related research results show high application value in many fields, such as human-computer interaction, intelligent emotional robot, fatigue driving detection, medical health, safety prevention and control, teaching evaluation and so on. However, the huge difference within the expression class still has a prominent impact on the expression recognition, and it is difficult to solve this problem by single feature and traditional feature fusion methods. Therefore, this paper uses neural network and feature fusion strategy to further characterize and classify the constructed features. Experiments show that this method can effectively overcome the interference caused by the differences within the expression class, and achieve ideal expression recognition results.

1. Introduction

Facial expression is a very important transmission medium of personal thoughts and perception of other people's behavior information, and it is also an important means of communication between people. With the development of artificial intelligence, people put forward higher requirements for human-computer interaction, hoping that the computer can recognize and understand the user's emotion from the facial expression like human, so as to make positive and accurate feedback.

The current problem facing face expression recognition is that there are large differences in expression classes, which are mainly manifested in the differences of expression attributes, the differences of expressions between different people and the differences of environmental factors. The different expression attributes show that different expressions have similar and different characteristics, and even the same kind of expression may have multiple expressions. The difference of expression attributes shows that the same kind of expression may have multiple ways of expression. Different people's expression of the same kind of expression may be different, or even quite different. To solve this problem, the common method of existing methods is to cut the face (remove the surrounding background, hair region), such as most of the artificial features and neural network models[1], but the results are not ideal. In addition, facial expression recognition is also more difficult due to similarities between facial expressions, such as joy and surprise, dislike and anger. In general, under the influence of the similarity between expression classes, the interference caused by the difference within expression classes is the most prominent problem and challenge in current facial expression recognition.

In this paper, we will improve the image preprocessing stage to remove the interference caused by the differences within the expression class, and combine with convolution neural network to extract
and classify the features, so as to improve the accuracy of expression recognition.

2. Basic principles

2.1 Basic framework of facial expression recognition

![Facial Expression Recognition Process](image)

Facial expression recognition refers to the recognition of people's emotional states based on face images. It's basic framework generally consists of three parts\cite{2}: face expression image preprocessing, expression feature extraction and selection, and expression recognition classification, as shown in figure 1. In the face image preprocessing stage, this paper will automatically filter the regions that are not representative of the expression features, and standardize the key areas, thereby reducing the interference for network training. Expression feature extraction and selection is the core part of expression recognition. The accuracy of expression feature extraction directly determines the accuracy of expression recognition system. In this paper, we will use three convolution subnetworks to extract features from the three key feature areas of face extracted in the previous step, and then integrate these features together. Expression recognition classification is to classify the features extracted in the previous step as input. Essentially, it measures the similarity between the features of the samples to be tested and the features of the training samples, and selects the category with the largest similarity value as the output result.

2.2 Convolutional Neural Network

Convolutional neural network (CNN)\cite{3} is used for feature extraction. Its basic structure includes data input layer, convolution layer, pooling layer, full connection layer and output layer. The operation process of convolutional neural network is introduced below.

First, input the image into the input layer in the form of matrix.
Second, feature extraction is performed under convolution operation of convolution layer.
Third, reduce the space size through the pooling layer.
Repeat the second and third steps several times.
Finally, the full connection layer classifies the images and outputs the corresponding values.

3. Expression Recognition Based on Convolution Neural Network and Multi-feature Fusion

The hair and background areas in face images can affect the generalization and robustness of classification models. If these areas cannot be filtered effectively during feature learning and classification model training, it is very likely to affect the final classification results. This is the main problem faced by face expression recognition. To solve this problem, this paper extracts and normalizes the key areas of face during the face image preprocessing stage, so as to automatically filter the above non-representative expression feature areas and reduce the interference caused by the differences of expression classes.
3.1 Algorithmic Overall Design

The architecture of expression recognition algorithm based on convolution neural network and multi-feature fusion mainly includes face key area extraction (left blue part) and block-based convolution neural network (right gray part). The block-based convolution neural network can be divided into backbone network (middle gray part) and classification section (right gray part). Backbone network is responsible for feature extraction, and classification section fuses and classifies the former features.

3.2 Face key area extraction

This paper uses facial muscle movement model [4] and face feature points extracted based on literature [5] method to design a segmentation algorithm for extracting three key areas of face: forehead, eyes and mouth. In the face image preprocessing stage, regions that are not representative of emoticon features are automatically filtered, and key regions are normalized to reduce the interference caused by intra-emoticon differences.

In this paper, two data sets, CK+ and kdef, are used to verify the results. Four samples were selected for each expression, and each sample had six expressions. Figure 3 shows the key areas of the face, which are divided according to the face key area segmentation algorithm: forehead (green box), eyes (blue box), and mouth (red box).
3.3 Block based convolutional neural network

After the extraction of the three key areas of the face, including the forehead, eyes and mouth, is completed, they correspond to the input of the three convolutional neural networks in the backbone network, so that the convolution layer can extract the features. The three convolutional neural networks are composed of six convolution layers (C1-C6 in figure 2), three pooling layers (M1-M3 in figure 2) and three fully connected layers (F1-F3 in figure 2). Then, the features of the three key areas are integrated in the full connected layer (F1-F3 in figure 2) and concatenation layer (CO in figure 2), and then transferred to the network classification. The top layer of network classification consists of two full connections (FC1 and FC2 in figure 2), two dropout layer (D1 and D2 in figure 2) and classification layer (CL in figure 2). The dropout layer is used to prevent over fitting, and the classification layer is used to classify the expression.

