Eurasia Journal of Science and Technology	ISSN: 2663-1024	Volume 4, Issue 1, pp 35-42	APR 2022

EXPRESSION RECOGNITION SYSTEM BASED ON DEEP LEARNING FRAMEWORK

Dan Li [1]*, Jinping Sun [1]*, Weiwei Liu [2] and Likai Wang [2]

1. School of Information Engineering(School of Big Data), Xuzhou University of Technology, Xuzhou, Jiangsu, China
2. Traffic police detachment of Xuzhou Public Security Bureau, Xuzhou Jiangsu, China
Corresponding Authors' Email: lidanonline@xzit.edu.cn, 313272361@qq.com

Abstract
Due to the changes of expression, background, position and noise, the automatic recognition of facial expression image is a challenge for computer. The system uses the face detection module in OpenCV and Dlib library, loads 68 key point detection models to detect faces, and annotates the key points on the image. The Fer2013 database is trained to get the position information of 68 key points on the face. The expression set is predicted by the classifier, and the predicted probability is displayed visually..
Keywords:facial expression recognition; convolutional neural network; feature extraction; Tenserflow framework.

1. INTRODUCTION

In recent years, with the continuous development of deep learning[1,2], the research of facial expression recognition has gradually stepped out of the laboratory and into practical application scenarios. At present, deep neural networks are widely used to learn the representation of facial expressions. Facial expressions can best reflect inner activities. With the in-depth research on facial expression recognition, how to obtain better facial expression representation has also become a hot research direction[3]. In addition, facial expression recognition also has many practical application scenarios. For example, after adding the facial expression recognition function, the question answering robot can provide more accurate services according to the owner's expression. In the medical field, doctors can judge the patient's tolerance according to the expression reported by the system. In fatigue driving monitoring, we can judge whether the driver is tired by facial expression.

The traditional method of extracting face features cannot extract deep features. In recent years, deep learning methods have made important contributions in the direction of face detection and registration. Compared with traditional methods, these models are more effective and can deal with face poses from a larger angle. Expression recognition based on deep learning method focuses on the face region, which requires preprocessing steps such as face detection, registration and size normalization. Some scholars have proposed a deep learning method[4,5] based on attention convolution network, which can focus on important parts of the face, and has achieved significant improvements over previous models on multiple data sets such as Fer2013, CK+, Ferg and Jaffe. Wang et al.[6] proposed a new method for static facial expression recognition. Through experiments on Fer2013 data set, a group of static images are divided into seven basic emotions by using CNN model, and effective classification is automatically realized. In the test set, the accuracy of facial expression recognition reached 68.79%. He et al.[7] proposed a multi-channel deep neural network, which learns and integrates space-time features. The basic idea of this method is to select the optical flow in the transformation process between the emotional face and the neutral face as a kind of temporal information of expression, and at the same time, it can also take an emotional gray-scale image generated by the face as spatial information. The recognition rate of this method is 98.38% in CK + database, 99.17% in Rafd database and 99.59% in MMI database. Inspired by the illumination invariant LBP descriptor, Levi et al.[8] extracted LBP features based on RGB images and mapped them to 3D space, used the LBP 3D mapping and RGB images to train convolutional neural networks respectively, and fused them at the decision level. In addition, Zhang et al.[9] extracted SIFT features based on 68 key points of the face, thereby converting the image into a 68*128 feature matrix as the input of CNN. Unlike ordinary CNN, because each row in the feature matrix represents SIFT features on a key point of the face, one-dimensional convolution and one-dimensional pooling are used to further learn the high-level semantic features related to expression in the feature matrix. Yu et al.[10] used multiple face detectors to detect faces and designed multiple CNN models. Each CNN model was first pre trained by other natural scene expression databases, and then migrated to the target natural scene expression database

Although the feature learning ability of deep learning is very strong, there are still some deficiencies in the application of human expression recognition. First of all, in order to avoid excessive fitting, the deep neural network has a lot of additional needs in the amount of training data. However, by analyzing the existing facial expression database, it cannot train a neural network with a certain depth architecture. In addition, based on the differences of personal attributes, such as actual age, gender, male and female, as well as ethnic background and expressive level, the differences between their subjects can still not be ignored. Of course, in addition to the differences caused by the main body, the changes of posture, illumination intensity and occlusion also have a greater impact on the results. The relationship between

these factors and facial expressions is nonlinear coupling. Therefore, in order to solve the problem of huge intra class variability and learn effective specific expression features, the requirements of deep network must be improved.

This paper designs a human face expression recognition model based on convolutional neural network. The model includes face recognition, face key point description, face expression recognition and face expression classification sub modules. The model is developed in Python language. Based on the deep learning Tensorflow framework, it is developed by using OpenCV, Dlib, Numpy, and Keras libraries.

