Journal of Computers Vol. 31 No. 2, 2020, pp. 12-19
doi:10.3966/199115992020043102002

Facial Expression Recognition Based on Deep Residual Network

Junsuo Qu[1*], Ruijun Zhang[2], Zhiwei Zhang[2], Jeng-Shyang Pan[3]

[1] Xi'an Key Laboratory of Advanced Control and Intelligent Process, School of Automation,
Xi'an University of Posts and Telecommunications, Xi'an 710061, China
qujunsuo@xupt.edu.cn

[2] Innovation Lab, Xi'an University of Posts and Telecommunications, Xi'an 710061, China
{ruijunzhang, zzwft}@stu.xupt.edu.cn

[3] School of Information Science and Engineering, Fujian University of Technolgy, Fuzhou 350000, China
jengshyangpan@fjut.edu.cn

Received 13 January 2019; Revised 7 March 2019; Accepted 12 March 2019

Abstract. Low accuracy of facial expression recognition for traditional methods, a facial expression recognition algorithm is proposed. Using the deep residual network model as the feature extractor, the residual block of the residual network is improved to enhance the information flow in the deep network. During training, apply some pre-processing techniques to extract only expression specific features from a face image and explore the presentation order of the samples and use softmax to classify and identify the extracted feature vectors. The experimental results show that a higher recognition rate is obtained on FER-2013.

Keywords: deep residual network, facial expression recognition, pre-processing techniques, softmax

1 Introduction

Facial expressions are the most natural way to reveal the inner world and play a vital role in our social interactions. Through facial expressions, you can express your feelings and infer others' attitudes and intentions. Facial expression recognition (FER) is an essential part of the dynamic analysis and can be used to identify inner human emotions. FER methods attempt to classify facial expression in a given image or sequence of images as one of six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) or as "neutral" [1]. Due to the complexity and subtlety of facial expressions and their relationship to emotions, accurate recognition of facial expressions still faces great difficulties.

A typical FER system consists of three stages: (1) Face detection and localization. (2) Extract expression information from the located face. (3) A classifier (like an SVM) is trained on the extracted information to output the final expression labels. The facial expression recognition process is shown in Fig. 1.

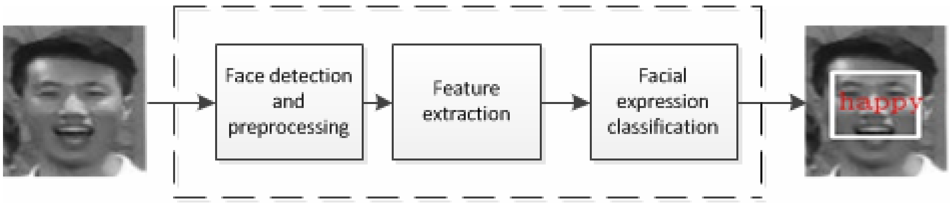
Face detection and preprocessing
Feature extraction
Facial expression classification
happy

Fig. 1. Facial expression recognition process

Traditional expression recognition is divided into three steps: feature learning, feature selection, and

[*] Corresponding Author

expression classification. Due to the ever-changing face images and complex backgrounds, it is increasingly difficult to meet the actual needs of manually extracted features. The deep learning method has the advantages of unsupervised learning, automatic feature extraction, and outstanding learning ability. It eliminates the traditional method of “first manual extraction of features and post-pattern recognition.” The three steps of expression recognition become a single step, and the input is an image rather than a set of manually encoded features. Liu et al. [2] proposed a boosted deep belief network (BDBN), which consists of a set of weak classifiers. Each weak classifier acts to increase the amount of training data for a particular table through the data enhancement, and in CK+ Experiments were performed on the data set and the three data sets created, and the expression recognition rate reached 93.5%.

The traditional convolutional neural network (CNN) has the problem that the training error increases with the increase of the number of network layers, but the deep residual network (ResNet) [3] can solve this problem with a deeper network. Our newly constructed facial expression recognition network improves the residual block of the residual network (FRES), using a deep residual network to extract features for facial expressions.

