Image Sequence Facial Expression Recognition Based on Deep Residual Network

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

[1] School of Automation, Xi'an Key Laboratory of Advanced Control and Intelligent Process, Xi'an University of Posts and Telecommunications, China
[2] Innovation Lab, Xi'an University of Posts and Telecommunications, China
[3] School of Information Science and Engineering, Fujian University of Technolgy, China
qujunsuo@xupt.edu.cn, {ruijunzhang, zzwft, qiaoning}@stu.xupt.edu.cn, jengshyangpan@fjut.edu.cn

Abstract

A sequence of facial expression images can provide rich texture information and motion information about facial expression changes. Combining traditional manual designed feature extraction methods with learning-based methods, this paper proposes an image sequence facial expression recognition algorithm based on deep residual network. Feature extraction is performed for each frame image, where the local binary pattern (LBP) map of the facial expression image is used as the input of the network, and the deep residual network model is used as the feature extractor for the image sequence. Then, each frame image feature is connected to a feature vector as the feature representation of the image sequence. Further, the image sequence is used as the input of the long short-term memory (LSTM) network, and the classification result is obtained through model training. Experimental results demonstrate the effectiveness of the proposed algorithm, where high recognition rates are observed based on both FER-2013 and AFEW6 datasets.

Keywords: Facial expression recognition, Local binary mode, Depth residual network, Long short-term memory

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" [1-2]. 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 Figure 1.

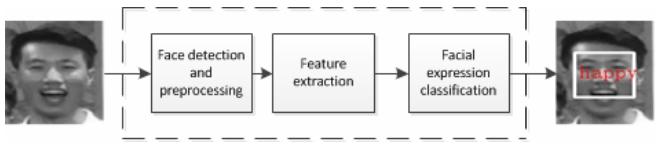

Figure 1. Facial expression recognition process

Facial expression recognition methods are divided into two categories based on static pictures and image-based sequences. Based on static images, you need to manually extract the direction for the specified feature. Common methods are based on face geometry [3], based on model matching methods, such as Active Appearance Models (AAM), based on frequency domain, pixel-based methods, such as Haar wavelet, Gabor wavelet transform, Local Binary Patterns (LBP), etc. These methods have two shortcomings: (1) Need to manually design a set of feature extraction directions, to some extent, will lose the feature information outside the artificial settings; (2) With a single static image as input, ignore the inter-frame information, reduce the recognition accuracy.

Since the facial features are more abundant than the facial features, false expression recognition may occur. The image sequence contains the local appearance features and motion information of facial expressions, which is closer to the essence of facial expression changes. Therefore, extracting dynamic features based on sequence images can eliminate various interference factors and become more adaptable. Yang et al. [4] extracted dynamic haar-like features from each frame image in the sequence image to form dynamic features,

*Corresponding Author: Junsuo Qu; E-mail: qujunsuo@xupt.edu.cn
DOI: 10.3966/160792642020112106001

including expression dynamic features and face appearance features. Zhao et al. [5] extended local binary pattern (LBP) to the VLBP (Volume Local Binary Patterns) feature, but only extracted the internal information of each frame, ignoring the time domain information of the expression action. Zhao et al. [6] further proposed the use of texture features of YT and XT planar images as dynamic features of the sequence. Because the time domain motion information of YT and XT plane images is more abundant, the effectiveness of dynamic features is improved to some extent.

Traditional expression recognition is divided into three steps: feature learning, feature selection, and expression classification. Due to the ever-changing face images and complex backgrounds, it is increasingly difficult to meet the actual need 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. [7] 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 gain, 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) method can only deal with the disadvantages of single-frame pictures. We have improved the structure of the original network and improved the residual block of the residual network. Using the LBP map of the image as the input to the residual network (LRES) allows the newly constructed network to identify the facial expression sequence.

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) Using timing features as input to the LSTM network, effectively utilizing sequence timing information. (3) We implemented the image sequence facial expression recognition on the AFEW 6.0 dataset. The results show that the LBP map of the image is beneficial to realize dynamic facial expression recognition.