1.1data set

A. Fer2013 dataset

In the challenge of feature representation of ICML (International Conference on Machine Learning) in 2013, Google Image Search API proposed Fer2013 dataset. This dataset is a large and less restricted data set collected online. The pictures in the dataset are unified to 48*48. The whole data set includes 28709 images for training, 3589 images for validation and 3589 images for testing. Each image is accompanied by seven expression tags, namely (anger, discrepancy, fear, happiness, safety, surprise and neutral).

Figure 1 Fer2013 database samples

B. CK+ dataset

The extended Cohn-Kanade (CK+) database contains 593 video sequences from 123 subjects, including the label of expression and the label of action units. The database is composed of 123 test objects, including all kinds of expressions and labels of various parts of motor units. The test objects are college students aged 18-30 years old. There are 593 dynamic image sequences, 65% of which are women. The last frame of the dynamic image sequence is used as action unit labels, and 327 sequences containing emotional labels.

C. WIDER FACE dataset

The wide face dataset is used for the evaluation of face detection algorithm. This dataset was proposed by the Chinese University of Hong Kong to solve the problem of saturation of face detection algorithm in the existing dataset. The whole dataset contains 32203 images and 393703 manually labeled faces. The number and scale of images are much larger than other datasets. These images include face images in different scenes, including face images with fuzzy image quality and different attitudes. An image contains dozens or even hundreds of faces.

2.OVERALL STRUCTURE OF SYSTEM MODEL

Figure 2 shows the flow chart of the whole model, which includes the following five major processes: image preprocessing, feature extraction, model training, expression classification and visualization results.

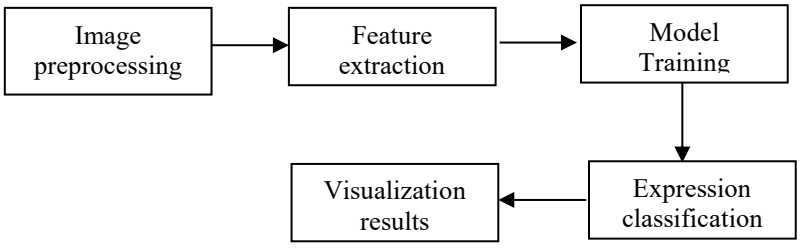


Figure 2 Model processing flow chart

The system mainly includes the following functions: 1) Face recognition sub function: when face collection information appears in the window after running the program, the background code will use the face detection module in OpenCv and Dlib library to implement face detection. 2) Face key point recognition: use the key point

detection model in Dlib library to detect the key points of the recognized face. Use Dlib library to load 68 key point detection model to detect face, as shown in Figure 3. 3) Facial expression recognition: directly load the pre-trained models in Python library. Here, Fer2013 database is used to fine tune the classification model and classify the expressions. The expressions are divided into seven kinds: anger, doubt, fear, happy, sad, surprise, neutral. The pixel values of Fer2013 database are shown in Figure 4. 4) Facial expression classification: classify the recognized expressions, give the highest probability to the most consistent expression words, and visually display the expression probability.

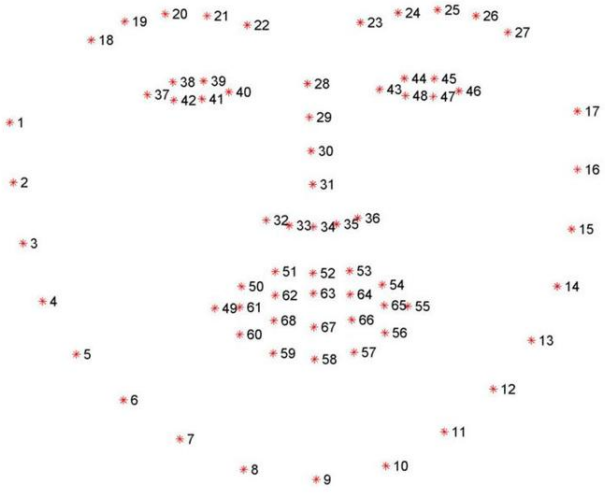
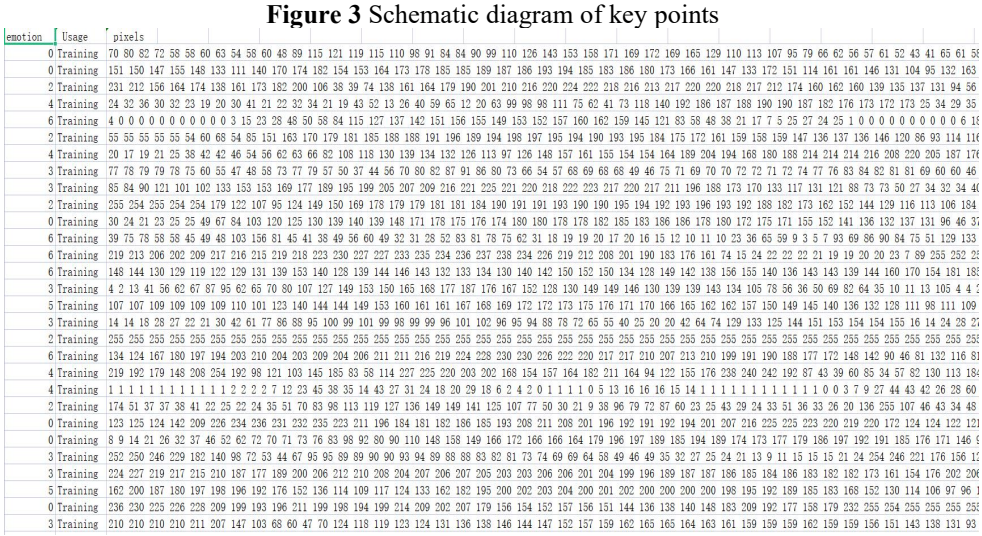


