IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 October 2023, accepted 11 November 2023, date of publication 16 November 2023,
date of current version 22 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3333381

RESEARCH ARTICLE

Convolutional Channel Attentional Facial Expression Recognition Network and Its Application in Human–Computer Interaction

JING PU[1] AND XINXIN NIE[2]
[1]School of Arts and Media, Sichuan Agricultural University, Yaan 625014, China
[2]College of Literature and Media, Chengdu Jincheng College, Chengdu 611731, China

Corresponding author: Jing Pu (pujing325@163.com)

ABSTRACT Currently, the use of robots has altered the way people live and their lifestyles. To realize a human-computer interaction system based on robots' comprehension of human emotions, this study chooses facial expressions as the research object and constructs a facial expression recognition model based on convolutional neural networks and channel attention. To deploy the recognition model to portable devices, a depth-separable convolution filter pruning algorithm based on principal component analysis is constructed. This model applies principal component analysis to reduce the dimensionality of similar filter matrices obtained by calculating geometric medians to prevent gradient explosion. The proposed algorithm for facial expression recognition in this study achieves a 99% recognition accuracy with an average of 80.39%, while using the least number of parameters among the compared algorithms. The verification experiment results of lightweight network show that when the model depth is 56 and the pruning rate is 40%, the correct rate of facial expression recognition of the network model based on the pruning strategy proposed in the study is 93.24%. When dimension is 0.85, the accuracy of model classification is the highest. The algorithm presented in this study exhibits excellent performance when recognizing facial expressions, demonstrating notable robustness and efficiency. The pruning strategy proposed in this study has a good model acceleration effect. It can not only reduce the memory occupied by about 41% of parameters, but also improve the classification accuracy, running time and calculation cost after pruning to a certain extent. The study applied deep separable convolution and PCA techniques to improve and reduce the dimensionality of the convolutional layer in the ResNet network structure, and improved the comparable filter matrix generated during the filter pruning process.

INDEX TERMS Channel attention, depth-separable convolution, facial expression recognition, human–computer interaction, filter pruning, principal component analysis.

I. INTRODUCTION

Recently, the research community has shown increasing interest in artificial intelligence. Human-Computer Interaction (HCI) is the technologies in which computers and humans communicate, understand, transmit information and operate with each other through certain communication methods [1]. Emotional interaction in human-computer interaction technology cannot be separated from Facial Expression Recognition (FER). FER is a technology to recognize human emotions by extracting facial features [2], [3]. FER methods are mainly split into traditional methods and Deep Learning (DL) methods. The former mainly relies on human experience for feature extraction, which requires high personal ability and is greatly affected by subjective factors [4]. The feature extraction and classification based on DL method is automatic and does not need to be designed separately. The output can be generated by simply inputting facial expression images [5], [6]. In DL's FER methods, Convolutional Neural Networks (CNNs) are usually used. CNN can extract facial

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Jin.

expression features that are not easily perceived by humans, and can represent more abstract expression features. Different from traditional expression recognition methods, this method realizes end-to-end communication through repeated training. It is no longer necessary to extract features manually, but only need to optimize and update parameters iteratively through back propagation algorithm and cost function [7]. CNN-based methods and their improvements offer considerable advantages for feature extraction on large-scale datasets. Nowadays, networks typically refer to CNN models with depths ranging from several layers to hundreds of layers. As the model's depth increases, CNN acquires robust learning capabilities. However, with the increasing demand for performance and complexity in network depth and recognition tasks, the demand for memory and hardware devices also increases, limiting the applications available. In addition, to ensure learning ability, well-designed neural networks often exhibit excessive parameterization, leading to wastage of resources. In order to optimize network models, save software and hardware resources, and maintain learning ability while streamlining the network, the concept of model compression has emerged. Many methods exist for model compression, including low rank decomposition, quantization, pruning, and lightweight network structures. Among them, pruning, as one of the most effective methods for compressing models, has been widely studied and applied. Therefore, starting from FER, this study proposed a new convolutional channel attention module, constructed a convolutional channel attention network model and pruned the branches. A deep separable convolutional filter pruning algorithm based on PCA has been proposed. Its primary innovation optimizes the filter-pruning algorithm based on geometric median. This is achieved through the use of deep separable convolution and PCA to enhance and decrease the dimensionality of the convolutional layer in the ResNet network structure. Additionally, this process improves the comparable filter matrix created during filter pruning. The research aims to complete the human-computer interaction system based on FER, and provide a new model for realizing natural human-computer interaction. The first section of the article is an introduction, which introduces the background, purpose, methods, and structure of the research. The second section summarizes the current research status of FER and model compression both domestically and internationally. The third section provides a detailed description of the FER method based on convolutional channel attention networks. The fourth section presents experiments to verify the effectiveness and superiority of the proposed expression recognition method and channel pruning algorithm, along with the experimental results. The fourth section introduces the main content, experimental results, shortcomings, and future prospects of this study.

II. RELATED WORK

FER is a method to obtain facial characteristics on the basis of face recognition, which can be divided into traditional method and DL method. Traditional recognition models rely on manually extracted features in the early stages, which are subject to weaknesses and limitations. Human factors also have a significant influence on these models. Therefore, the method based on DL is commonly used currently, which mainly includes expression feature extraction and expression classification. Li et al. combined attention mechanism and local binary pattern feature extraction method to make the Neural network (NN) focus on more effective feature information. The feasibility and effectiveness of this method had been verified on self-made data sets and 4 commonly used data sets [8]. Wang et al. proposed a NN that can adaptive capture face regions, and used regional bias loss to improve the weight of important regions to achieve accurate recognition of facial expressions under occlusion and posture changes [9]. Borgalli et al. designed three different architectures based on deep CNN for face expression recognition across data sets, and obtained an accuracy rate of 50.96%∼68.81% on CK+ data sets [10]. Jiang et al. introduced an expression recognition model that utilized a local feature clustering loss function to reduce interference resulting from image differences [11].To realize face recognition in small sample data sets, Li et al. constructed a recognition model based on simplified CNN to accurately extract effective face features. The model achieved a mean recognition precision rate of over 97% on two frequently used datasets [12].

To facilitate the smooth deployment of deep CNN on mobile devices with limited resources, many model compression schemes have been constructed. The concept of network pruning involves removing the less important or redundant parts of the NN. Although the learning ability of deep CNN is obvious, in fact, after the training process, not all parameters or structures are useful for the training results. Currently, pruning can be broadly classified as unstructured pruning and structured pruning. Zhao et al. proposed a deep reinforcement learning method based on multi-layer sparse coding and non-convex regular pruning. This method promoted strong sparsity through non-convex regularization, and then removed weights of low importance, achieving more than 80% parameter reduction [13]. Shen et al. designed a new pruning algorithm for data classification algorithms, utilizing past base classifiers that contribute to the current classification to carry out forward supplementary integration, so as to complete the sequential integrated pruning of multiple classifiers [14]. Chen et al. calculated the importance of filters by using the discrete cosine transform to preserve the more important filters in the feature map. Through this pruning technique, the parameters in the ResNet-50 network decreased by 70%, with a mere 3.84% reduction in accuracy [15]. Wang et al. proposed the use of loss improvement to evaluate the weak learner in the gradient propulsion model, and proposed two methods of simple pruning and statistical pruning as well as a dynamic switching scheme to optimize pruning strategies [16]. Zheng et al., aiming at efficient network generation methods, extracted architecture samples from joint classification distribution and conducted dynamic

pruning of search space. This method achieved up to 97% accuracy and improved the speed of the search [17].

Based on the research status of expression recognition and model compression at home and abroad, it can be found that most of the current face recognition methods are DL methods. These methods automatically extract and classify features, providing greater convenience and intelligence. In the face expression recognition based on DL, CNN and its improved method are the most commonly used. It can extract more detailed expression features, and can represent more abstract expression features. To compress the deep CNN model so that it can be deployed on small devices with limited resources, it is essential to research model pruning algorithms in the field of DL. Considering the difference in importance of feature channels, this study uses the channel attention module combined with deep separable convolution to raise the accuracy of face recognition, and uses the principle of Principal Component Analysis (PCA) to prure the filter. To complete the deployment of recognition model in human-computer interaction system.