2 Facial Expression Recognition Model

Our facial expression recognition model constructs a single-frame expression recognition network that uses datasets for training. In the face sequence expression recognition, the trained single frame expression recognize network is used as the feature extractor to recognite the feature extraction of the sequence expression.

2.1 Single Frame Facial Expression Recognition System

2.1.1 Single Frame 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 LBP map of the face and their respective expression ids, and learns a set of weights of 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 LBP map of 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 Figure 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 LBP map is extracted from 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 use the same methodology as the training phase: the training phase is spatial normalization, cropping, down-sampling, and LBP spectral 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).

2.1.2 Image Preprocessing

2.1.2.1 Face Detection

Face detection is performed by using the DNN classifier model that comes with OpenCV. The DNN model is implemented according to the algorithm proposed in the literature and trained by ResNet-10 as the backbone network, which has high speed and

accuracy of face detection. Suitable for different face directions (up, down, left, right, side, etc.), even working under severe occlusion, can detect faces of various scales, such as Figure 3.

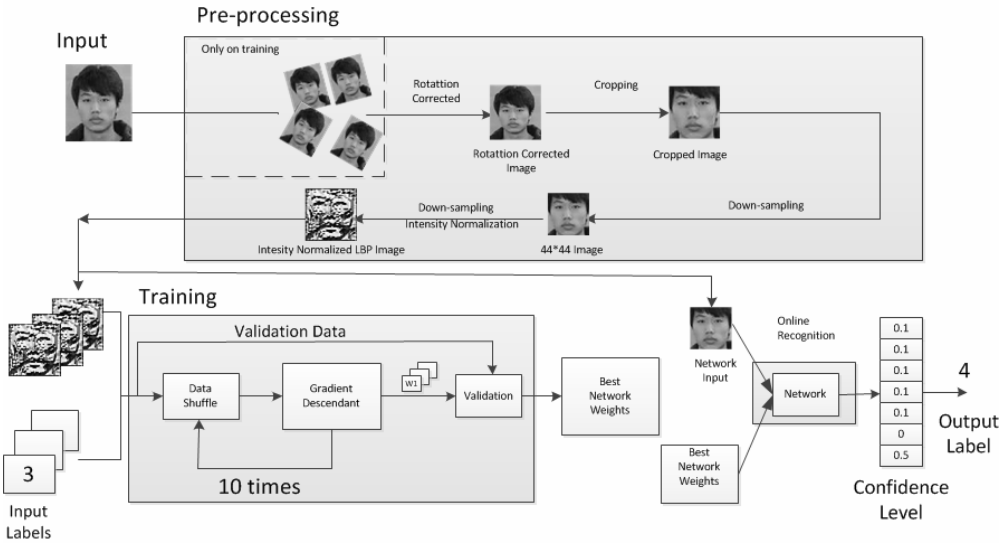

Figure 2. Single frame facial expression recognition system flow

Figure 3. Face detection result

2.1.2.2 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 Figure 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.

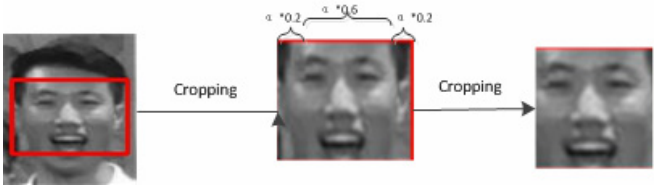

Figure 4. Image cropping process

2.1.2.3 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 (eyes, mouth, eyebrows, etc.) 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 the downsampling process aims to reduce the running time of the GPU to perform the convolution process.

2.1.2.4 Local Binary Mode

The LBP algorithm is used to reduce the manifold surface of the input image from a high dimension to low dimension. Because the LBP algorithm has gray invariance and Figure 5 can clearly distinguish the features of the face interest region, and at the same time dilute the smooth region with little research value. At the same time, the dimension of the feature is reduced and the running time of the large volume data of the deep convolutional neural network is reduced.