Figure 3 Schematic diagram of key points



emotion	Usage	pixels
0	Training	70 80 82 72 58 58 60 63 54 58 60 48 89 115 121 119 115 110 98 91 84 84 90 99 110 126 143 153 158 171 169 172 169 165 129 110 113 107 95 79 66 62 56 57 61 52 43 41 65 61 59
0	Training	151 150 147 155 148 133 111 140 170 174 182 154 153 164 173 178 185 185 189 187 186 193 194 185 183 186 180 173 166 161 147 133 172 151 114 161 161 146 131 104 95 132 163
2	Training	231 212 156 164 174 138 161 173 182 200 106 38 39 74 138 161 164 179 190 201 210 216 220 224 222 218 216 213 217 220 220 218 217 212 174 160 162 160 139 135 137 131 94 56
4	Training	24 32 36 30 32 23 19 20 30 41 21 22 32 34 21 19 43 52 13 26 40 59 65 12 20 63 99 98 98 111 75 62 41 73 118 140 192 186 187 188 190 190 187 182 176 173 172 173 25 34 29 35
6	Training	4 0 0 0 0 0 0 0 0 0 0 3 15 23 28 48 50 58 84 115 127 137 142 151 156 155 149 153 152 157 160 162 159 145 121 83 58 48 38 21 17 7 5 25 27 24 25 1 0 0 0 0 0 0 0 0 0 0 6 18
2	Training	55 55 55 55 55 54 60 68 54 85 151 163 170 179 181 185 188 188 191 196 189 194 198 197 195 194 190 193 195 184 175 172 161 159 158 159 147 136 137 136 146 120 86 93 114 116
4	Training	20 17 19 21 25 38 42 42 46 54 56 62 63 66 82 108 118 130 139 134 132 126 113 97 126 148 157 161 155 154 154 164 189 204 194 168 180 188 214 214 214 216 208 220 205 187 176
3	Training	77 78 79 79 78 75 60 55 47 48 58 73 77 79 57 50 37 44 56 70 80 82 87 91 86 80 73 66 54 57 68 69 68 68 49 46 75 71 69 70 70 72 72 71 72 74 77 76 83 84 82 81 81 69 60 60 46
3	Training	85 84 90 121 101 102 133 153 153 169 177 189 195 199 205 207 209 216 221 225 221 220 218 222 223 217 220 217 211 196 188 173 170 133 117 131 121 88 73 73 50 27 34 32 34 46
2	Training	255 254 255 254 254 179 122 107 95 124 149 150 169 178 179 179 181 181 184 190 191 191 193 190 190 195 194 192 193 196 193 192 188 182 173 162 152 144 129 116 113 106 184
0	Training	30 24 21 23 25 25 49 67 84 103 120 125 130 139 140 139 148 171 178 175 176 174 180 180 178 178 182 185 183 186 186 178 180 172 175 171 155 152 141 136 132 137 131 96 46 31
6	Training	39 75 78 58 58 45 49 48 103 156 81 45 41 38 49 56 60 49 32 31 28 52 83 81 78 75 62 31 18 19 19 20 17 20 16 15 12 10 11 10 23 36 65 59 9 3 5 7 93 69 86 90 84 75 51 129 133
6	Training	219 213 206 202 209 217 216 215 219 218 223 230 227 227 233 235 234 236 237 238 234 226 219 212 208 201 190 183 176 161 74 15 24 22 22 22 21 19 19 20 20 23 7 89 255 252 23
6	Training	148 144 130 129 119 122 129 131 139 153 140 128 139 144 146 143 132 133 134 130 140 142 150 152 150 134 128 149 142 138 156 155 140 136 143 143 139 144 160 170 154 181 183
