Hindawi
Journal of Robotics
Volume 2022, Article ID 4969883, 9 pages
https://doi.org/10.1155/2022/4969883

Research Article

Expression Recognition Using Improved AlexNet Network in Robot Intelligent Interactive System

Yifeng Zhao and Deyun Chen

School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, Heilongjiang 150080, China

Correspondence should be addressed to Deyun Chen; chendeyun@hrbust.edu.cn

Received 2 December 2021; Accepted 15 January 2022; Published 8 February 2022

Academic Editor: Shan Zhong

Aiming at the insufficient feature extraction in the expression feature extraction stage of traditional convolutional neural network and the misclassification of mislabeled samples, an expression recognition and robot intelligent interaction method using deep learning is proposed. First, in image preprocessing, the dimension of the color image is reduced by image gray adjustment to reduce the amount of calculation, the shadow interference is eliminated by the average method, and the image is enhanced by histogram equalization. Second, multichannel convolution is used to replace the single convolution size in the second convolution layer in AlexNet, the Global Average Pooling layer is introduced to replace the fully connected layer, and Batch Normalization is introduced to improve the feature extraction ability of the model and avoid gradient explosion. Finally, the Focal Loss is improved by setting the probability threshold to avoid the impact of mislabeling samples on the classification performance of the model. The experimental results show that the recognition accuracy of the model on FER2013 data set is 98.36%. The effectiveness of the algorithm is verified on the intelligent interactive system of service robot based on expression recognition. Compared with other expression recognition methods, the proposed method can extract more expression features and recognize facial expression more accurately.

1. Introduction

With the progress of society and the continuous development of science and technology, people pay more and more attention to intelligent robots and related fields. The wide application of robots and people's demand for robots promote the development of robot technology in a more intelligent direction [1]. In recent years, according to the requirements of China's national high-tech research and development plan in "13th five-year plan," in order to implement the "made in China 2025," robots will be taken as the key development field, and the overall deployment has been made Moreover, the research on service robot has been favored by many researchers and relevant research institutions, and it has become a hotspot in robot research [2–4]. As a key technology in the field of robotics, human-robot interaction (HRI) is a technology involving multidisciplinary fields, including human-computer interaction technology, artificial intelligence, robotics, and natural language understanding. HRI is a technology to study how to conduct friendly information interaction between human and robot [5–7]. At present, most of the interaction technology between human and robot mainly depends on human giving the specified commands to the robot to complete specific tasks. The robot cannot perceive and understand the user's purpose and intention, nor infer the user's psychological state according to the user's current external state to provide better interactive services. Robots still lack active awareness. In order for service robots to provide users with more convenient and humanized service experience, the traditional human-computer interaction technology must be upgraded so that robots can reasonably infer emotional state according to people's external state performance, so as to perceive and understand people's intentions and emotions and provide relevant services [8–11].

In recent years, the computing power of computer hardware has been greatly improved, including CPU (central processing unit), GPU (graphics processing unit), and artificial intelligence chips specially developed to accelerate the computing power of deep neural network, such as TPU (tensor processing unit) and NPU [12–15]. Because the field of computer vision needs strong computing power, especially the data-driven deep learning algorithm, which benefits from the improvement of computing power, it has achieved breakthrough results. Facial expression recognition technology is an interdisciplinary and comprehensive subject, which mainly includes behavior, psychology, sociology, and computer science [16–19]. Through the research of facial expression and emotional analysis, robots can perceive human psychological state and interpret human behavior intention. Especially in the field of human-computer interaction, intelligent interaction can be realized, so facial expression recognition technology has high scientific research and application value.

With the development of deep learning, more and more researchers pay attention to this hot field [20]. Facial expression recognition integrates many technologies such as machine learning, pattern recognition, and computer vision. Its application scenarios are mainly as follows. (1) In the field of intelligent transportation, it can detect the driver's expression in real time, judge the driver's mental state, and observe whether the driver is driving tired, or whether the driver is in negative emotional states such as anger and anxiety. If the driver is in poor condition, give an alarm and warning immediately. (2) In the field of service robots, with the continuous development of robot technology, service robots will eventually enter ordinary families. Human computer interaction technology based on emotion analysis divides the robot emotion interaction technology into three parts: robot emotion recognition, robot emotion model, and robot emotion feedback. In the whole closed-loop interaction technology, robot emotion perception technology is the basic research field.

As a humanized interaction mode of service robot, facial expression recognition technology collects, detects, and recognizes human facial expression information through the vision system and then analyzes human psychological behavior. In order to improve the accuracy of expression recognition, this paper improves the model on the basis of traditional AlexNet. The main innovations are as follows.

(1) Four different convolution kernels are used for multichannel convolution to extract features to a greater extent. The Global Average Pooling layer is used to replace the fully connected layer to avoid the problem of too many parameters.