III. FER METHOD BASED ON CONVOLUTIONAL CHANNEL ATTENTION NETWORK

Currently, the classic DL networks used for FERface several issues, including large network parameters, low recognition accuracy, slow processing time, complex model training and large computation. To effectively solve the above problems and optimize the accuracy of FER, this study proposed a FER algorithm with convolutional channel attention. To enhance its suitability for human-computer interaction, this paper proposes a PCA-based algorithm for depth-separable convolution filter pruning.

A. IMPROVEMENT OF CNN BASED ON CONVOLUTIONAL CHANNEL ATTENTION XRS MODULE

To pay more attention to the characteristics of each channel when training FER Networks, we combine the deep separable convolutional network mechanism in Xception network, the residual mechanism proposed by ResNet network and the Squeeze and Excitation Networks (SEnet) attention module. A Convolutional channel attention (Xception-ResNet-SEnet (XRS) network module is designed. Depth-separable convolution is utilized to divide the ordinary convolution into two steps: deep convolution and point-by-point convolution [18], [19]. When deep convolution is used for spatial feature extraction, only one two-dimensional convolution kernel is used in each input channel. This approach differs from the traditional convolution method that employs three-dimensional kernels and consequently reduces the number of required kernels without impacting the number of channels. Therefore, in subsequent point-by-point convolution, 1×1.3D convolution kernel set is used to extract channel features, and the quantity of channels in the output feature map is changed meanwhile. The detailed structure of a depth-separable convolution is shown in Figure 1.

SEnet is a module based on the channel attention mechanism, which can emphasize useful features in the image while suppressing non-significant features by adjusting the channel weights of feature maps. To obtain the weights of each channel, the Block unit of SEnet first uses the traditional convolution feature graph U_c as the input to carry out the Squeeze process. The essence of Squeeze process is to compress each feature channel in spatial dimension $H \times W$ by using the global average pooling function, as shown in Equation (1).

$$Z = F_{sq}(U_C) = \frac{1}{H \times W} \sum_{H}^{i=1} \sum_{W}^{j=1} U_C(i, j) \tag{1}$$

In Equation (1), Z is the weight after compression. The number of channels in the feature diagram U_C denotes C; The length and width are H and W. The compression operation is performed after the compression operation. The compression weight of $1 \times 1 \times C$ is fully connected by two nonlinear layers. The mathematical expression is shown in Equation (2).

$$S_C = F_{ex}(Z, W) \tag{2}$$

In Equation (2), S_C is the adjusted weight. Finally, the weight S_C is fused with the original input U_C to achieve the adjustment of channel importance, which is calculated as shown in Equation (3).

$$\bar{X} = F_{scale}(U_C, S_C) = U_C \otimes S_C \tag{3}$$

In Equation (3), S_C represents element-by-element multiplication; $\bar{X}$ represents the features obtained after processing by SEnet. SE as a submodule needs to be used in conjunction with other models. Figure 2 illustrates the resulting XRS module structure.

In Figure 2, the XRS module uses a deep separable convolutional network instead of a normal convolutional network to widen the network and reduce the parameters and operation costs. Then the channel attention module SEnet is introduced into the output of the separable Convolution Layer (CL), and the weights of the output channels are reassigned according to the importance. The sigmoid function normalizes the c channels' weights, which fall within the range of 0-1. By performing a point multiplication operation between the normalized channel weight and the feature map tensor, the channel weight of the feature map tensor has been reassigned. Finally, in order to suppress the phenomenon of gradient vanishing, the residual mechanism in the ResNet network was introduced. That is to sum the feature tensor with the feature map tensor that redistributes weights, and obtain the final feature map tensor. By introducing XRS module into the CNN, the convolutional channel attention network model for face recognition is obtained. Figure 3 illustrates the structure of the network model.

In Figure 3, the network consists of three main parts. The network is divided into three main parts. The stacking of CL and Pooling Layer extracts the primary texture characteristics. The XRS module reassigns weights to the extracted

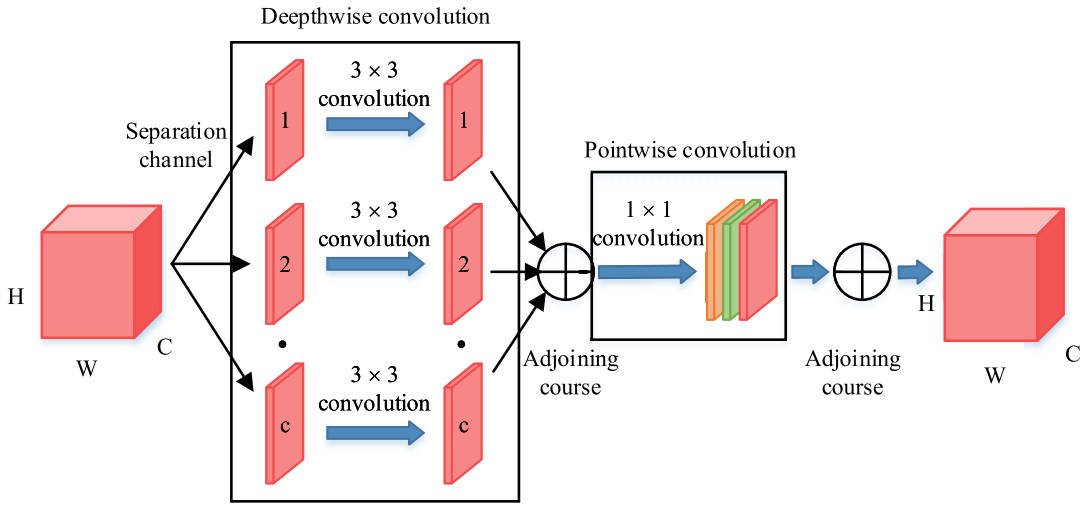


FIGURE 1. Schematic diagram of deep separable convolution.

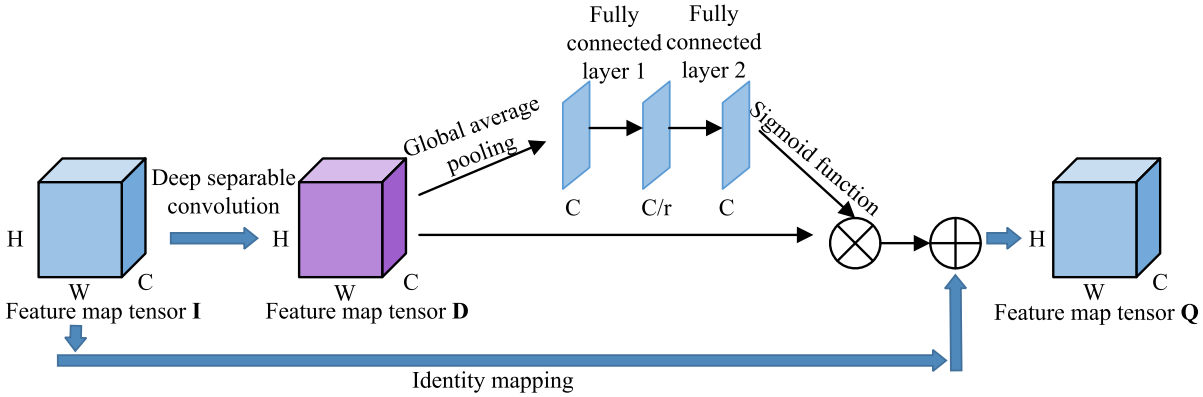


FIGURE 2. Schematic diagram of deep separable convolution.

abstract features and basic texture features, assigning more weights to the main facial expression features to reduce the weight of irrelevant information. The fully connected layer classifies the recognized feature information graphs to enhance the precision. Batch normalization and ReLU nonlinear activation functions are added to each layer of the convolutional structure. ReLU function's expression is shown in Equation (4).

$$f(x) = relu(x) = \max(x, 0) \tag{4}$$

Batch normalization adjusts the distribution of the data so that the output of each layer is normalized to a distribution with a mean of 0 and a variance of 1. This will avoid the problem of gradient explosion during training, reduce the dependence on parameter initialization, and make training faster and more efficient. Finally, a fully connected layer is applied to convert the feature tensor data into vectors, and the Dropout layer randomly drops some data to reduce overfitting. Meanwhile, the facial expressions are classified using the softmax activation function. 7 vectors are input to represent the probabilities of 7 facial expressions respectively.