3	Training	4 2 13 41 56 62 67 87 95 62 65 70 80 107 127 149 153 150 165 168 177 187 176 167 152 128 130 149 149 146 130 139 139 143 134 105 78 56 36 50 69 82 64 35 10 11 13 105 4 4 3
5	Training	107 107 109 109 109 109 110 101 123 140 144 144 149 153 160 161 161 167 168 169 172 172 173 175 176 171 170 166 165 162 162 157 150 149 145 140 136 132 128 111 98 111 109
3	Training	14 14 18 28 27 22 21 30 42 61 77 86 88 95 100 99 101 99 98 99 99 96 101 102 96 95 94 88 78 72 65 55 40 25 20 20 42 64 74 129 133 125 144 151 153 154 154 155 16 14 24 28 27
2	Training	255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255
6	Training	134 124 167 180 197 194 203 210 204 203 209 204 206 211 211 216 219 224 228 230 230 226 222 220 217 217 210 207 213 210 199 191 190 188 177 172 148 142 90 46 81 132 116 81
4	Training	219 192 179 148 208 254 192 98 121 103 145 185 83 58 114 227 225 220 203 202 168 154 157 164 182 211 164 94 122 155 176 238 240 242 192 87 43 39 60 85 34 57 82 130 113 184
4	Training	1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 7 12 23 45 38 35 14 43 27 31 24 18 20 29 18 6 2 4 2 0 1 1 1 1 0 5 13 16 16 16 15 14 1 1 1 1 1 1 1 1 1 1 1 0 0 3 7 9 27 44 43 42 26 28 60
2	Training	174 51 37 37 38 41 22 25 22 24 35 51 70 83 98 113 119 127 136 149 149 141 125 107 77 50 30 21 9 38 96 79 72 87 60 23 25 43 29 24 33 51 36 33 26 20 136 255 107 46 43 34 48
0	Training	123 125 124 142 209 226 234 236 231 232 235 223 211 196 184 181 182 186 185 193 208 211 208 201 196 192 191 192 194 201 207 216 225 225 223 220 219 220 172 124 124 122 121
0	Training	8 9 14 21 26 32 37 46 52 62 72 70 71 73 76 83 98 92 80 90 110 148 158 149 166 172 166 166 164 179 196 197 189 185 194 189 174 173 177 179 186 197 192 191 185 176 171 146 9
3	Training	252 250 246 229 182 140 98 72 53 44 67 95 95 89 89 90 90 93 94 89 88 88 83 82 81 73 74 69 69 64 58 49 46 49 35 32 27 25 24 21 13 9 11 15 15 15 21 24 254 246 221 176 156 12
3	Training	224 227 219 217 215 210 187 177 189 200 206 212 210 208 204 207 206 207 205 203 203 206 206 201 204 199 196 189 187 187 186 185 184 186 183 182 182 173 161 154 176 202 206
5	Training	162 200 187 180 197 198 196 192 176 152 136 114 109 117 124 133 162 182 195 200 202 203 204 200 201 202 200 200 200 198 195 192 189 185 183 168 152 130 114 106 97 96 1
0	Training	236 230 225 226 228 209 199 193 196 211 199 198 194 199 214 209 202 207 179 156 154 152 157 156 151 144 136 138 140 148 183 209 192 177 158 179 232 255 254 255 255 255 255
3	Training	210 210 210 210 211 207 147 103 68 60 47 70 124 118 119 123 124 131 136 138 146 144 147 152 157 159 162 165 165 164 163 161 159 159 159 162 159 159 156 151 143 138 131 93