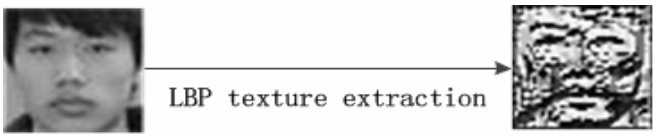

Figure 5. LBP extract texture features

2.1.2.5 Intensity Normalization

The image brightness and contrast can vary even in images of the same person in the same expression. Therefore, increasing the variation in the feature vector. Such variations increase the complexity of the problem that the classifier has to solve for each expression. To reduce these issues, intensity normalization was applied. A method adapted from a bio-inspired technique described in [8], called contrastive equalization, was used. Basically, the normalization is a two step procedures: 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 Figure 6.

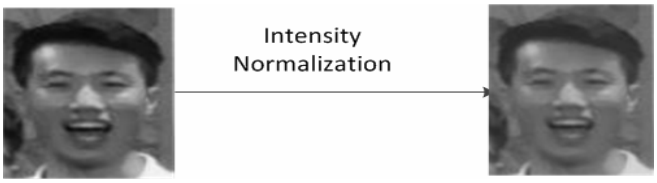

Figure 6. Illustration of the intensity normalization

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.

2.1.3 Network Structure

The classic improvement of the CNN framework using ResNet [9] and DenseNet [10] 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 Figure 7.

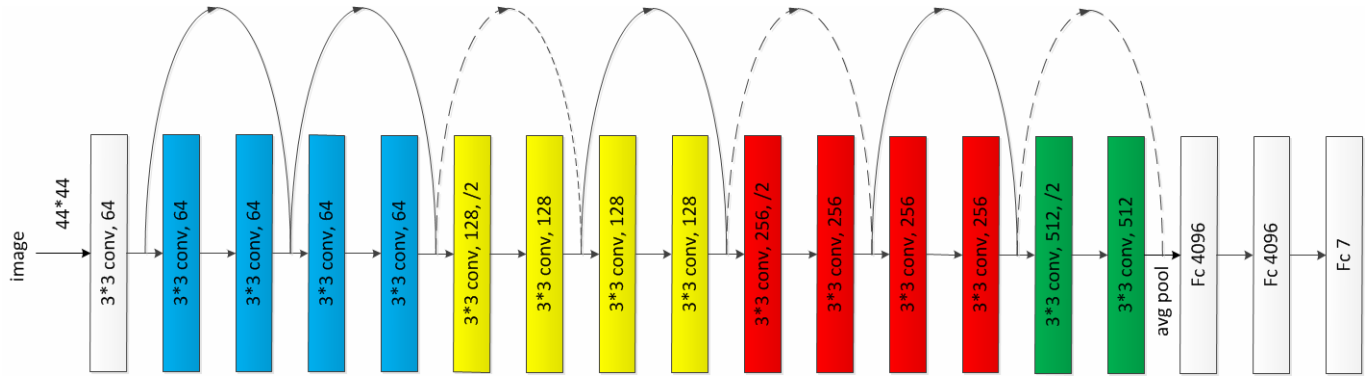

Figure 7. Pipeline of the CNN framework

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. The network model has a total of about 1.8 million parameters, and the total number of neuron connections in the network is about 468 million. Compared with AlexNet's 60 million 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 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. Softmax regression is a general form of logistic regression. The mathematical formula is shown in Equation (2):

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

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 $k = 7$.

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):

$$F(x) = H(x) + x \tag{3}$$

This combination may hinder the flow of information through deep networks. 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 input, but concatenated the two feature mappings. The output function of the feature mapping is shown in Equation (4).

$$F(x) = [H(x), x] \quad (4)$$

Figure 8 shows the structure of the traditional residual block and our modified residual block.