A loss function that is appropriate can improve the facial expression recognition performance of the model. Due to the blurry background of facial expression images and the high similarity of backgrounds between different classes, the inter class distance of facial expression images is small, and the cross entropy loss function is not capable of constraining the inter-class distance [20]. The extracted facial expression features are prone to confusion between classes. Therefore, this study added ArcFace loss to the loss function of the network. In the Angle domain, the Angle penalty term is added for different classes, the distance between classes is increased, and the intra-class distance is further condensed. To accelerate the speed of network convergence, standard cross entropy loss $Loss_{CrossEntropy}$ is combined with ArcFace loss $Loss_{arcface}$, and the total loss $Loss$ is shown in Equation (5).

$$Loss = Loss_{CrossEntropy} + Loss_{arcface} \tag{5}$$

The ArcFace loss is the introduction of an additional Angle margin penalty in Softmax loss. Softmax loss is good for closed set problems, but it is not good for open set problems like expression recognition, and it lacks of constraints on

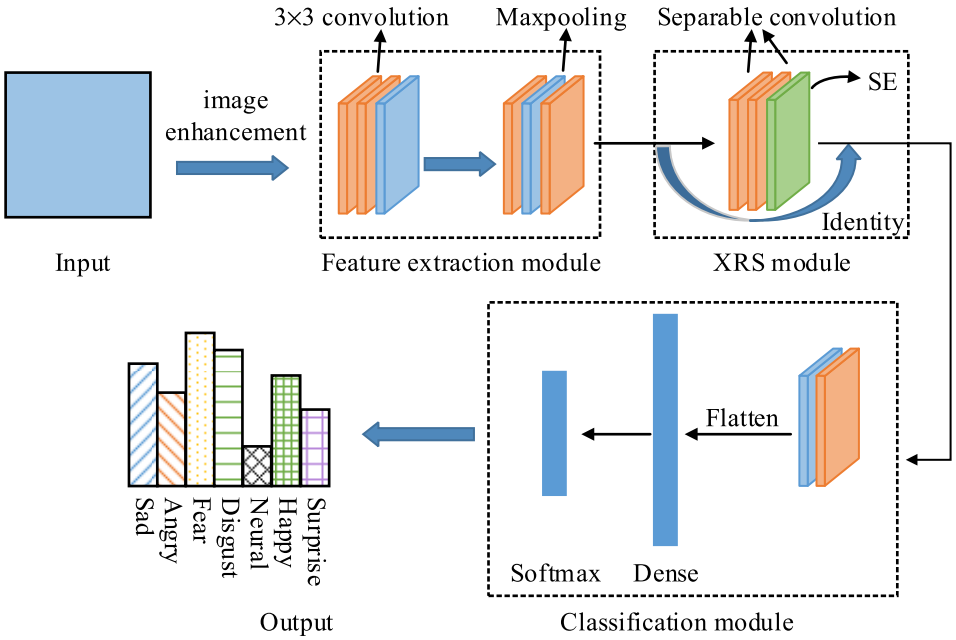


FIGURE 3. Schematic diagram of improved convolutional neural network based on XRS module.

intra-class and inter-class distances. Equation (6) shoes the expression of ArcFace loss.

$$Loss_{arcface} = -\log \frac{e^{s\cos(\theta_{R_i}+m)}}{e^{s\cos(\theta_{R_i}+m)} + \sum_{j=1, j\neq R_j}^{N} e^{s\cos\theta_j}} \tag{6}$$

In Equation (6), the depth feature of sample j actually belongs to class R_j; It's the Angle between the embedded features. The penalty term m of the hypersphere radius s and Angle θ are the two hyperparameters introduced by the ArcFace loss. By adjusting the hyperspherical radius s, ArcFace loss enables embedding a multitude of feature categories to satisfy the requirements of extensive facial expression training. Through the constraint effect of Angle penalty term m, the distribution of expression feature vectors is more concentrated in the weight center, and the inter-class difference and intra-class convergence are enhanced. The NN model is optimized using the optimizer, and the optimization process in DL is shown in Figure 4.

In DL, the dependence of update parameters on the objective function varies. If only a unified global learning rate is set, for the learning process with a large gradient, the convergence speed will be slow if the step size is too long. When the learning rate is set too big, the optimized parameters will oscillate and become unstable [21], [22]. Therefore, some optimization algorithms should be selected in the training process to realize the self-adaptation of the learning rate. Classical optimization algorithms include AdaGrad algorithm, the Root Mean Square prop (RMSprop) algorithm and the Adaptive moment estimation (Adam) algorithm. The

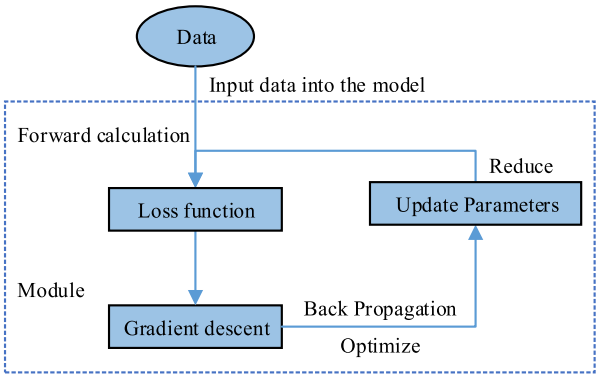


FIGURE 4. Schematic diagram of the optimization process of deep learning networks.

selection of a specific optimizer is determined through experimentation.

B. CONVOLUTION CHANNEL ATTENTION RECOGNITION COMPRESSION MODEL FOR HUMANCOMPUTER INTERACTION SYSTEM

After designing the improved CNN based on the convolutional channel attention XRS module, the FER function can be realized, which makes it possible for the robot to recognize emotions. To achieve harmonious human-computer interaction, this sectionuses the NAO robot as the emotional interactive platform. The features of the NAO robot include rich sensors, video acquisition, and voice output functions. Consequently, this study designs seven basic emotional expression actions and voice dialogues based on the requirements of human-computer interaction. Figure 5 shows

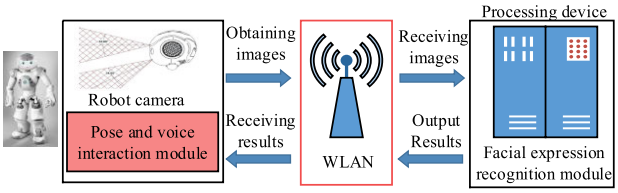


FIGURE 5. Schematic diagram of the architecture of a human-machine interaction system embedded in a FER model.

the architecture of the human-computer interaction system embedded with the FER model.

In the human-machine interaction system shown in Figure 5, the main function of the NAO robot is to collect images and achieve emotional interaction, while the main function of the computer is to preprocess the received images and calculate and output facial expression results through the facial expression recognition network model. The transmission and reception of information between the NAO robot and the computer are carried out through a WLAN wireless network.Today is the era of mobile Internet, smart phones and other portable devices occupy people's lives. To enable the system to be deployed on portable electronic devices, lightweight NN is designed for FER models in this study. The development of lightweight networks has two directions: manual design and model compression.Through model compression, network redundancy is reduced, resulting in a decrease in both time and space complexity of algorithms. Common model compression operations include pruning, quantization, low-rank decomposition and knowledge distillation. The use of depth-separable convolution as an alternative to conventional convolution mentioned earlier is also a model compression method. Changing the convolution mode can effectively speed up network training and increase the depth of layers in deep CNNs. In addition, a depth-separable convolution filter pruning with PCA is constructed. The algorithm uses PCA to reduce the dimensionality of the similar filter matrix generated in the filter pruning. The flow of the algorithm is shown in Figure 6.