(2) By setting the probability threshold, Focal Loss is improved and its confidence is changed, so as to reduce the attention of FL to this kind of samples and improve the classification performance of the model.

2. Related Researches

The research of expression recognition can be mainly divided into traditional feature extraction methods and deep learning methods. The traditional feature extraction mainly depends on the manually designed extractor, which requires a lot of professional knowledge. At the same time, the generalization and robustness are slightly insufficient compared with the deep learning method. The feature extraction of deep learning method updates and iterates the weights through back-propagation and error optimization algorithm to extract deeper and more abstract features in the process of learning a large number of samples. In recent years, many scholars have applied deep learning method to facial expression recognition and achieved good results. Reference [21] proposed a microexpression recognition method based on Wasserstein Generative Adversarial Network, established facial expression recognition network and facial identity recognition network, and improved the accuracy and robustness of facial expression recognition by suppressing intraclass changes. Reference [22] converted the expression image into Local Binary Pattern (LBP) feature map and then used the LBP feature map as the input of CNN for training, which had achieved good results. However, it will lead to low accuracy and insufficient robustness in unknown environment. Reference [23] combined ResNet-50 model and VGG16 model to form a new combined model for facial expression recognition and achieved good results on KDEF data set. In addition to improving the basic model and network structure, many researchers have also studied and improved the loss function. Reference [24] proposed the attention mechanism network SENet (Squeeze-and-Excitation Networks), which automatically obtained the importance of each feature channel through learning and then enhanced the features important to the current task and suppresses the features that are not useful to the current task according to the importance. Reference [25] used the method of spatial attention combined with multichannel connection to improve the convolutional neural network. First, finetune the pretrained model to obtain the feature map, add the spatial attention mechanism to highlight the expression area, and then classify the feature vectors with obvious discrimination. Reference [26] proposed a method to improve the fusion of convolutional neural network and attention mechanism, which integrated the global image features with multiple unobstructed facial regions of interest features, so as to improve the expression ability of regional features. Reference [27] proposed a facial expression recognition algorithm based on local features and deep belief network, extracted the nonuniform illumination invariant features of LSH in facial expression images, and extracted the edge detail features of facial expressions by using Gauss-Laplace operator. Reference [28] proposed a high-frequency edge digital signature feature extraction framework combined with histogram features, which obtains the dynamic digital signature descriptor of human face by projecting edge pixels vertically and horizontally. Digital signature uniquely and completely describes facial expression.

In this paper, the classical AlexNet network model is improved. Aiming at the insufficient feature extraction in the expression feature extraction stage of traditional convolutional neural network and the misclassification of mislabeled samples, an expression recognition and robot intelligent interaction method using deep learning is proposed. The

multichannel convolution is used to replace the single convolution size in the second convolution layer in the classical AlexNet. Global Average Pooling layer is used to replace the fully connected layer, and Batch Normalization operation is introduced to avoid the gradient explosion phenomenon. Focal Loss is improved by setting the probability threshold to avoid the impact of mislabeling samples on the classification performance of the model.

3. Facial Expression Recognition Based on Convolutional Neural Network

3.1. Overall Network Structure. Facial expression recognition is a typical image classification problem in computer vision, which mainly includes four parts. The flow chart of the designed facial expression recognition system is shown in Figure 1. In the facial expression recognition system, the first part is image acquisition, face detection, and image preprocessing. Fast and accurate face detection is carried out on the collected images. After the face region is detected, in order to reduce the influence of inconsistent scale size, image preprocessing is carried out. Then, Mobile Net is used to extract expression features, and expression features are extracted from the face region after image preprocessing. The effectiveness of features directly affects the accuracy of expression classification. After a lot of training and learning, the feature extraction algorithm can effectively extract expression features. The last is expression recognition. Through the final judgment and classification of the extracted expression features, the facial expression recognition is completed and the recognition results are output.

3.2. Image Preprocessing. When recognizing face information, it is necessary to collect face image through camera, and the image may be affected by factors such as illumination change, expression change, and partial shadow. Direct use of the collected image for subsequent processing will increase the amount of calculation and reduce the recognition accuracy. Therefore, it is necessary to preprocess the image before detecting face.

Image gray adjustment is to reduce the dimension of color image. After gray processing, the amount of calculation can be reduced, and the shadow interference can be eliminated by using average method. The average value of the three channels of each pixel point is taken as the processing result, that is:

$$\text{Gray}(i,j) = \frac{R(i,j) + G(i,j) + B(i,j)}{3}. \quad (1)$$

In order to improve the observability of the original image and improve the contrast of the image, histogram equalization is used to enhance the image. Its basic principle is to make the gray value of the image as evenly distributed as possible, so that the number of each gray level is basically the same. The histogram equalization function should meet two requirements.

(1) The value range of $f(x)$ is $[0, L-1]$, where L indicates gray level.