Figure 4 picture pixel value

2.1Facial expression recognition and classification

D. Data preprocessing and data enhancement

Before the feature training of deep neural network learning, it needs to complete the preprocessing, in order to adjust the visual semantic information expressed by the face. The model uses cross validation to construct training set and test set, and uses face alignment, video normalization, data set expansion and other technologies for preprocessing. On different data sets, the established online classroom student emotion recognition model is compared with different baseline models. In addition, the self built video emotion database is studied. Although the facial muscle movements in the expression database played by students are large, the expression characteristics are strong, and the duration is long, so it is relatively easy to collect, but in reality, most students are in normal mood, and the expression change range is not too large, and the expression unconsciously revealed by students is closer to the real learning state. Therefore, this paper collects natural expressions in different states through problem-solving induction. In addition, in view of the insufficient size of the distance learning sample database, which makes the training depth neural network prone to over fitting, the data can be enhanced through image preprocessing and image transformation methods, such as the rotation, noise, rotation, translation, contrast change of the training samples, so that the enhanced student expression recognition network model has stronger robustness to environmental interference.

E. Expression recognition based on improved convolutional neural network

1) Residual network (RESNET)

As the application of convolutional neural network is more and more extensive, its potential problems are

gradually exposed. When the number of layers of neural network is deepened, the effect of learning is not positively correlated with the change of the number of layers. In other words, the convolutional neural network model appears degradation. Researchers have proposed three possibilities for such problems: over fitting and gradient disappearance or gradient explosion. The increase of network layers cannot improve the classification performance, but reduce the convergence rate and accuracy.

ResNet[11,12] is a feature extraction network model proposed by three Chinese scientists, and it is also one of the most widely used CNN networks at present. From the perspective of information theory, ResNet introduces direct mapping, and the network must contain more information than the previous layer of network, avoiding the problem that the amount of information will decay with the deepening of layers in the forward propagation process caused by DPI (data processing inequality).

2) Residual block

The residual block is the constituent unit of the residual network, which is composed of residual and direct mapping. It is expressed by formula as follows:

$$y_l = h(x_l) + F(x_l, W_l) \quad (1)$$

$$x_{l+1} = f(y_l) \quad (2)$$

For deeper layers, it can be expressed as:

$$x_L = x_l + \sum_{i=1}^{L-1} F(x_i, W_i) \quad (3)$$

The structure of the residual block is shown in Figure 5, where weight refers to convolution:

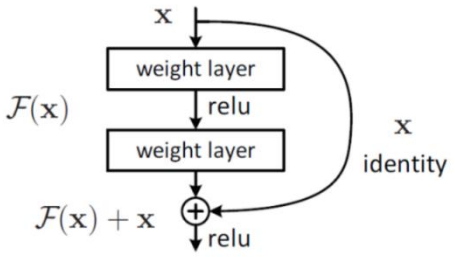


Figure 5 Residual block structure

The loss function loss is expressed as:

$$\frac{\partial loss}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \left[1 + \frac{\partial}{\partial x_L} \sum_{i=1}^{L-1} F(x_i, W_i) \right] \quad (4)$$

3) Bottleneck module

The bottleneck module skillfully controls the dimension of feature mapping by using a 1*1 convolution kernel, so that the classification number of the subsequent 3*3 convolution kernel will not be affected by the previous results.

4) Residual network structure

ResNet network is an improvement of VGG19 network, which integrates residual units into the network through short circuit mechanism. The important change is that the convolution with step size of 2 is used for down sampling, and GPA (global average pooling) is used to replace the full connection layer. When designing ResNet network, in order to ensure the complexity of network layer, the product of feature map size and number is set as a fixed value. When the size is halved, the number will double. Figure 6 shows the comparison between ResNet structure and other network structures:

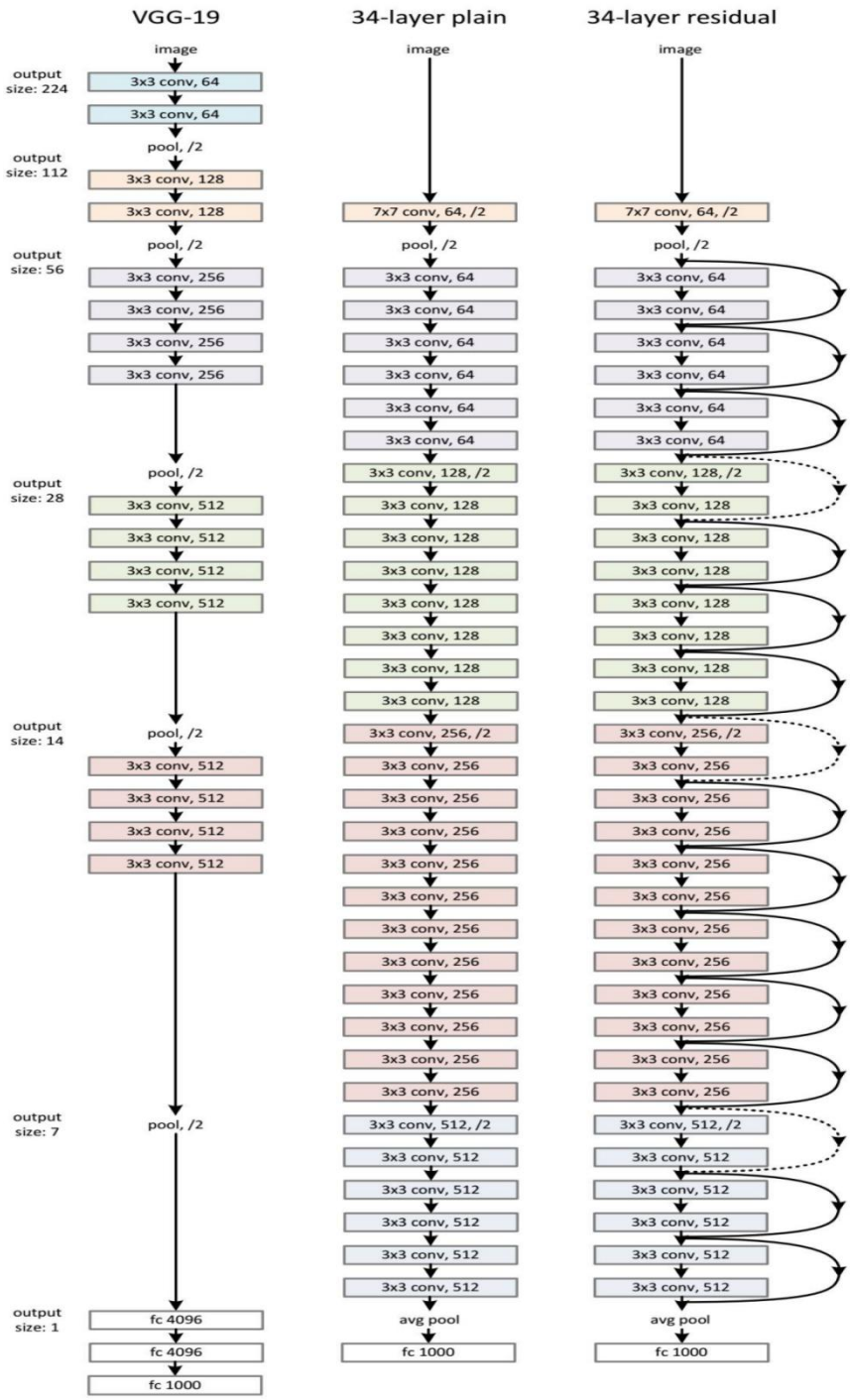


Figure 6 Comparison between ResNet network and other networks

F. Facial expression classification

The cascade classifier of OpenCV is used to load Haar face detection model and torch is used to load CNN or ResNet model respectively. Turn up the camera to enter the shooting state, continue to read the captured video, analyze the video frame by frame, and obtain the gray-scale image of each frame to judge whether there is a face in the image. If one or more human faces are detected, locate and extract the face, and do gray-scale and binary processing on the detected image, and input the processed image into the classifier for analysis. All faces in the camera shooting area are locked by the square box, and the probability of showing the expression of the current facial state is displayed. For the computer, when the face is recognized, the face data is expanded and normalized immediately, and the neural network model is used to recognize the expression, predict and classify it, and return it to the user.

3.Experimental results and analysis

The model shows the final running results in visual form. The purple box is used to locate the face, and the key points of the face are depicted at the same time, and the corresponding expression probability is displayed on the

interface. The effect of the system is tested under different conditions.

1) Light intensity is different: as shown in Figure 7. There are certain differences in the accuracy of the same expression under different illumination intensity. When the illumination intensity is good, the accuracy of expression recognition will be improved. When the illumination intensity is weak, although the expression can still be recognized, the accuracy will be reduced.

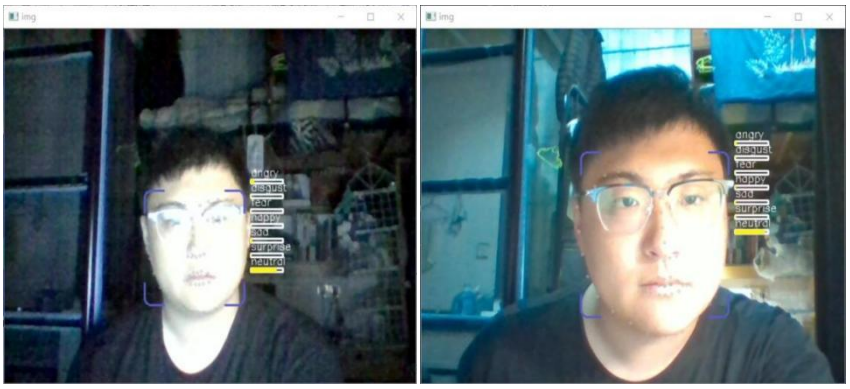


Figure 7 Results of different light intensities

2) Different scenes: the detection accuracy will also affect face recognition in different scenes, as shown in Figure 8. In the indoor test, the accuracy will decline, which is not accurate enough, but in the outdoor test, the accuracy will be more accurate.