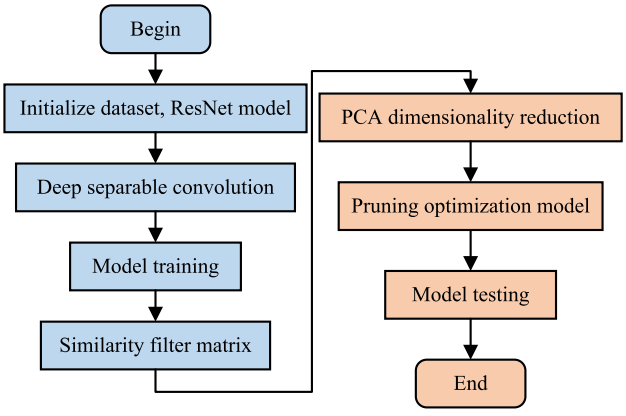


FIGURE 6. Flow chart of deep separable convolutional filter pruning algorithm based on PCA.

PCA regards the similar filter matrix calculated by NN as a matrix X consisting of p samples and q index variables. In this matrix, p samples are represented by the observed values of q indicators in the form shown in Equation (7).

$$X = \begin{bmatrix} x_{11} & \cdot & x_{1q} \\ \cdot & \cdot & \cdot \\ x_{p1} & \cdot & x_{pq} \end{bmatrix} = [x_1, x_2, \cdot, x_q] \tag{7}$$

The covariance matrix $MC_{q\times q}$ of q index variables is calculated as shown in Equation (8).

$$MC_{q\times q} = Cov\,(X)_{q\times q} = E\,(X - EX)\,(X - EX)^T \tag{8}$$

In Equation (8), E denotes the identity matrix. The purpose of PCA is to calculate the index variables, so as to obtain q new indicators, which are independent of each other. The linear combination involved in the calculation is shown in Equation (9).

$$y_i = Xa_i = a_{1i}x_1 + a_{2i}x_2 + \cdot + a_{qi}x_m, i = 1, 2, \cdot, q \tag{9}$$

In Equation (9), a_i is the coefficient of decreasing variance of q new index y_i obtained by linear transformation. The new indicator y_i obtained is the principal component $1, 2, \cdot, q$ of the original indicator. In practical application, if the first h principal component can reflect the original q index variable to the maximum extent, then the h principal component $y_1, y_2, \cdot, y_h$ can be subsequent analysis to achieve the purpose of reducing the filter dimension and easing the gradient explosion. a_i $(i = 1, 2, \cdot, h)$ is both the coefficient of the first h principal component and the eigenvector corresponding to the first h large eigenvalue λ_i $(i = 1, 2, \cdot, h)$ of the covariance matrix. Therefore, the matrix after dimensionality reduction is shown in Equation (10).

$$Y_{p\times h} = X_{p\times q}A_{q\times h} \tag{10}$$

In the process of dimensionality reduction of matrix, the choice of dimensionality reduction h is very important. The appropriate dimensionality reduction is determined by calculating the comprehensive evaluation value. Firstly, the information contribution degree and the sum of the contribution degree of the eigenvalue λ_i $(i = 1, 2, \cdot, h)$ are calculated. For principal component y_i, the degree of its information contribution is calculated as shown in Equation (11).

$$con_i = \lambda_i / \sum_{k=1}^{q} \lambda_k, i = 1, 2, \cdot, q \tag{11}$$

In Equation (11), con_i is the degree of information contribution of principal component y_i. The sum of contribution rates h of the first α_h principal components is calculated as shown in Equation (12).

$$\alpha_h = \sum_{k=1}^{h} \lambda_k / \sum_{k=1}^{q} \lambda_k \tag{12}$$

When $\alpha_h = 0.85, 0.90, 0.95$ is calculated, the first h principal component index can be used to replace the original B index for analysis. Finally, the comprehensive score of

each principal component is carried out, and the calculated comprehensive score value is used to evaluate. The overall score is calculated as shown in Equation (13).

$$Z_{score} = \sum_{i=1}^{h} con_i y_i \tag{13}$$

To minimize the information loss in the filter pruning, the similarity of the filter is calculated, and then the dimensionality of the similar filter is reduced by PCA. The calculation of similarity is optimized by using the cosine method to measure similarity between filters as a cosine value. The cosine similarity is calculated as shown in Equation (14).

$$Sim = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{p} (A_i \times B_i)}{\sqrt{\sum_{i=1}^{p} (A_i)^2} \times \sqrt{\sum_{i=1}^{p} (B_i)^2}} \tag{14}$$

In Equation (14), A and B represent the filter vector of dimension n. A_i and B_i represent the components of vectors A and B. The cosine similarity calculation treats any two filters as two vectors, and the closer their cosine values are to 1, the more similar they are.

IV. APPLICATION TEST OF FER WITH CONVOLUTIONAL CHANNEL ATTENTION NETWORK IN HUMAN-COMPUTER INTERACTION SYSTEM

After establishing the channel attention network, data training is necessary to evaluate its performance of the network. The data sets used in this study are CK+, FER2013 and RAF-BD datasets. The optimal optimizer is first selected through experiments, and then ablation experiments were conducted on the RAF-BD dataset to verify the performance of the XRS network, and common recognition algorithms were used on multiple datasets for identification and comparison. Finally, a single label subset of the RAF-BD dataset is utilized to validate the credibility and consistency of the compact network.

A. FER EFFECT OF ATTENTION MODULE OF CONVOLUTIONAL CHANNEL

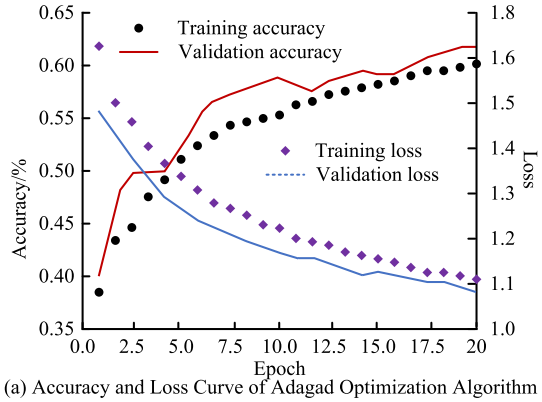
The CK+ dataset is a laboratory-standard facial expression dataset containing 981 training images of eight types of expressions. The FER2013 dataset consists of 28,709 images of faces and non-faces exhibiting different postures and illumination, facilitating the assessment of the algorithm's anti-interference performance. The RAF-BD dataset contains a wealth of human facial expression images, among which the number of two sets are 12,271 and 3,068, respectively. There are seven types of expressions in the single label subset of RAF-BD dataset, while the compound label subset includes 12 types of compound expressions, and the study only applies to the single label subset. The original learning rate for the network is 0.001. In the process of model training, if the ideal convergence effect is not achieved after 5 rounds, the learning rate is adjusted to reduce, but the minimum learning rate is 0.00001. To mitigate overfitting in the early stages of training, a dropout layer with a parameter of 0.3 is added. The network Batch is 64. To select the optimizer most suitable for facial expression data set, the VGG19 network is built using transfer learning. Three optimizers, AdaGrad, RMSProp and Adam, were tested separately on the RAF-DB dataset with a training epoch of 20. The accuracy and loss curves obtained by the experiments of the three optimization algorithms are shown in Figure 7. Compared with Figure 7 and Figure 7 (b), the accuracy of epoch 1 to epoch 2 using RMSprop optimization algorithm is improved by 0.08; The improvement speed is 5% faster than that of AdaGrad optimization algorithm. The precision of every epoch surpasses that of the AdaGrad optimization algorithm. In Figure 7 (c), although the Adam algorithm has a relatively fast lifting speed, its loss oscillates in the late training period, making it difficult for the model to converge. Comparing the performance of the three algorithms, RMSProp algorithm has the best accuracy and loss, so this algorithm is used as the network optimizer.