Our contributions can be summarized as follows:

(1) According to the scale of facial expression dataset, a simple and effective face feature extraction CNN model based on Resnet18 is proposed, which effectively solves the problem of facial feature extraction gradient disappearing and enhances the information flow in a deep network. (2) Design the cross-entropy loss function of the model to improve the training speed of the model. (3) The experimental results of our proposed algorithm on FER-2013 datasets show that the algorithm’s effectiveness and superiority over other state-of-the-art approaches.

2 Facial Expression Recognition Model

2.1 Facial Expression Recognition System Structure

Our facial expression recognition system only completes three learning stages of FER in one classifier (CNN). System operation is divided into two main phases: training and testing. During the training process, the system receives training data, which includes the grayscale of the face images and their respective expression ids, and learns a set of weights for the network. To ensure that training performance is not affected by the order in which samples are represented, some images are separated into validation and used, samples are presented in a different order, and the final set of optimal weights is selected in a set of training performed. During the test, the system receives the grayscale the face and its resolution and outputs the predicted result by using the final optimal weight set learned during the training process.

An overview of the system is shown in Fig. 2. In the training phase, new images are synthetically generated to increase the database size. After that, a rotation correction is carried out to align the eyes with the horizontal axis. Subsequently, the image is cropped to remove background information and to keep only expression specific features. A down-sampling procedure is carried out to get the features in different images in the same location. Then, the image normalized, and the convolutional neural network is trained using the normalized image. The output of the training phase is the set of weights of the round that achieved the best result with the validation data after a few training rounds considering data in different orders. The testing phase uses the same methodology as the training phase: the training phase is spatial normalization, cropping, down-sampling, and grayscale image intensity normalization. Its output is a single number - the id - of one of the seven basic expressions. The expressions are represented as integer numbers (0 - angry, 1 - disgust, 2 - fear, 3 - happy, 4 – sad, 5 – surprise, and 6 –neutral).