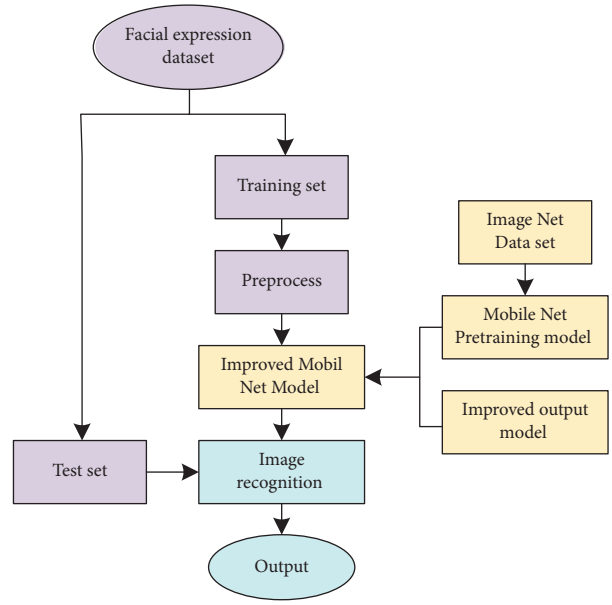

Figure 1: Overall framework of the proposed method.

(2) $f(x)$ is monotonically increasing in $[0, L-1]$.

Equation (2) is often used for histogram equalization in image processing, and this function can meet the above two requirements.

$$y = f(x) = \sum_{i=0}^{x_i} (L-1)\frac{h(x_i)}{n}, \quad (2)$$

where $h(x_i)$ is the vertical height of the gray value x_i in the histogram and n represents the total number of pixels.

3.3. Improved AlexNet Network. The convolution kernel size of convolution layer in AlexNet network is a single value, and the generated feature map features are not diverse, which will lead to insufficient feature extraction. This section makes improvements to the three defects: single size convolution kernel, too many parameters in the fully connected layer, and change of data distribution.

The traditional convolution layer uses a single-sized convolution kernel to convolute the input data for obtaining several feature maps. The multichannel convolution technology can be regarded as an extension of the traditional convolution, adding several convolution kernels of different sizes to a single convolution layer. It will make the generated feature map features more diverse, as shown in Figure 2.

Change the convolution kernel depth of 96 in the first layer to 64, and replace the convolution kernel with a single size of $5 * 5$ in the third layer with four convolution kernels. The sizes of the four convolution kernels are $1 * 1$, $3 * 3$, $5 * 5$, and $7 * 7$, respectively, and set the depth of the four convolution kernels to 64, respectively. Since the depth of the original convolution kernel with a single size of $5 * 5$ is 256, four convolution kernels with different sizes are replaced to combine the four convolution kernels with different sizes. Four different convolution kernels are used

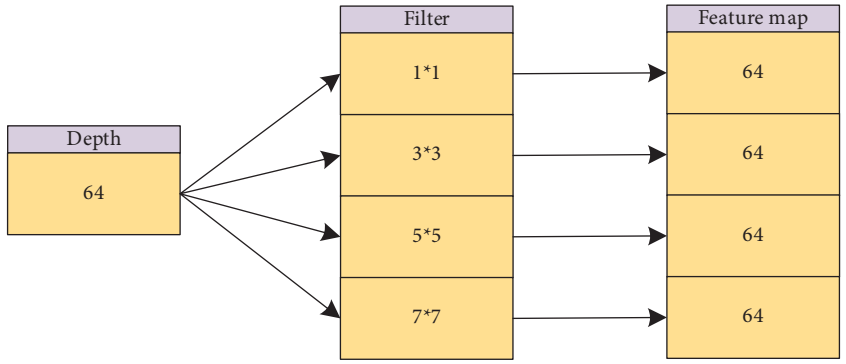

Figure 2: Multiscale convolution kernel operation.

for multichannel convolution to replace the third convolution layer of classical AlexNet network. Compared with the convolution kernel with single size, it strengthens the diversity of convolution kernels and extracts more features.

In order to avoid the problem of too many parameters, the Global Average Pooling (GAP) layer is used to replace the fully connected layer. After the fully connected layer expands the convolution layer into a vector, it is necessary to classify each feature map, and the idea of GAP is to combine the above two processes, as shown in Figure 3. The GAP is different from the average pooling. The average pooling has a filter size. Generally, it averages a subregion of the feature map and then uses the step size to slide the region. However, the GAP has no filter size, which is for the whole feature map.

In order to apply the Global Average Pooling to the classical AlexNet network, first, a convolution layer is added. The kernel size is $3 * 3$, the stripe is 1, the padding is 1, and the depth is 64. At this time, the output is $6 * 6 * 64$. Then, add a maximum pooling layer with filter of $3 * 3$ and stripe of 2, and then add an LRN layer. At this time, the output is $3 * 3 * 64$. Add a convolution layer with kernel size of $3 * 3$, stripe of 1, padding of 1, and depth of 5. At this time, the output is $3 * 3 * 5$. Finally, add GAP, that is, Global Average Pooling for the whole five feature maps, so as to obtain five features.