To evaluate the XRS's effectiveness, an ablation experiment was carried out with the RAF-DB dataset. NN is constructed using conventional convolutional layer and depth-separable convolutional layer respectively. Meanwhile, based on CNN, XR module, XS module and XRS module are introduced respectively. In Table 1, the recognition network using conventional convolutional layers has the lowest accuracy; The accuracy was 75.38% after training. With the use of the separable CL, although the training parameters is increased, the training time per epoch remains the same and the accuracy is improved by 0.71%. The convolutional channel attention network suggested in this study has the largest number of parameters among the five methods, because the number of modules added is the largest. The training time for each epoch on the RAF-DB dataset is 11 seconds, with an achievable accuracy rate of 77.80%. The result is better than a single XR or XS module. Adding XRS network can enhance the accuracy of FER, and the effect is the best.

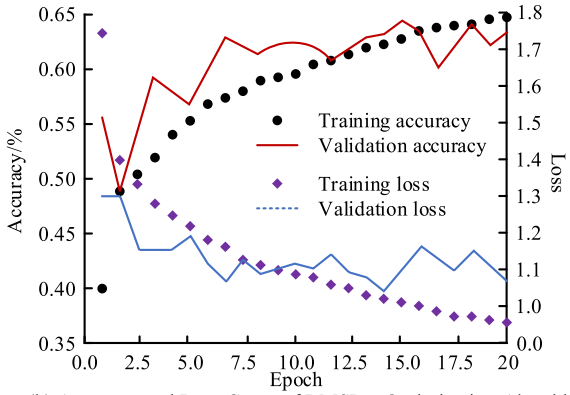
TABLE 1. Ablation experiment for validation of XRS module effectiveness.

Serial Number	Method	Accuracy/%	Parameter quantity(B)	Single epoch runtime/s
1	Conventional convolution	75.38	1814602	10
2	Separable convolution	76.09	1883465	10
3	XR module	77.58	1884490	11
4	XS module	77.74	1891649	11
5	XRS module	77.80	1892681	11

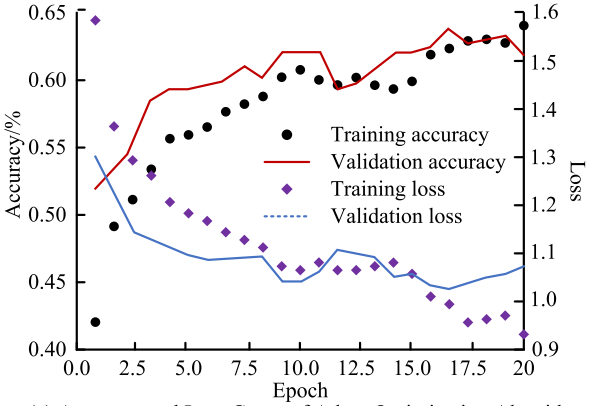
The confusion matrix is a NXN matrix for classifying the results of DL models and evaluating the effect of DL models. To more intuitively understand the real labels and predicted labels of each type of expression in the training process, the confusion matrix was used to assess the precision of the training model on the RAF DB validation set. The confusion matrix obtained by the experiment is shown in Figure 8. In Figure 8, the columns of the confusion matrix represent



(a) Accuracy and Loss Curve of Adagad Optimization Algorithm



(b) Accuracy and Loss Curve of RMSPro Optimization Algorithm



(c) Accuracy and Loss Curve of Adam Optimization Algorithm

FIGURE 7. Experimental accuracy and loss curves of different optimization algorithms.

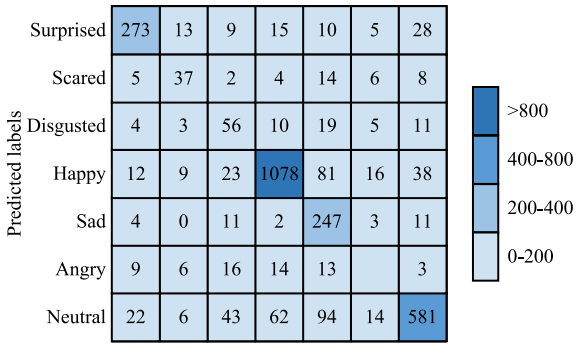
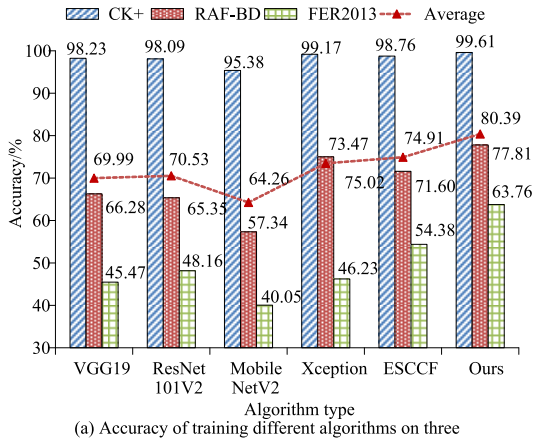
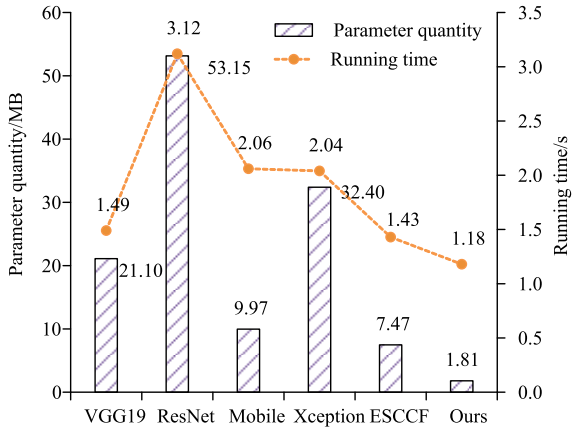


FIGURE 8. Confusion matrix obtained by training with RAF-DB dataset.



(a) Accuracy of training different algorithms on three datasets

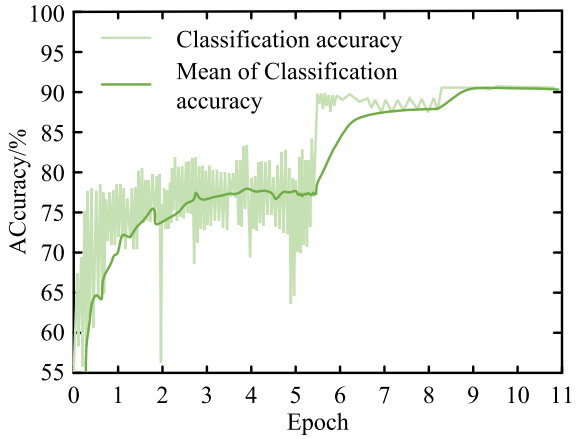


(b) The number of parameters trained by different algorithms on three datasets

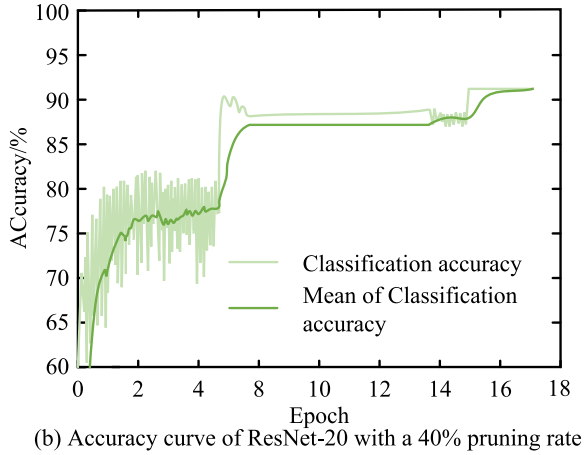
FIGURE 9. Training results of different algorithms on three datasets.

the real categories of expressions, while the rows represent the predicted expression categories.The sum of each column displays the accurate count of emoticons, while the sum of each row indicates the total of predicted emoticon labels. The values in the confusion matrix are data quantities. From Figure 8, it can be seen that 273 data of the surprised expression were correctly recognized.The recall rate for happy is 90%, while surprise has a recall rate of 83% and fear has a recall rate of 50%. Gas has a recall rate of 70% and nausea has a recall rate of 35%. Heartbreak has a recall rate of 52%, and natural has a recall rate of 85%. According to the recall rate, the most effective expression is happy, and the least effective expression was disgusting. The number of correctly classified expressions in the whole verification set represents the precision of the model, with a value of 77.73%. The amount of data in the whole data set is less balanced in different expressions, which may lead to poor recognition results.