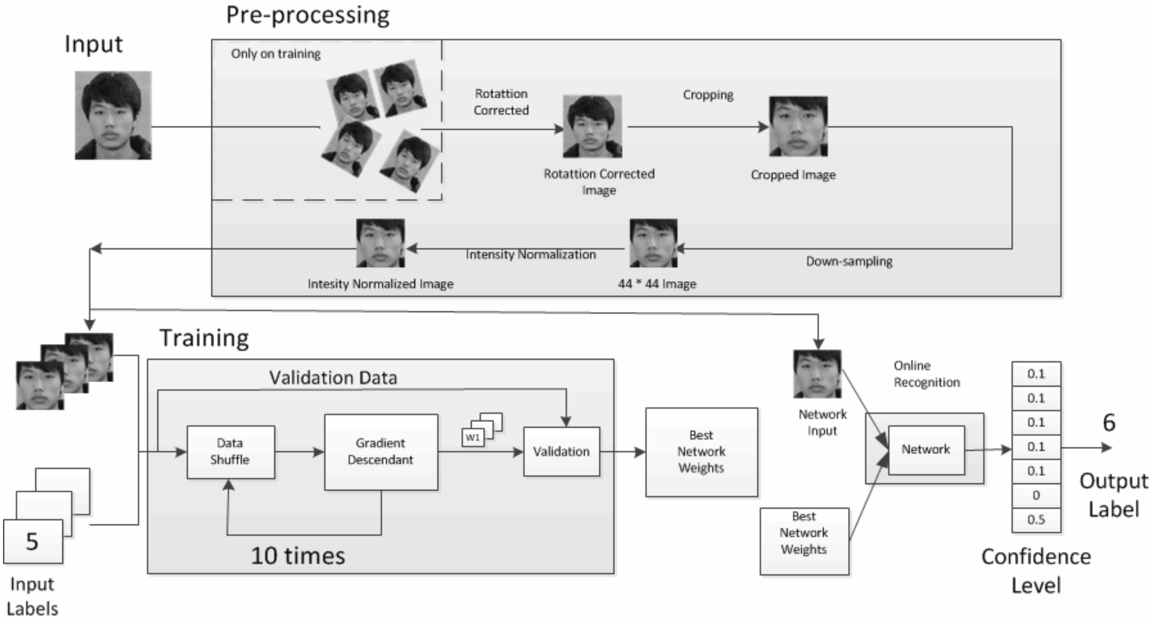

Fig. 2. Facial expression recognition system structure

2.2 Image Preprocessing

2.2.1 Face Detection

Face detection is performed by using the deep neural network (DNN) classifier model that comes with OpenCV. The DNN model is implemented according to the algorithm proposed in the literature [4] and trained by ResNet10 as the backbone network, which has high speed and accuracy of face detection. Suitable for different face directions, even working under severe occlusion, can detect faces of various scales, such as Fig. 3.

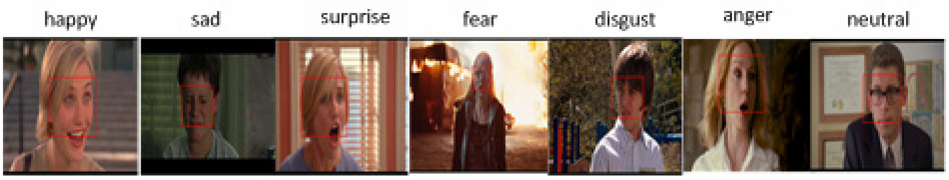

Fig. 3. Face detection result

2.2.2 Image Cropping

The information in the face frame (such as the ear, forehead, etc.) is also not important for the classification of facial expressions. This information may reduce the classification rate of facial expressions. Because the classifier still has a problem to solve, it is to distinguish between background and foreground. After cropping, all parts of the image without expression-specific information are deleted. The cutting process is shown in Fig. 4. The ear information is removed in the figure pass, and the coefficient is 0.2. Since the face is symmetrical, only the face image width is 0.6 times. These factor values are determined based on face structure and experience.

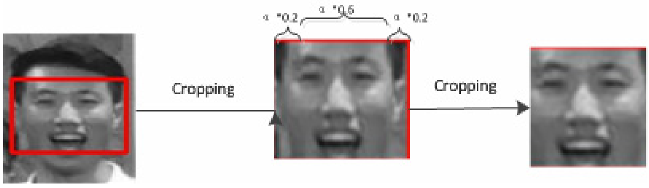

Fig. 4. Image cropping process

2.2.3 Downsampling

The operation of downsampling ensures the size of the image in the deep neural network and ensures the standardization of the scale, the position of the facial components in all images is the same. Downsampling uses linear interpolation to ensure that after resampling which is ensured the facial components in the image are in the same position, and it aims to reduce the running time of the GPU to perform the convolution process.

2.2.4 Intensity Normalization

The image brightness and contrast can vary even in images of the same person in the same expression. To reduce these variations issues, intensity normalization was applied. A method adapted from a bio-inspired technique described in [5], called contrastive equalization. The normalization is a two step procedure: firstly subtractive local contrast normalization is performed; and secondly, divisive local contrast normalization is applied. In the first step, the value of every pixel is subtracted from a Gaussian-weighted average of its neighbors. In the second step, every pixel is divided by the standard deviation of its neighborhood. The neighborhood for both procedures uses a kernel of 7 * 7 pixels (empirically chosen). An example of this procedure is illustrated in Fig. 5.

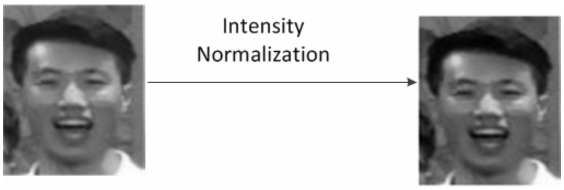

Fig. 5. Illustration of the intensity normalization

Equation (1) shows how each new pixel value is calculated in the intensity normalization procedure:

$$\chi' = \frac{\chi - \mu_{nhgx}}{\sigma_{nhgx}}. \tag{1}$$

Where χ' is the new pixel value, χ which is the original pixel value, μ_{nhgx} is the Gaussian weighted average of the neighbors of χ, and σ_{nhgx} is the standard deviation of the neighbors of χ.