In the process of network training, the update of the parameters of the previous layer will affect the change of the data distribution of the latter layer. In order to solve the change of data distribution and ensure that the output data of each layer are on the same distribution, the normalized preprocessing is carried out first, and then the normalized data enter the next layer of the network. Therefore, Batch Normalization (BN) is introduced. The mean value of the data processed by BN algorithm is 0 and the variance is 1. The data normalization formula is shown in the following equation:

$$x^{(k)} = \frac{x^{(k)} - E\left[x^{(k)}\right]}{\sqrt{\mathrm{Var}\left[x^{(k)}\right]}}, \tag{3}$$

where $x^{(k)}$ is the input neuron and $E[x^{(k)}]$ is the average value of neurons in training data.

In order to improve its expression ability, learnable parameters γ and β are introduced.

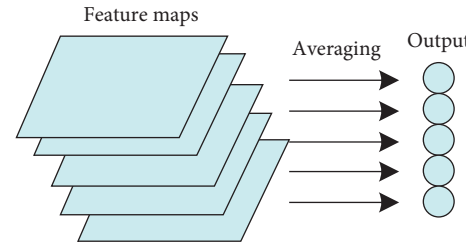

Figure 3: Schematic diagram of Global Average Pooling operation.

$$y^{(k)} = \gamma^{(k)} x^{(k)} + \beta^{(k)} \tag{4}$$

If the data set contains massive data, the calculation will be too complex. A simplified way is used to replace the mean and variance of the whole data set with the mean and variance of a batch, which greatly reduces the amount of calculation.

The calculation formulas of the mean and variance of a batch are shown in the following equations, respectively:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{5}$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2, \tag{6}$$

with the mean and variance, the normalization operation can be carried out. The process is shown in formula (8):

Normalization operation:

$$x_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}}, \tag{7}$$

where μ is the mean and σ is the variance.

The output of Batch Normalization can be obtained according to equations (4) and (8):

$$y_i = \gamma x_i + \beta. \tag{8}$$

3.4. Sample Imbalance and Focal Loss Improvement. Sample imbalance is an important problem in the field of machine learning. This problem will cause minor samples to

drown in major samples and reduce the importance of minor samples. In practical classification problems, major samples are simple and easy to classify, while minor samples account for only a few, which are difficult to classify. Because of simple samples with dominant position, the loss of simple samples should be small, but the number is large, which makes a major contribution to the loss. The samples that are difficult to classify are easy to be ignored by the model. These uncertain samples often lead to overfitting of the network, nonconvergence of the network in the initial stage, and damage to the model learning useful information.

In the expression recognition task, Cross Entropy (CE) is the commonly used loss measurement function, as follows:

$$\mathrm{CE} = -\ln\ p_i, \tag{9}$$

$$p_i = \frac{\exp(a_i)}{\sum_{j=1}^{7}(a_i)}, \tag{10}$$

where p_i is the probability of the label corresponding to the model prediction result and a_i represents the output of the neuron corresponding to the Softmax layer.

Focus Loss (FL) function reduces the weight of the samples easy to classify, making the model focus more on the samples difficult to classify during training. FL formula is as follows:

$$\mathrm{FL} = -\alpha(1 - p_i)\gamma \ln\ p_i. \tag{11}$$

Among them, the role of balance parameter α is to control the weight of unbalanced samples to the total loss and balance the number of samples in different categories. Focusing parameter γ is used to control the weights of the samples easy to classify and the samples difficult to classify.

In the calculation of FL, if there are some errors in the sample label of the data set, or the noise itself is very large, the model will learn the wrong information and reduce the accuracy of the model due to the increase of weight. Aiming at the problem that FL cannot deal with mislabeled samples, the threshold judgment is set through the confidence of samples and real labels, and the mislabeled samples are selected to change their confidence, so as to reduce FL's attention to such samples and improve the classification performance of the model.

$$\begin{cases} p_{\mathrm{top}} = \varepsilon, p_{\mathrm{top}} > c, \text{ and } y_p \neq y_t, \\ \mathrm{FL}, \text{ others}, \end{cases} \tag{12}$$

where p_{top} is the maximum probability that the prediction is true in several types of samples, the superparameter c is the probability threshold, y_t is the real label of the sample, and y_p is the prediction label of the sample.