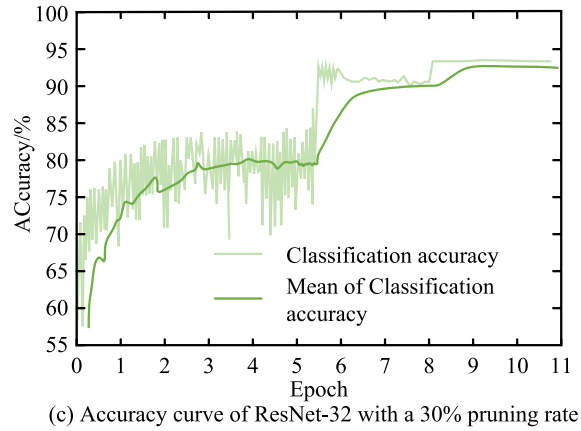
To further prove the effectiveness of the proposed algorithm, a lightweight CNN expression recognition model using VGG19, ResNet101V2, MobileNetV2, Xception and



(a) Accuracy curve of ResNet-20 with a 30% pruning rate

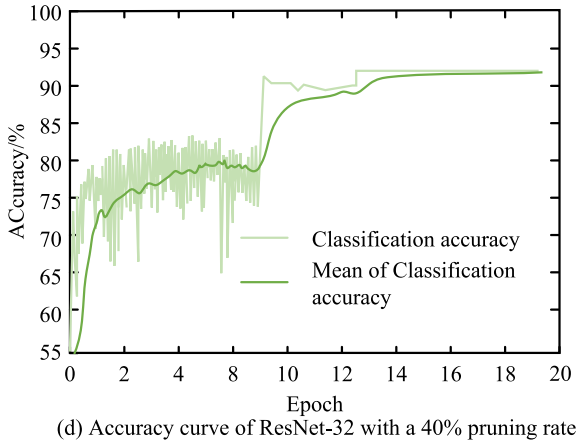


(b) Accuracy curve of ResNet-20 with a 40% pruning rate

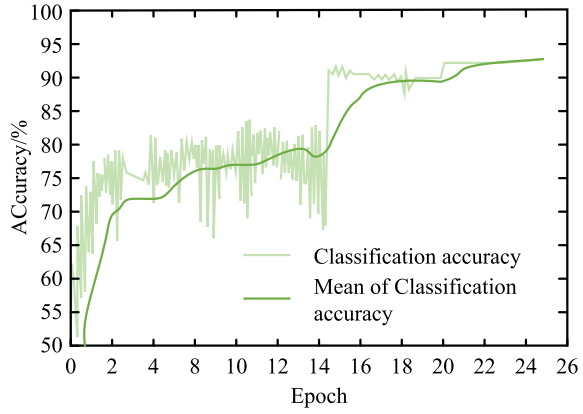


(c) Accuracy curve of ResNet-32 with a 30% pruning rate

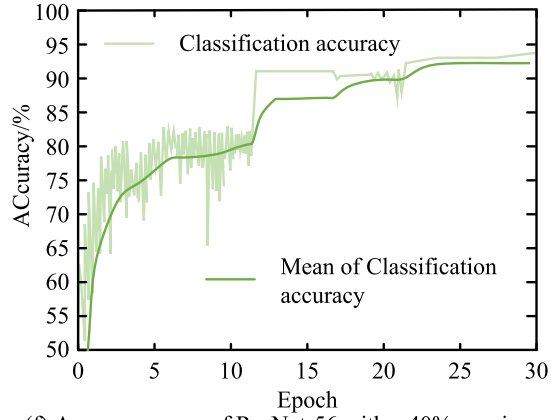
FIGURE 10. Accuracy curves of networks with different depths under different pruning rates.



(d) Accuracy curve of ResNet-32 with a 40% pruning rate



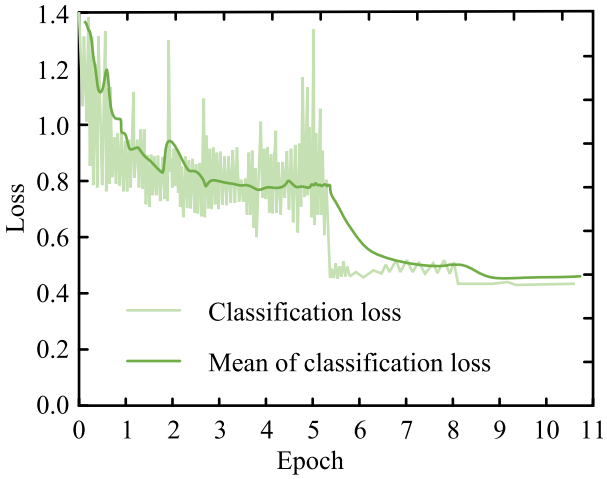
(e) Accuracy curve of ResNet-56 with a 30% pruning rate



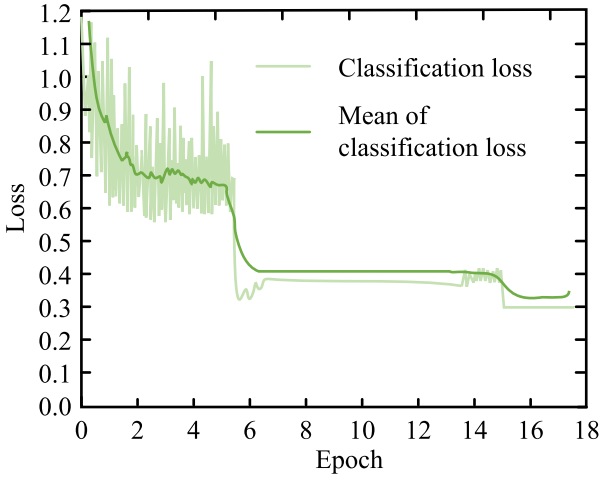
(f) Accuracy curve of ResNet-56 with a 40% pruning rate

FIGURE 10. *(Continued.)* Accuracy curves of networks with different depths under different pruning rates.

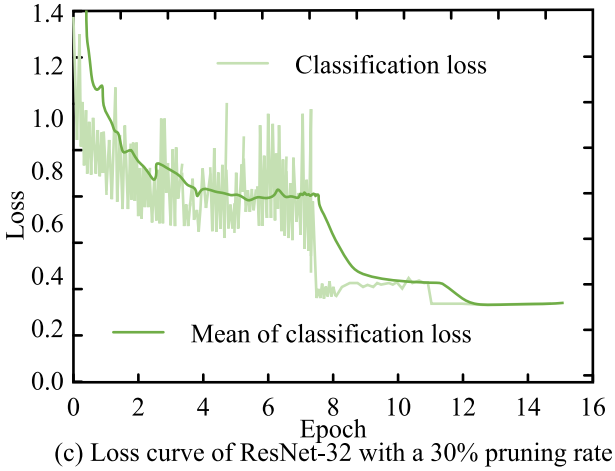
Enhanced Separable convolutional channel features is studied on CK+, RAF-BD and FER2013 datasets Convolutional Channel Features (ESCCF) conducted a comparative experiment. The comparison results of accuracy and number of parameters obtained by 6 different models after training on the 3 data sets are shown in Figure 9. In Figure 9, the MobileNetV2 network has the lowest average accuracy of 64.26%. The proposed network has the highest accuracy rate, with an average accuracy of 80.39%, which is about 25.10% higher than that of the MobileNetV2 network. The FER2013 dataset contains non-face images, so there is interference in the training process, resulting in low performance of the network. However, the accuracy of the proposed algorithm is still 63.76%, and the accuracy of training in CK+ data set is as high as 99.61%. These results indicate good FER performance and a certain level of robustness. In Figure 9 (b), the ResNet101V2 network with the largest number of parameters has 53.15MB. The proposed algorithm has the lowest parameter count, only about 1.18MB. As for the speed of FER, the fastest is the research algorithm, predicting image