The convolution core size of convolution layer is 3×3, step length is 1, padding parameter is same, and the activation function of relu is marked as conv_{filter_size}, filter_size is the number of filters. The pooling layer uses the maximum pooling layer with step size of 2 and core size of 2×2, labeled max_pool. The full connection layer is marked fc_{activation}_{filter_size}. The main parameters of activation function are tanh and relu. The dropout layer is marked 'drop_ratio', where ratio is the probability of dropout. C in the classification layer CL represents the total number of facial expression categories. In each subnetwork, the receptive field of convolution kernel will become four times of the original after pooling layer, and the output sizes of the last pooling layer are 4×7, 8×4 and 7×4, respectively. Finally, using the stacking of convolution layer and pooling layer, and the complementarity of three different levels of facial expression features extracted from the key areas of human face (covering the visual bottom features to semantic level features), the information transformation from visual bottom features to semantic level features is modeled and learned.
4. Experimental process

4.1 Facial expression database
In order to verify the effectiveness of the algorithm in this paper, two data sets, KDEF and CK+, are used for experimental test and analysis. 840 front face photos were selected from KDEF dataset and divided into 10 groups of subject-exclusive, that is, the same person can only appear in one group, not in two or more groups at the same time. The experiment used a common 10-fold cross validation strategy. For each validation, select one group from 10 groups as the test set, and the other 9 groups as the training set, and run 10 times with the current settings, and select the best result as the report result of the revalidation. Repeat this process until 10 validations are completed, and then average the results of 10 validations as final results.

The algorithm in this paper divides CK+ into two different configurations:
· CK+ 107 (CK+ 10-fold 7 classes): This set divides 327 sequences with 7 expressions into 10 groups of subject-exclusive.
· CK+ 106 (CK+ 10-fold 6 classes): This set ignores contemptuous expressions, and the remaining 309 expression sequences from 107 characters are divided into 10 groups of subject-exclusive.

The last three images are extracted from each expression sequence of CK+ as the samples of training set and test set. The algorithm also uses 10-fold cross validation strategy in CK+ 107 and CK+ 106.

4.2 Analysis of experimental results
In order to predict the accuracy of the expression recognition algorithm in this paper, the following formula (1) is used:

\[ P = \frac{\text{Hit}}{\text{N}} \]  

Among them, hit represents the hit number of real expression tags in the test samples, and N represents the total number of test samples.

| method | KDEF     | CK+107   | CK+106   | average |
|--------|----------|----------|----------|---------|
| use the full face as input  | 89.56    | 94.55    | 95.86    | 93.48   |
| using the cropped face as input | 91.60    | 97.86    | 97.71    | 95.62   |
| using three key face regions as input | 91.10    | 98.01    | 98.20    | 96.07   |

Table 1 shows the experimental results of the algorithm in this paper on KDEF, CK+107 and CK+106 data sets. The experimental results show that the algorithm in this paper has the highest recognition accuracy among the three methods.

Compared with the whole face and the cropped face, the algorithm in this paper can effectively avoid the interference from the non-representative regions by extracting the key regions of human face, and achieves relatively good experimental results. The method based on convolutional neural network and multi-feature fusion can greatly reduce the interference caused by the difference within expression classes, and effectively recognize and classify facial expressions.

5. Conclusions
In this paper, we propose an expression recognition algorithm based on convolutional neural network and multi-feature fusion. By extracting three different levels of features from three key regions of human face, the interference of non representative regions of facial expression features on subsequent feature learning and recognition classification can be avoided from the source of feature learning, and it can also eliminate the interference caused by the differences in the size and position of facial organs. At the same time, it can normalize the key areas of face and provide fine-grained expression information, so as to increase the difference between expression classes and reduce the difference
within expression classes. Finally, the three key facial features are fused and classified based on convolution neural network, and the experimental results show that this method has higher recognition rate.

References
[1] Zhao X, Liang X, Liu L, et al. (2016) Peak-Piloted Deep Network for Facial Expression Recognition. In: Computer Vision – ECCV 2016, 9906: 425–442.
[2] Lu Y. (2019) Analysis and Research on Key Technologies of facial expression image recognition. Jilin University.
[3] Lu G, He J, Yan J, et al. (2016) Convolutional neural network for facial expression recognition. Journal of Nanjing University of Posts and Telecommunications, 36 (1): 16-22.
[4] Li Y, Zeng J B, Liu X and Shan S G. (2020) Progress and challenges in facial action unit detection. Journal of Image and Graphics, 25(11): 2293-2305.
[5] Kazemi V, Sullivan J. (2014) One millisecond face alignment with an ensemble of regression trees. In:IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1867-1874.
[6] Srivastava N, Hinton G, Krizhevsky A, et al. (2014) Dropout: A Simple Way to Prevent NeuralNetworks from Overfitting. J. Mach. Learn. Res., 15(1): 1929–1958.