2.3 Network Structure

The classic improvement of the CNN framework using ResNet [3] and DenseNet [6] not only eases the problem of gradient disappearance in deep network back propagation but also significantly improves the performance of image classification. To this end, under the incentive of the above network architecture, the dynamic facial expression recognition algorithm based on a convolutional neural network is based on the improved ResNet18 network framework, as shown in Fig. 6.

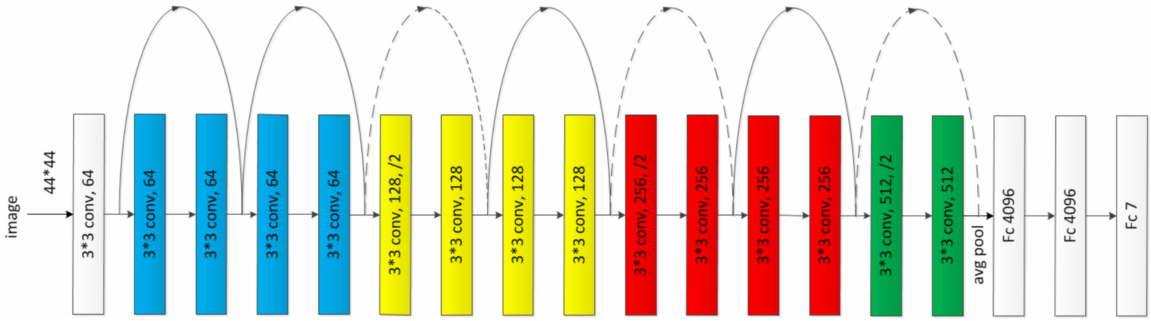

Fig. 6. Pipeline of the CNN framework

Our CNN framework consists of multiple convolutional layers, modified residual work blocks, and fully connected layers. The BatchNorm layer is set after each convolutional layer to increase the capacity of the model. The activation function uses relu and downsamples through the maximum pooling layer maxpool. Compared with AlexNet's parameters, the model has obvious advantages in the number of network layers and the number of parameters to be trained. The dimensions of the output vectors of the three fully connected layers Fc16, Fc17, and Fc18 are 4096, 4096, and 7 respectively. Among them, the size of the output of the Fc18 layer is the same as the number of categories to be classified. The Dropout layer was added after the Fc16 layer and the Fc17 layer, respectively, to reduce over-fitting of the network. The last layer is the softmax regression layer, which follows the Fc18 layer and is used to output the probability that the picture is divided into classes. The mathematical formula is shown in Equation (2), where k is the number of categories, and when k=2, Softmax degenerates into Logistic. There are 7 categories of models trained in this model, so k=7.

$$h_{\theta}(x^{(i)}) = \begin{bmatrix} p(y^{(i)}) = 1 \left| x^{(i)}; \theta \right| \\ p(y^{(i)}) = 2 \left| x^{(i)}; \theta \right| \\ \vdots \\ p(y^{(i)}) = k \left| x^{(i)}; \theta \right| \end{bmatrix} = \begin{bmatrix} e^{\theta_1^T x(i)} \\ e^{\theta_2^T x(i)} \\ \vdots \\ e^{\theta_k^T x(i)} \end{bmatrix}. \quad \textbf{(2)}$$

In RestNets, the output of the feature mapping of the residual block consists of a non-linearly transformed composite function H(x) and an identity function x, which are combined as in Equation (3):

$$F(x) = H(x) + x. \quad \textbf{(3)}$$

In order to improve the flow of information between layers, we improved the combination mode of the residual block. Motivated by DenseNet, We no longer summated the two inputs, but concatenated the two feature mappings. The output function of the feature mapping is shown in Equation (4).

$$F(x) = [H(x), x]. \quad \textbf{(4)}$$

Fig. 7. shows the structure of the traditional residual block and our modified residual block.