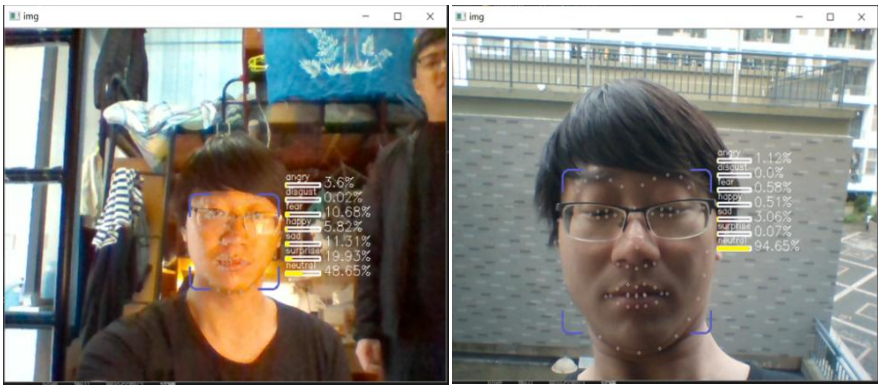


Figure 8 Indoor and outdoor results

3) Different shooting angles: in the same scene and under the same light, test the face from different angles, as shown in Figure 9. It is found that when the face is not in the lens, the face cannot be detected, because the characteristic value of the face cannot be extracted at this time, so it cannot be detected.

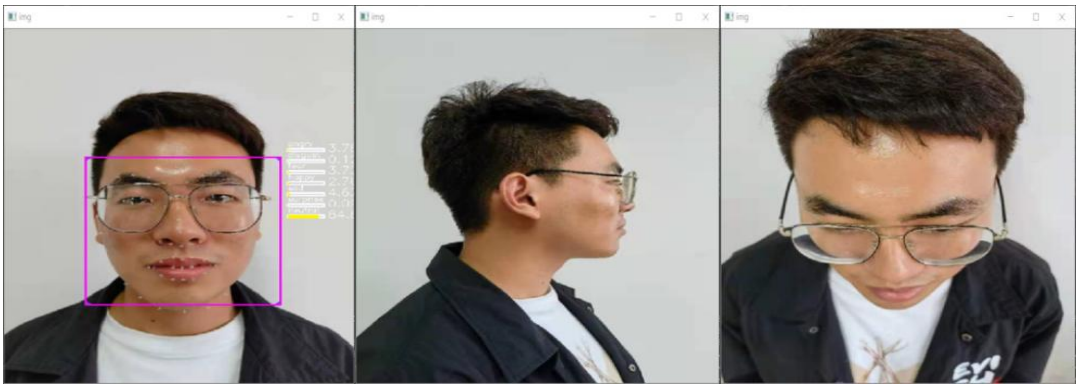


Figure 9 Results at different angles

(4) Gender difference: when using the test set to train the model, the gender of the user is not distinguished, but the gap between the male and female ratio is not large, so the impact of gender on the accuracy is not particularly obvious in the final test. Under the same expression, the

accuracy of male and female testers is almost the same, as shown in Figure 10.

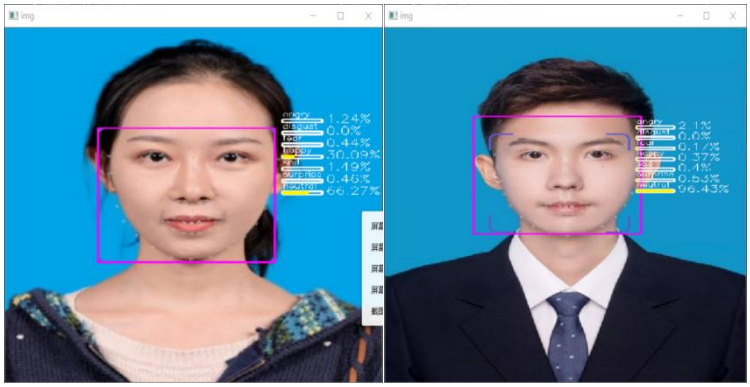


Figure 10 Female and male test results

5) Different ethnic backgrounds: because most of the training sets use pictures of white and black people, and the facial features of different people are also different, the accuracy of the system in detecting white and black people will be higher during the test. As shown in Figure 11.