If the maximum probability p_{top} mapped by the sample is greater than this threshold c, it is considered that the confidence of the sample is very high, and then the sample prediction label is compared with the real label. If the comparison shows that the sample prediction label is equal to its real label, it indicates that the sample is a simple sample with high confidence, and execute Focal Loss. If the sample prediction label is not equal to its real label, it indicates that the sample is a mislabeled sample with high confidence. Set its prediction probability to the minimum; that is, discard the sample. Aiming at the problem of mislabeled samples, this algorithm improves Focal Loss by setting threshold parameters to judge the discrete probability of Softmax output. It selects and discards mislabeled samples and improves the classification performance of the model. The flow chart of the improved Focal Loss algorithm is shown in Figure 4.

4. Design of Intelligent Interactive System for Service Robot Based on Expression Recognition

In the design of service robot human interaction system, if robot emotional interaction with people is achieved, it is a key link for robot to perceive and understand human emotional expression. Human emotional expression mainly includes expression, language, and body action. This paper mainly uses the way of recognizing facial expression to make the robot understand human emotional expression and give corresponding interactive feedback. The intelligent interactive system of service robot designed in this paper mainly consists of expression recognition, voice interaction, action interaction, and so on. The overall design block diagram of the interactive system is shown in Figure 5.

This paper adopts a facial expression recognition method based on deep learning. In the process of expression recognition, the most important step is to extract expression features, so after detecting the face, locating and tracking the key feature points of the face should be done, so as to extract the effective features representing the changes of facial expression. When recognizing face information, it is necessary to preprocess the image before detecting face. In feature extraction, four different convolution kernels are used for multichannel convolution to replace the third convolution layer of classical AlexNet network, which can extract picture features in a higher dimension. Sample imbalance is an important problem in the field of machine learning. Using the Focus Loss function to reduce the weight of the samples easy to classify makes the model focus more on the samples difficult to classify in training. The voice interaction module NAOqiAPIs provides rich application module interfaces, and voice-related application modules are provided by the Audio module. The main modules and functions provided by Audio are shown in Table 1, which implements recognition of the speaker's emotion through speech. ALDialog and ALTextToSpeech modules are used to create a knowledge base to make speech interaction strategies according to different expression categories and convert the text knowledge base into speech. The Motion application module provided by NAOqiAPIs can be used to set various poses and actions of the robot. The main modules and functions provided by Motion module are shown in Table 2.

5. Experiment and Analysis

5.1. Experimental Environment and Data Set. FER2013 facial expression data set consists of 35886 facial expression images, including 28708 training images, 3589 public

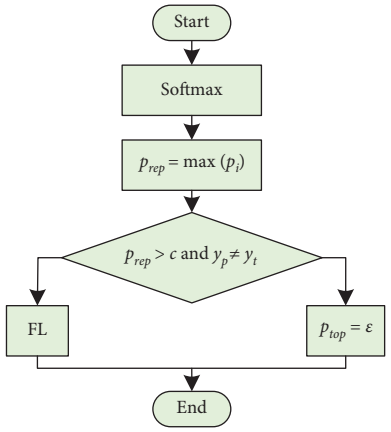

Figure 4: Flow chart of improved FL algorithm.

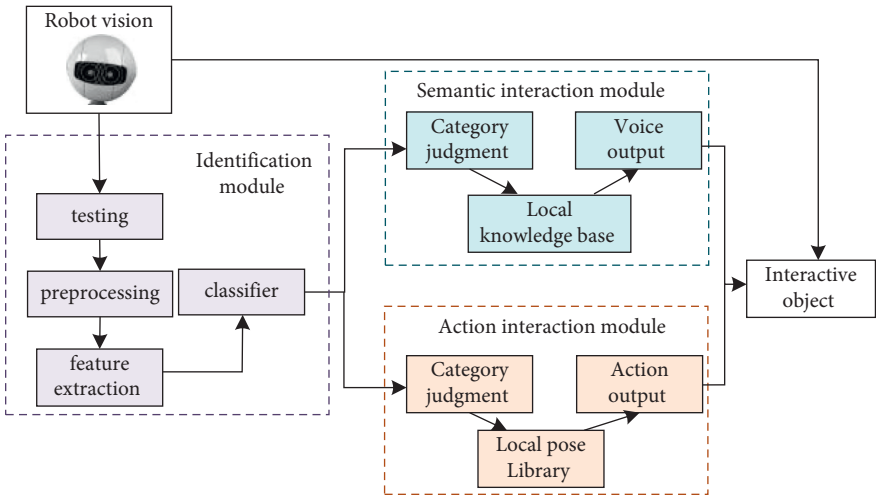

Figure 5: Overall block diagram of interactive system.

Table 1: Audio module function introduction.

AL audio player and AL audio recorder	Audio recording and playback
AL sound detection and AL sound localization	Sound detection and student source location
AL speech recognition	Speech recognition
AL text to speech	Text to speech
AL dialog	Create a basic knowledge base for communication dialogue
AL voice emotion analysis	Speaker's emotion through speech recognition

Table 2: Motion module function introduction.