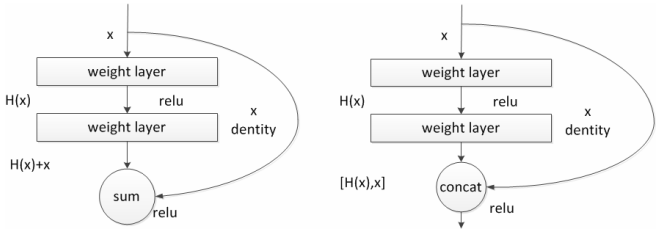

Figure 8. Left: Traditional residual block. Right: Modified residual block

2.2 Image Sequence Facial Expression Recognition System

2.2.1 Image Sequence Facial Expression Recognition System Structure

The image sequence facial expression recognition task is mainly divided into three stages: image sequence preprocessing, feature extraction, and expression classification. In the image sequence preprocessing stage, face detection is performed on each frame image of the input image sequence, and is directly discarded for images that do not include a human face. Next, detected face images are geometrically normalized so that all images remain the same size. Then, LBP calculation is performed on each frame image to obtain a corresponding LBP map and used as an input of the network structure. In the feature extraction stage, the trained network model is used as a feature extractor to extract features for each frame of image. Then, as the input of the Long short-term memory (LSTM) network in units of image sequences, the classification results are obtained through model training. The image sequence facial expression recognition process is shown in Figure 9.

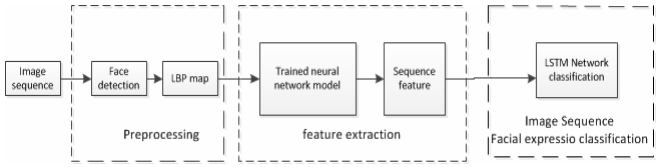

Figure 9. Algorithm framework

2.2.2 Feature Extraction and Classification

Different from the traditional feature extraction method, the proposed algorithm uses the trained convolutional neural network model for extraction. The process of feature extraction is shown in Figure 10. For an image sequence, the LBP map corresponding to each frame of its image is sequentially input into the network. Then, the trained network model is used as a feature extractor to extract features from the input LBP map. The feature map of the Fc16 layer in the network structure is selected as the feature vector of the input image, and the dimension size is 4096. In this way, the feature vector of each frame of the image sequence can be obtained.

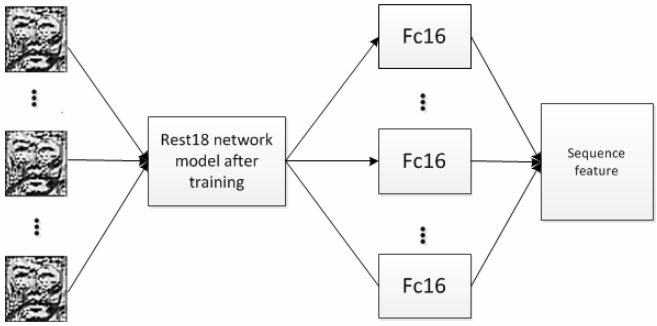

Figure 10. Feature extraction process

In order to obtain the feature representation of the image sequence, these feature vectors need to be analyzed and utilized. The algorithm adopted in this chapter is to combine the time dimension of the feature vector into a sequence feature in the image sequence as the feature representation of the image sequence. Thus, a 20 * 4096 feature vector can be obtained for an image sequence. This method is adopted because the fc6 layer feature map is the high-level feature representation of the image, and the advantage of LSTM processing timing information.

2.2.3 Image Sequence Classification Network

LSTM is a special RNN that solves the problem of gradient disappearance and gradient explosion during long sequence training. Compared with ordinary RNN, LSTM performs better in longer sequences. Input the extracted sequence features into the designed LSTM neural network, input the feature sequence 20 * 4096 into the network and train the model, and finally obtain the classification result through the output of softmax, so that the good combination and sequence timing information will be The timing information of one frame is utilized to realize facial expression recognition of the image sequence. The image sequence facial expression recognition process is shown in Figure 11.