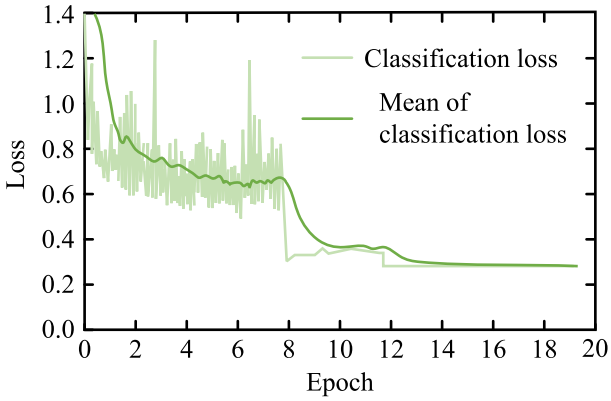
(a) Loss curve of ResNet-20 with a 30% pruning rate



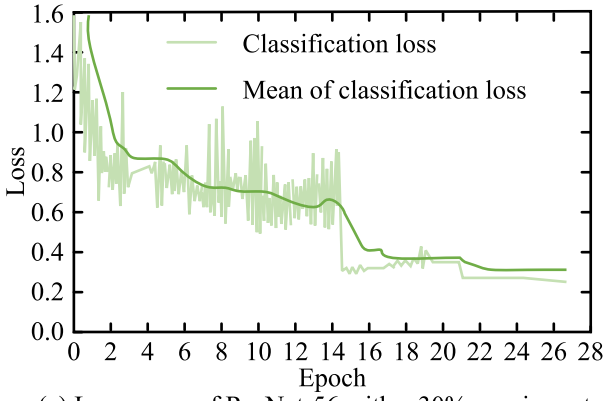
(b) Loss curve of ResNet-20 with a 40% pruning rate



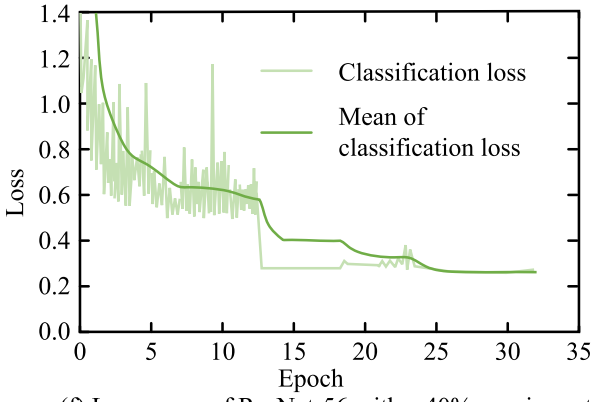
(c) Loss curve of ResNet-32 with a 30% pruning rate



(d) Loss curve of ResNet-32 with a 40% pruning rate



(e) Loss curve of ResNet-56 with a 30% pruning rate



(f) Loss curve of ResNet-56 with a 40% pruning rate

FIGURE 11. *(Continued.)* **Accuracy curves of networks with different depths under different pruning rates.**

FIGURE 11. **Accuracy curves of networks with different depths under different pruning rates.**

results in 1.81 seconds, which is about 41.99% faster than the prediction time of the ResNet101V2 network. Therefore, the XRS module added in this study does not bring a lot of parameters to the network. This achieves network model compression to some extent, and therefore, results in relatively high operational efficiency.

B. IMPLEMENTATION EFFECT OF LIGHTWEIGHT IDENTIFICATION NETWORK FOR HUMAN-COMPUTER INTERACTION SYSTEM

The single label expression subset in RAF-DB is used as the data set to verify the lightweight implementation of the convolutional channel attention recognition network. The experiment utilized a residual network with a depth of

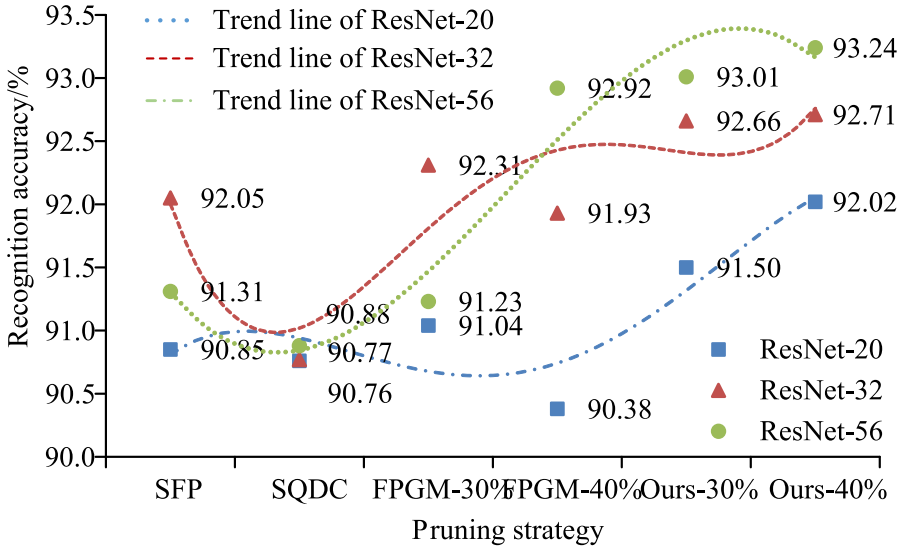


FIGURE 12. Expression recognition accuracy of lightweight convolutional channel attention networks based on different pruning strategies.

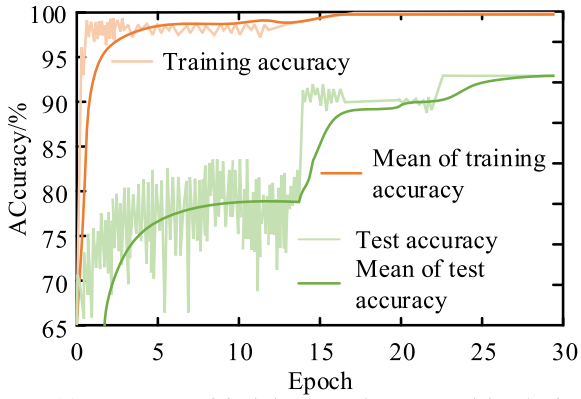
TABLE 2. Accuracy results of pruning strategies with different dimensionality reduction dimensions.

Network model	Dimensionality reduction	Accuracy/%
ResNet-56	Original	93.01
	0.95	93.03
	0.90	93.05
	0.85	93.52
	0.80	93.24

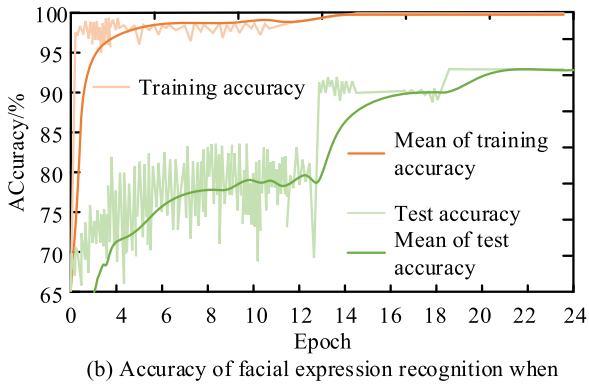
20,32,56 was selected as the experimental object. In the process, the model was pruned on the data set by setting the pruning rate of 30% and 40%. When the smoothness is set to 0.972, the training and testing accuracy changes and loss changes of ResNet-20, ResNet-32 and ResNet-56 on the RAF-DB dataset are obtained. The change of accuracy rate of model classification after pruning is shown in Figure 10. After pruning the model, the classification accuracy of the model gradually increases and tends to be stable when the accuracy reaches a certain degree.

In the process of pruning the model, the accuracy of the model will inevitably be lost. The loss changes of models ResNet-20, ResNet-32 and ResNet-56 with different depths are shown in Figure 11 when the pruning rate is 0.3 and 0.4 respectively. The model loss consistently reduces and can reach as low as 0.3. The loss of models with different depths and different pruning rates generally leveled off after iteration up to 15 epochs, with the fastest iteration requiring only 6 epochs. The model constructed in this study can improve the accuracy of model classification.