AL robot posture	Provides a predefined pose for the robot
AL navigation	Make the robot move safely and stop when an obstacle is detected
AL motion	Write your own way of movement according to different methods

validation images, and 3589 private validation images. Each image is a grayscale image with fixed size of 48 ∗ 48. There are 7 kinds of expressions in the data set, which, respectively, correspond to digital labels 0–6. The labels corresponding to specific expressions are as follows: 0: Angry, 1: Disgust, 2: Fear, 3: Happy, 4: Sad, 5: Surprised, and 6: Neutral. Figure 6 shows an example of FER2013 expression database.

5.2. Curve of Accuracy and Loss Value. The batch size of the training network is set to 64, the epoch of training is set to 5000 steps, the image size of facial expression is 48 ∗ 48, and the starting value of learning rate is 0.001. The network is optimized by Adam optimizer.

Figure 7 is the accuracy change diagram of the facial expression experiment based on improved AlexNet network, and Figure 8 is the loss change diagram of the facial

Figure 6: Example of FER2013 expression database.

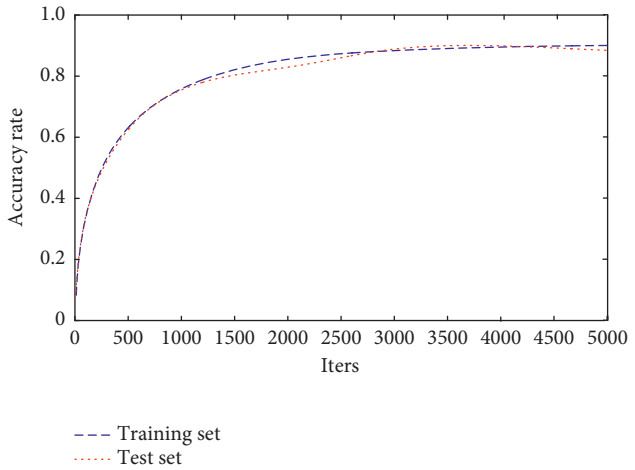

Figure 7: Experimental accuracy change diagram of facial expression recognition based on improved AlexNet network.

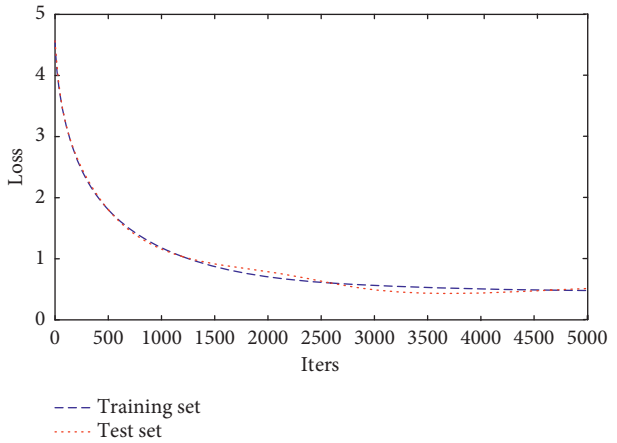

Figure 8: Experimental loss change diagram of facial expression recognition based on improved AlexNet network.

expression experiment based on improved AlexNet network. The experiment took about 13 hours. As can be seen from Figure 7, the curve changes obviously before step 2000, and the accuracy curve of the experiment rises rapidly. After step 2000, the change is slow and finally tends to converge. In step 3000, the accuracy rate of the training set is 90.26% and that of the test set is 89.30%. As shown in Figure 8, the value of the loss function decreases gradually with the increase of the number of training times. After reaching a certain number of training times, the loss value tends to be constant.

5.3. Identification Accuracy Comparison and Confusion Matrix. In order to prove the accuracy of the proposed method, [27] and [28] are selected for comparison. The comparison result is shown in Table 3. The accuracy of the proposed algorithm reaches 98.36%, while the accuracy of the comparative algorithm in [27, 28] reaches 96.69% and 97.17%, respectively. It is proved that the accuracy rate of the proposed method in this paper is better than the comparison algorithm on FER2013 data set. References [27, 28] only extract the edge detail features of facial expression, but lack feature processing, so the accuracy is low. The proposed method introduces the multichannel convolution technology and replaces the classical single-sized convolution kernel with four different-sized convolution kernels of $1 * 1$, $3 * 3$, $5 * 5$, and $7 * 7$. Compared with the single-sized convolution kernel, multichannel convolution strengthens the diversity of convolution kernel and can extract more features. In addition, Global Average Pooling and Batch Normalization are added. Therefore, the accuracy of facial expression recognition has been greatly improved by the improved AlexNet network.

In order to further verify the algorithm, the confusion matrix is drawn according to the experimental results on FER2013 data set. The column represents the predicted class, the row represents the real class, the diagonal value is the prediction accuracy of this class, and the rest is the probability of prediction error. It can be seen from the confusion matrix in Figure 9 that the classification results of the algorithm in this paper are evenly distributed, and all kinds of expression samples are more likely to be classified into their classes. Comparing the accuracy rates of seven kinds of expressions, it is found that the accuracy rates of Sad, Happy,

Table 3: Comparison of different algorithms.