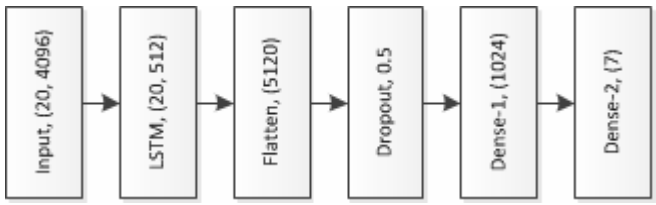

Figure 11. LSTM classification network structure

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. Another benefit of data manipulation is that it expands the amount of data in the database, making the trained network more robust. 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 Figure 12, the bottom images show their corresponding pre-processed faces. Each column represents the same expression: left to right, anger, disgust, fear, happiness, sadness, surprise and neutral.

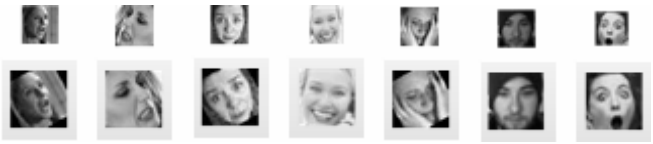

Figure 12. Examples of pre-processed faces

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).

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

3.3 Network Parameter Settings

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

3.4 Network Construction Enviroment

The experiment used the Pytorch framework to build the network. Experimental hardware platform: Inter(R) Core (TM) i5-8400 CPU, 32GB RAM, NVIDIA GeForce GTX 1080Ti GPU memory 11GB. The software platform is: the operating system is

Table 1. Training parameters

Parameter	Parameter values
Learning rate	0.0001
Momentum	0.9
Epochs	100
Metric	Accuracy
Loss function	cross-entropy
Dropout	0.5

configured with Linux Ubuntu 18.04, Python 3.5, NVIDlA CUDA Framework 9.0 and cuDNN libraries are installed.

4 Experimental Results

To verify the performance of the image sequence recognition algorithm based on the residual depth network, the experiments were performed on the FER-2013 dataset [11] and AFEW 6.0 dataset [12].

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 images are 48 * 48 pixels. AFEW 6.0 is the official dataset provided by the EmotiW2016 Emotional Recognition Challenge. It is containing 773 training samples, 383 validation samples, and 593 test samples. Since only the video clips of the training set and the verification set have emoticons, the performance only on the verification set.

4.1 FER-2013 Performance Results

The confusion matrix of LRES on the FER-2013 dataset is shown in Figure 13. The confusion matrix between the ground-truth class label and the most likely inferred class label information provides a better understanding of LRES's limitations.

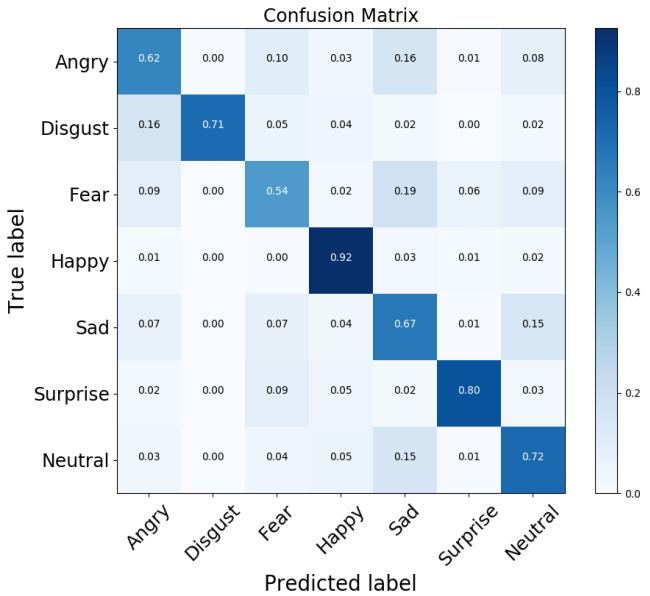

Figure 13. LRES classification confusion matrices on the FER-2013 test set (trained on FER-2013 training set)