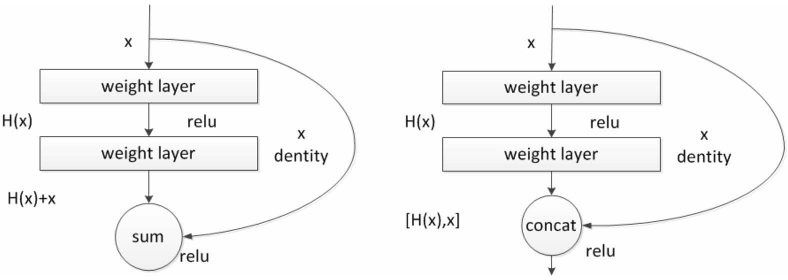

Fig. 7. Left: Traditional residual block. Right: Modified residual block

3 Network Training

3.1 Data Enhancement

To prevent the network from over-fitting too quickly, some image transformations such as flipping, rotating, cutting, etc. can be artificially performed, and the above operation is called data enhancement. In this experiment, during the training phase, we randomly cut the image into 44 * 44, and the image is randomly mirrored and then sent to the model training. In the testing phase, this article uses an integrated approach to reduce outliers. We cut and mirror the image in the upper left corner, the lower left corner, the upper right corner, and the lower right corner. This operation enlarges the database by 10 times and then sends the 10 images into the model. Then the obtained probability is averaged, and the largest output classification is the corresponding expression. This method effectively reduces the classification error.

The top images are 7 original faces in Fig. 8, the bottom images show their corresponding pre-processed faces.

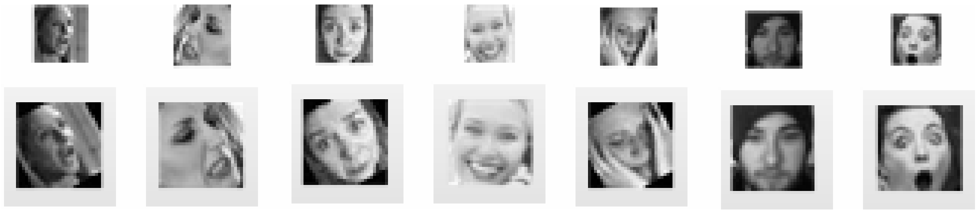

Fig. 8. Examples of pre-processed faces

3.2 Loss Function

In the design, we explored the calculation method of the cross-entropy loss function. After the model is fully connected, the output probability of each class is obtained, but the probability is not normalized. We normalize the probability to 1 through a softmax layer, which is easier to process. The cross entropy loss function is calculated as Equation (5). In softmax regression, we solve the multi-classification problem by the normalized probability. The class y can take k different values (instead of 2).

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{i}\log(h_{\theta}(x^{i})) + (1-y^{i})\log(1-h_{\theta}(x^{i}))] \cdot \tag{5}$$

3.3 Network Parameter Settings

The network is trained by Stochastic Gradient Descent (SGD), which training parameters are shown in Table 1.

Table 1. Training parameters

Parameter	Parameter values
Learning rate	0.0001
Momentum	0.9
Epochs	100
Metric	Accuracy
Dropout	0.5

4 Experimental Results

To verify the performance of the expression recognition algorithm based on the residual deep network, the experiments were performed on the FER-2013 dataset [7]. The FER-2013 dataset contains 27,809 training images, 3,589 validation images, and 3,589 test images. The face is marked with any of six basic expressions or neutral: the number of images representing the six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) and neutral is 4953, 547, 5121, 8989, 6077, 4002 and 6189. The images are 48 * 48 pixels.

The confusion matrix of FRES on the FER-2013 dataset (trained on FER-2013 training set) is shown in Fig. 9. The confusion matrix between the ground-truth class label and the most likely inferred class label information provide a better understanding of FRES's limitations. As expected, confusion frequently occurs between "anger", "fear", and "sadness" because they create similar motions.