Model	Accuracy rate (%)
The proposed method	98.36
Reference [27]	96.69
Reference [28]	97.17

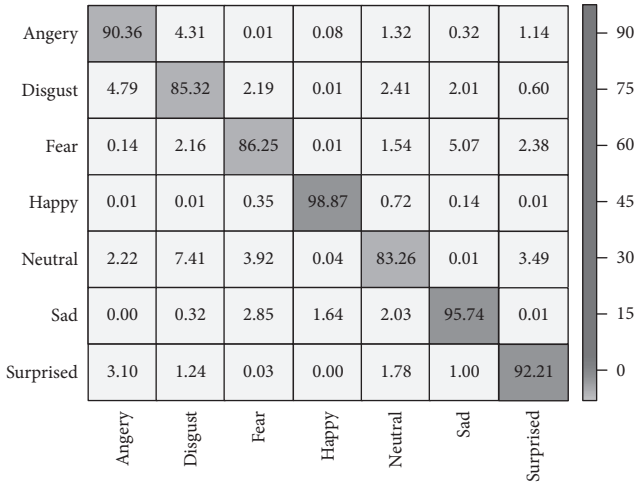

Figure 9: Confusion matrix of FER2013.

and Surprised are higher, the accuracy rate of Neutral is the lowest, and they are easy to be disturbed by Disgust expressions. It also verifies the similarity of the feature map of the two expressions. The feature separation is not obvious in pixel space, and it is also in line with human cognition. This is because this paper improves the Focal Loss by setting the probability threshold to change its confidence, so as to reduce FL's attention to this kind of samples and improve the classification performance of the model.

5.4. Interaction Effect with Robot. After the Facial Expression Rec instruction box is executed, the Switch Case instruction box distributes the expression recognition results to the eight Animated Say instruction box corresponding to the eight expression classes. The Animated Say instruction box stores predefined interaction languages and actions. For example, the expression class of the recognition interaction object is Happy, and the voice text stored in the instruction box is "You look happy, can you share it with me." The robot will make an action while talking. The expression class is Surprised, and the voice text stored in the instruction box is "Why are you so surprised? What's wrong with me." Similarly, the robot makes an action while talking. In the process of robot local test, the processing time of 200 facial expressions by the system is shown in Figure 10. Because the processor performance of the robot is worse than that of the PC, the average time for expression recognition is about 1.24 s, which is worse than that of the PC. Although the robot can detect and track the face of interactive object, the illumination of the face photographed by the camera changes greatly with the rotation of the face, which is a main reason for incorrect expression recognition.

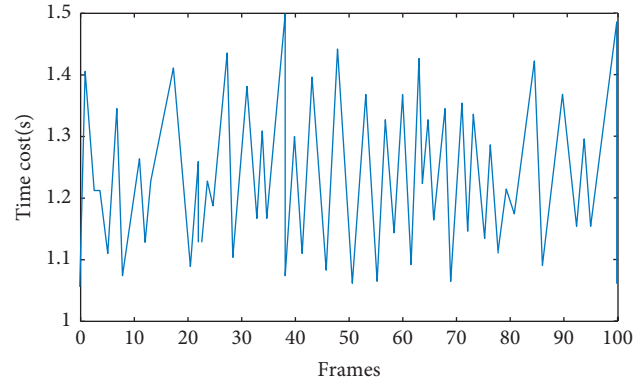

Figure 10: System running time on Robot.

6. Conclusion

Based on the above statement, this paper proposes an expression recognition and robot intelligent interaction method using deep learning. The proposed method uses four different convolution kernels for multichannel convolution to implement feature extraction to a greater extent, replaces fully connected layer with Global Average Pooling layer to avoid the problem of too many parameters, and improves Focal Loss by changing its confidence and improves the classification performance of the model. The experimental results show that the model improves the accuracy and can recognize facial expression more accurately.

In the design of robot intelligent interaction system, in order to make the robot interact with people more accurately, context analysis can be carried out in the future, so that the robot can provide more suitable interactive feedback information according to the context.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of Heilongjiang Province (no. LH2021F029) and Higher Education Reform Project of Heilongjiang Province (no. SJGY20200311).

References

[1] S. Kwak, K. San, and J. Choi, "Can robots be sold? The effects of robot designs on the consumers' acceptance of robots," in *Proceedings of the 2014 9th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 220-221, IEEE, Sapporo, Japan, March 2014.

[2] E. Kwon and G. J. Kim, "Humanoid robot vs. projector robot: exploring an indirect approach to human robot interaction,"

in *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 157-158, Osaka, Japan, March 2010.