The CNN feature extraction method based on LRES 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 [11]	69.26
Maxim [11]	68.82
Tang [13]	71.16
Mollahosseini [14]	66.40
Liu [15]	65.03
DNNRL [16]	70.60
FC3072 [17]	70.58
CPC [18]	71.35
Proposed Approach	**73.00**

4.2 AFEW 6.0 Performance Results

In the experiment, the trained network model is used to extract the features of the training set and verification set of the AFEW 6.0 dataset. Then, the LSTM network is used to train the feature vector of the training set. After the classification model is obtained, the verification set is tested. The accuracy of the expression recognition is 58.38%. Besides, the training set and verification set of the AFEW 6.0 dataset are combined, using 5 fold cross-validation method, 10 fold cross-validation method, 15 fold cross-validation method and leave-one-way cross-validation method for this experiment. The performance of the algorithm was verified. The experimental results are shown in Table 3. The experimental results show that when the 15-fold cross-validation method is adopted, the recognition accuracy is the highest.

Table 3. Expression recognition results based on AFEW 6.0 dataset

Authentication method	Accuracy (%)
5 Fold cross-validation	56.45
10 Fold cross-validation	57.56
15 Fold cross-validation	58.53
Leave-one-way cross-validation	57.37

Compare the algorithms presented in this chapter with several algorithms in the EmotiW2016 Challenge. The results are shown in Table 4. It should be noted that since the AFEW 6.0 dataset only provides the training set and the verification set. The emotional label does not provide an emotional tag for the test set. Therefore, the comparison result is based on the table of the validation set Emotional recognition accuracy rate.

Table 4. Comparison of facial expression recognition based on AFEW 6.0 Verification set

Authentication method	Accuracy (%)
Baseline	38.81
Fan [19]	51.96
Yao [20]	51.96
Bargal [21]	59.42
Proposed Approach	**58.38**

The recognition rates in Table 4 are the results in EmotiW2016 Recognition Challenge official. The baseline is 38.81%, and our proposed approach LRES is 58.38%, 19.57% higher than this benchmark. From the comparison results, the recognition accuracy of various algorithms is generally low, and it is difficult to identify the data set from the side. From the experimental results, the algorithm is still effective. Also, we used the original expression image as the input of the network to conduct experiments, and the recognition accuracy obtained on the verification set was 51.62%. This shows that the LBP map of the image is more conducive to the realization of dynamic facial expression recognition, and also shows that different types of input have different effects on network performance. Compare the algorithms presented in this chapter with several algorithms in the EmotiW2016 Challenge. The results are shown in Table 4. It should be noted that since the AFEW 6.0 dataset only provides the training set and the verification set. The emotional label does not provide an emotional tag for the test set. Therefore, the comparison result is based on the table of the validation set Emotional recognition accuracy rate.

5 Conclusion

In this paper, a dynamic facial expression recognition algorithm based on a deep residual network was proposed. As a special feature of this algorithm, the input of the network used the LBP map of the facial expression image, and the trained deep residual network was used as the feature extraction to obtain the sequence feature results from the LSTM network. In the representative facial expression picture datasets, FER-2013 and AFEW6.0, the proposed algorithm had a good effect on recognizing facial expressions in terms of average recognition accuracy. However, the algorithm failed to meet the requirements of autonomy, accuracy and real-time. Therefore, future work may focus on optimizing the proposed algorithm, improving the preprocessing operation, and integrating various 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 numbers [2018JM-6120 and 2019JM-606], and [Xi'an Science and Technology Projects] grant number [201805040YD18CG24(6)].

References

[1] P. Ekman, W. V. Friesen, Constants Across Cultures in the Face and Emotion, *Journal of Personalit and Social Psycholog*, Vol. 17, No. 2, pp. 124-129, February, 1971.

[2] J. S. Qu, L. C. Hou, R. J. Zhang, Q. P. Zhang, K. M. Ting, An Improved Measurement Variable Estimation Model for Positioning Mobile Robot, *Interaction Studies*, Vol. 20, No. 1, pp. 78-101, July, 2019.