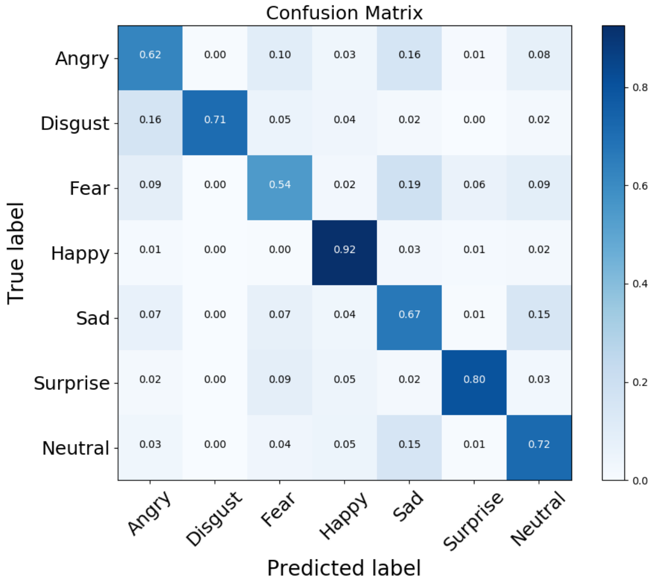

Fig. 9. FRES classification confusion matrices on the FER-2013 test set (trained on FER-2013 training set)

The CNN feature extraction method based on our algorithm is compared with the existing FER feature extraction method. Table 2 shows the recognition accuracy of different FERs, and our proposed algorithm achieves competitive results in the Fer-2013 dataset.

Table 2. Recognition rate of state-of-the-art methods on FER-2013 datasets (%)

Method	Recognition rate (%)
Unsupervised [7]	69.26
Tang [8]	71.16
Mollahosseini [9]	66.40
Liu [10]	65.03
DNNRL [11]	70.60
FC3072 [12]	70.58
CPC [13]	71.35
Proposed Approach	**73.00**

5 Conclusion

In this paper, we proposed a facial expression recognition algorithm based on the deep residual network. This algorithm improves the residual of the traditional residual network and enhances the information flow in the deep network, so its can extract features are more effective. In the representative picture FER-2013 dataset shows that our algorithm has a good effect on facial expression in terms of average recognition accuracy. However, the algorithm still fails to meet the requirements of autonomy, accuracy and real-time. Therefore, future research work will focus on optimizing and improving the proposed algorithm and try to change the input image type, improve the network structure, and fuse multiple deep neural network structures.

Acknowledgements

This work is partially supported by [International Cooperation and Exchange Program of Shaanxi Province] grant number [2018KW-026], [Natural Science Foundation of Shaanxi Province] grant number [2018JM6120], and [Major Science and Technology Projects of Xian Yang Bureau] grant number [2017k01-25-12], [Xi'an Science and Technology Projects] grant number [201805040YD18CG24(6)].

References

[1] P. Ekman, Contacts across cultures in the face and emotion, J Pers Soc Psychol 17(2)(1971) 124-129.

[2] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, in: Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE Computer Society, 2014.

[3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE Computer Society, 2016.

[4] D.E. King, Max-margin object detection. <https://arxiv.org/abs/1502.00046>, 2015.

[5] B.A. Wandell, Foundations of Vision, Sinauer Associates, 1995.

[6] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE Computer Society, 2017.

[7] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, Y. Zhou, Challenges in representation learning: a report on three machine learning contests, in: Proc. 2013 International Conference on Neural Information Processing, 2013.

[8] G. Zhao, M. Pietikäine, Boosted multi-resolution spatiotemporal descriptors for facial expression recognition, Pattern recognition letters 30(12)(2009) 1117-1127.

[9] A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: Proc. 2016 IEEE winter conference on applications of computer vision (WACV), 2016.

[10] K. Liu, M. Zhang, Z. Pan, Facial expression recognition with cnn ensemble, in: Proc. 2016 international conference on cyberworlds (CW), 2016.

[11] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, D. Tao, Deep neural networks with relativity learning for facial expression recognition, in: Proc. 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2016.

[12] B.K. Kim, J. Roh,, S.Y. Dong, S.Y. Lee, Hierarchical committee of deep convolutional neural networks for robust facial expression recognition, Journal on Multimodal User Interfaces 10(2)(2016) 173-189.

[13] T. Chang, G. Wen, Y. Hu, J. Ma, Facial expression recognition based on complexity perception classification algorithm. <https://arxiv.org/abs/1803.00185>, 2018.