[3] K. Haring, D. Silvera, and T. Takahashi, "How people perceive different robot types: a direct comparison of an android, humanoid, and non-biomimetic robot," in *Proceedings of the 2016 8th International Conference on Knowledge and Smart Technology*, pp. 265–270, IEEE, Chiang Mai, Thailand, February 2016.

[4] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib, "Progress and prospects of the human-robot collaboration," *Autonomous Robots*, vol. 42, no. 5, pp. 957–975, 2018.

[5] N. Wang, Y. Zeng, and J. Geng, "A brief review on safety strategies of physical human-robot interaction," in *Proceedings of the ITM Web of Conferences*, pp. 155–163, EDP Sciences, Moscow, Russia, November 2019.

[6] R. Liu and X. Zhang, "A review of methodologies for natural-language-facilitated human–robot cooperation," *International Journal of Advanced Robotic Systems*, vol. 16, no. 3, pp. 1729–1736, 2019.

[7] A. R. Wagner, P. Robinette, and A. Howard, "Modeling the human-robot trust phenomenon," *ACM Transactions on Interactive Intelligent Systems*, vol. 8, no. 4, pp. 1–24, 2018.

[8] M. Awais, M. Y. Saeed, and M. S. A. Malik, "Intention based comparative analysis of human-robot interaction," *IEEE Access*, vol. 12, no. 5, pp. 1996–2008, 2020.

[9] N. Mavridis, "A review of verbal and non-verbal human-robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, no. 1, pp. 165–175, 2014.

[10] T. Xue, W. Wang, and J. Ma, "Progress and prospects of multimodal fusion methods in physical human–robot interaction: a review," *IEEE Sensors Journal*, vol. 2, no. 9, pp. 10–19, 2020.

[11] X. Liu, S. Ge, and F. Zhao, "A dynamic behavior control framework for physical human-robot interaction," *Journal of Intelligent and Robotic Systems*, vol. 101, no. 1, pp. 1–18, 2021.

[12] D.-X. Zhou, "Deep distributed convolutional neural networks: Universality," *Analysis and Applications*, vol. 16, no. 6, pp. 895–919, 2018.

[13] H. Patel, A. Thakkar, M. Pandya, and K. Makwana, "Neural network with deep learning architectures," *Journal of Information and Optimization Sciences*, vol. 39, no. 1, pp. 31–38, 2018.

[14] W. Luo, J. Lu, X. Li, L. Chen, and K. Liu, "Rethinking motivation of deep neural architectures," *IEEE Circuits and Systems Magazine*, vol. 20, no. 4, pp. 65–76, 2020.

[15] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, no. 6, pp. 53040–53065, 2019.

[16] X. Zhao and S. Zhang, "A review on facial expression recognition: feature extraction and classification," *IETE Technical Review*, vol. 33, no. 5, pp. 505–517, 2016.

[17] N. V. K. Medathati, H. Neumann, G. S. Masson, and P. Kornprobst, "Bio-inspired computer vision: towards a synergistic approach of artificial and biological vision," *Computer Vision and Image Understanding*, vol. 150, no. 2, pp. 1–30, 2016.

[18] S. Alyamkin, M. Ardi, A. C. Berg et al., "Low-power computer vision: status, challenges, and opportunities," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 411–421, 2019.

[19] A. Voulodimos, N. Doulamis, and A. Doulamis, "Deep learning for computer vision: a brief review," *Computational Intelligence and Neuroscience*, vol. 8, no. 5, pp. 126–175, 2018.

[20] A. Gupta, "Current research opportunities for image processing and computer vision," *Computer Science*, vol. 20, no. 3, pp. 15–26, 2019.

[21] C. Xu, Y. Cui, Y. Zhang, P. Gao, and J. Xu, "Person-independent facial expression recognition method based on improved Wasserstein generative adversarial networks in combination with identity aware," *Multimedia Systems*, vol. 26, no. 1, pp. 53–61, 2020.

[22] H. Zhang, Z. Qu, and L. Yuan, "A face recognition method based on LBP feature for CNN," in *Proceedings of the 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference*, pp. 544–547, IEEE, Chongqing, China, March 2017.

[23] P. Dhankhar, "ResNet-50 and VGG-16 for recognizing facial emotions," *International Journal of Innovations in Engineering and Technology*, vol. 13, no. 4, pp. 126–130, 2019.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *IEEE*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[25] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recognition*, vol. 92, no. 2, pp. 177–191, 2019.

[26] Y. Li, J. Zeng, and S. Shan, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.

[27] Y. Yaermaimaiti, "Facial expression recognition based on local feature and deep belief network," *Journal of Decision Systems*, vol. 3, no. 15, pp. 1–13, 2021.

[28] K. Talele and K. Tuckley, "Facial expression recognition using digital signature feature descriptor," *Signal, Image and Video Processing*, vol. 14, no. 4, pp. 701–709, 2020.