[3] D. Ghimire, J. Lee, Geometric Feature-based Facial Expression Recognition in Image Sequences Using Multi-class Adaboost and Support Vector Machines, *Sensors*, Vol. 13, No. 6, pp. 7714-7734, June, 2013.

[4] P. Yang, Q. Liu, D. N. Metaxas, Boosting Coded Dynamic Features for Facial Action Units and Facial Expression Recognition, *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, 2007, pp. 1-6.

[5] G. Zhao, M. Pietikainen, Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 6, pp. 915-928, April, 2007.

[6] G. Zhao, M. Pietikäinen, Boosted Multi-resolution Spatio-temporal Descriptors for Facial Expression Recognition, *Pattern Recognition Letters*, Vol. 30, No. 12, pp. 1117-1127, September, 2009.

[7] P. Liu, S. Han, Z. Meng, Y. Tong, Facial Expression Recognition via a Boosted Deep Belief Network, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, Ohio,USA, 2014, pp. 1805-1812.

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

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, 2016, pp. 770-778.

[10] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, K. Keutzer, Densenet: Implementing Efficient Convnet Descriptor Pyramids, *Computer Science*, https://arxiv.org/pdf/1404.1869.pdf, 2014.

[11] 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, *International Conference on Neural Information Processing (ICONIP)*, Springer, Berlin, Heidelberg, 2013, pp. 117-124.

[12] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Acted Facial Expressions in The Wild Database, *Technical Report TR-CS-11*, September, 2011.

[13] Y. Tang, Deep Learning Using Linear Support Vector Machines, *Computer Science*, https://arxiv.org/pdf/1306.0239.pdf, 2013.

[14] A. Mollahosseini, D. Chan, M. H. Mahoor, Going Deeper in Facial Expression Recognition using Deep Neural Networks, *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, 2016, pp. 1-10.

[15] K. Liu, M. Zhang, Z. Pan, Facial Expression Recognition with CNN Ensemble, *2016 International Conference on Cyberworlds (CW)*, Chongqing, China, 2016.

[16] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, D. Tao, Deep Neural Networks with Relativity Learning for Facial Expression Recognition, *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Seattle, WA, USA, 2016, pp. 1-6.

[17] 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*, Vol. 10, No. 2, pp. 1-17, June, 2016.

[18] T. Chang, G. Wen, Y. Hu Y, J. Ma, Facial Expression Recognition Based on Complexity Perception Classification Algorithm, *Computer Science*, https://arxiv.org/pdf/1803.00185.pdf, 2018.

[19] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks, *The 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2016, pp. 445-450.

[20] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha,, Y. Chen, HoloNet: Towards Robust Emotion Recognition in The Wild, *The 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2016, pp. 472-478.

[21] S. A. Bargal, E. Barsoum, C. C. Ferrer, C. Zhang, Emotion Recognition in The Wild from Videos Using Images, *The 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2016, pp. 433-436.

Biographies

Junsuo Qu received his M.S. degree in the communication and information system from Xidian University, Xi'an, China, in 1998. He is a faculty member of the school of Automation from Xi'an University of Posts & Telecommunications as a Full Professor.

Ruijun Zhang received his B.S. degree in Electronic Information Science and Technology from Jilin Agricultural University, Changchun, China, in 2016. Now he is studying for a master's degree in Xi'an University of posts & Telecommunications.

Zhiwei Zhang received his B.S. degree in Communication Engineering from Xi'an Shiyou University, Xi'an, China, in 2017. Now he is studying for a master's degree in Xi'an University of Posts & Telecommunications.

Ning Qiao received his B.S. degree in Communication Engineering from Xi'an University of Posts & Telecommunications, Xi'an, China, in 2016. Now he is studying for a master's degree in Xi'an University of Posts & Telecommunications.

Jeng-Shyang Pan, male, Kaohsiung, Taiwan. He is currently Assistant to the President of Fujian Engineering, Dean of the School of Information Science and Engineering, Ph.D., Professor, Doctoral Supervisor. Mainly engaged in information security, cloud computing, and identification research.