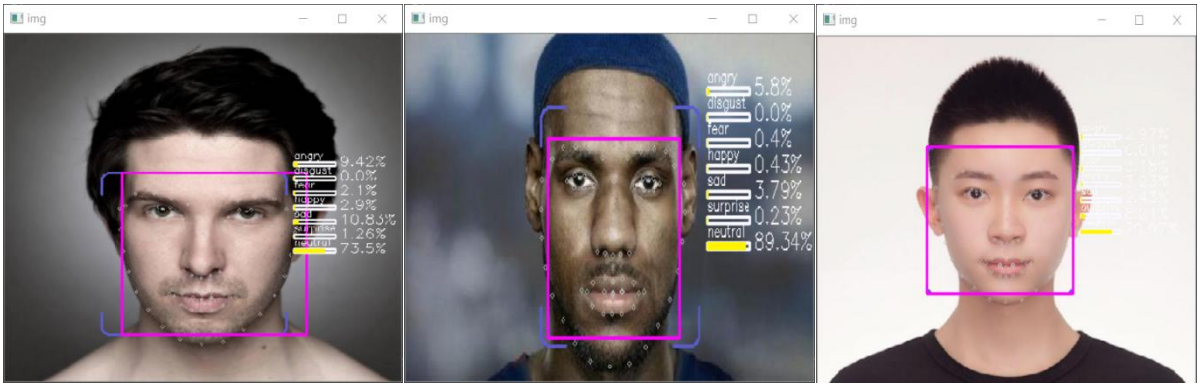


Figure 11 schematic diagram of test results of white, black and yellow people

6) Face occlusion: during the test, it is found that whether to wear glasses has a certain impact on the results. As shown in Figure 12, because facial occlusions such as glasses and masks will hinder the extraction of facial feature values, some key points cannot be detected, and although the pictures in the training set used have pictures of people wearing glasses, the number is not large. These factors together lead to the deviation of the test results.

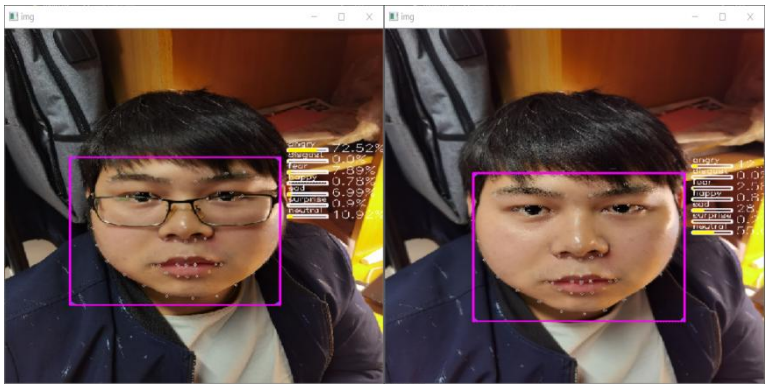


Figure 12 schematic diagram of results with and without glasses

4.CONCLUSION

A facial expression recognition model based on convolutional neural network is developed. The model designs convolutional neural network and carries out facial expression recognition training on Fer2013 and other datasets. Then, the trained model is used to complete the required detection for the real-time face. The experimental results show that the facial expression recognition system developed based on Tensorflow deep

learning framework is feasible and reliable, and can better meet the expression recognition needs of a certain environment.

ACKNOWLEDGMENT

This work was supported in part by the Xuzhou Science and Technology Plan Project (Grant: KC21303), Jiangsu Industry University Research Cooperation Project (Grant:BY2021159), Jiangsu Educational Science "14th five year plan" Project(Grant:C-c/2021/01/65), the Sixth "333 project" of Jiangsu Province, Natural Science Research Projects of Colleges and Universities in Jiangsu Province(Grant: 22KJA520012).

REFERENCES

[1] Bo, C. , et al. "The Hierarchical Beta Process for Convolutional Factor Analysis and Deep Learning." ICML Omnipress, 2020.

[2] Memon, F. A. , et al. "Predicting Actions in Videos and Action-Based Segmentation Using Deep Learning." IEEE Access PP.99(2021):1-1.

[3] Shahid, A. R. , S. Khan , and H. Yan . "Contour and region harmonic features for sub-local facial expression recognition." Journal of Visual Communication and Image Representation 73.2(2020):102949.

[4] Zhang, F. , et al. "A Unified Deep Model for Joint Facial Expression Recognition, Face Synthesis, and Face Alignment." IEEE Transactions on Image Processing 29(2020):6574-6589.

[5] Trimech, I. H. , A. Maalej , and N. Amara . "Facial Expression Recognition Using 3D Points Aware Deep Neural Network." Traitement du Signal 38.2(2021):321-330.

[6] Wang X, Huang J, Zhu J, Yang M, Yang F. "Facial expression recognition with deep learning. " Proceedings of the 10th International Conference on Internet Multimedia Computing and Service. New York, NY, USA: Association for Computing Machinery, 2018

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[8] Levi G and Hassner T. "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. " in Proceedings of the 2015 ACM on international conference on multimodal interaction. 2015. ACM: 503-510

[9] Zhang T, Zheng W, Cui Z, et al., "A deep neural network-driven feature learning method for multi-view facial expression recognition. " IEEE Transactions on Multimedia, 2016.18(12):2528-2536

[10] Yu Z and Zhang C. "Image based static facial expression recognition with multiple deep network learning. " in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. 2015. ACM: 435-442

[11] Tachibana, R. , et al. "Comparative performance of self-supervised 3D-ResNet-GAN for electronic cleansing in single- and dual-energy CT colonography." Imaging Informatics for Healthcare, Research, and Applications 2021.

[12] Ikechukwu, A. V. , et al. "ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images." Global Transitions Proceedings 2. 2(2021):375-381.