To better illustrate the classification effect of lightweight models on data sets, Soft Filter Pruning (SFP), Sparse and Quantization Driven Compression (Sparse and Quantization Driven Compression) were studied. SQDC and Filter Pruning via Geometric Median (FPGM) were compared. The accuracy of the lightweight convolutional channel attention network based on different pruning strategies in expression recognition is shown in Figure 12. When the model depth reaches 56 and the pruning rate is set at 40%, the network model utilizing the pruning strategy proposed in this study exhibits the highest FER accuracy (93.24%).



(a) Accuracy of facial expression recognition before adjusting dimensionality reduction



(b) Accuracy of facial expression recognition when dimensionality reduction is 0.85

FIGURE 13. Accuracy changes of lightweight convolutional channel attention recognition models before and after dimensional changes.

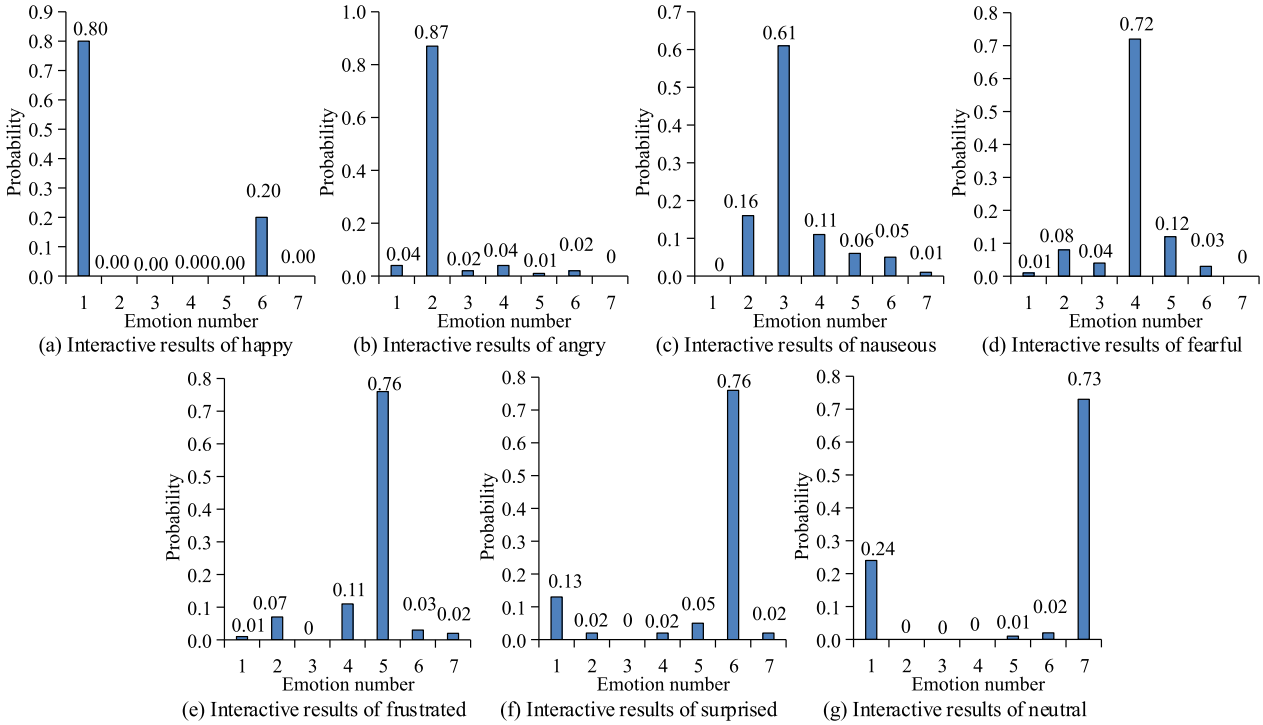


FIGURE 14. Results of human-computer interaction experiments.

The depth-separable convolution filter pruning algorithm, which is based on PCA, changes the Euclidean distance to cosine similarity, and adjusts the dimensionality of PCA reduction. To determine the most appropriate dimensionality reduction, the study conducted several tests. The pruning rate is set to 40% during the experiment of determining the dimensionality reduction. The accuracy rate of expression recognition obtained by the model based on ResNet-56 is shown in Table 2. In Table 2, when dimension is 0.85, the accuracy of model classification is the highest.

The change in accuracy of the lightweight convolutional channel attention recognition model before and after the change in dimension is shown in Figure 13.The pruning strategy proposed in the study has a significant acceleration effect on lightweight recognition models. This strategy reduces the parameter memory occupancy by approximately 41% and improves the classification accuracy, runtime, and calculation cost of the pruned model to some extent.

In order to verify the application effect of the network in real cases, the study conducted human-computer interaction experiments using 7 facial expressions. The expressions are happy, angry, nauseous, fearful, frustrated, surprised, and neutral, and are numbered 1-7 in order. The study designed interactive actions corresponding to expressions using keyframe methods. The interactive experiment is conducted through wireless connection. During the experiment, the images sent by the robot can be seen in real-time through the computer. The computer processes the received images using a convolutional channel attention network and displays the facial expression results directly on the facial image in white font. The computer can also output the predicted probability values of the convolutional channel attention network, At the same time, the maximum probability value calculated by the convolutional channel attention network is sent to the robot as the final expression result. The robot receives this result and engages correspondingly. The findings of the interactive experiment are presented in Figure 14. From Figure 14, it can be seen that in human-computer interaction, robots perform better in recognizing expressions such as joy, anger, fear, frustration, surprise, and neutrality, with recognition probabilities above 0.7. However, the probability of recognizing disgusting expressions is only 0.61, which is due to the limited number of training samples, resulting in poor recognition performance and easy misjudgment as anger and fear. Therefore, the efficacy of robot-human interaction is not ideal. Overall, robots can accurately recognize facial expressions to a certain extent, and the research and design system can meet the basic human-computer interaction expression recognition effect.

V. CONCLUSION

To make robots recognize facial expressions like humans, a FER algorithm for human-computer interaction system is designed. The XRS module is proposed to reduce the parameters in the model and the degradation of the network. Moreover, the SEnet module can weigh channels and filter out significant feature channels.To deploy the recognition model to portable devices, a depth-separable convolution filter pruning algorithm with PCA is suggested to compress the network so that it can be more easily deployed to portable

devices. With the use of separable CL, although thetraining parameters is increased, the training time for each epoch remains the same, and the accuracy is improved by 0.71% compared to conventional convolution. The suggested model can reach the highest recognition accuracy of 99.61% and an average of 80.39%, and the number of parameters is only 1.88MB. The proposed algorithm is also the fastest among the comparison algorithms, with a running time of only 1.81 seconds. The experimental results of the validity verification of network pruning show that when the model depth is 56 and the pruning rate is 40%, the correct rate of FER of the network model based on the pruning strategy is 93.24%. The model's classification accuracy is highest at dimension 0.85. The suggested model has better performance of FER, and has certain robustness and efficiency. The pruning strategy has a good acceleration effect on the model, and has little influence on the model recognition accuracy. However, the recognition speed of the suggestedFERmodel still cannot achieve the ideal effect, and the processing speed of the model needs to be further optimized in the future. The pruning algorithm proposed in this study has a limited optimization effect on the multi-layer model. Further optimization strategies, like parallel pruning, could be considered.

REFERENCES

[1] L. Sheng and C. Li, "Weakly supervised coarse-to-fine learning for human action segmentation in HCI videos," *Multimedia Tools Appl.*, vol. 82, no. 9, pp. 12977–12993, Dec. 2022, doi: 10.1007/s11042-022-13792-1.

[2] M. Yang, "Research on vehicle automatic driving target perception technology based on improved MSRPN algorithm," *J. Comput. Cogn. Eng.*, vol. 1, no. 3, pp. 147–151, Jan. 2022, doi: 10.47852/bonviewJCCE20514.

[3] N. Dua, S. N. Singh, V. B. Semwal, and S. K. Challa, "Inception inspired CNN-GRU hybrid network for human activity recognition," *Multimedia Tools Appl.*, vol. 82, no. 4, pp. 5369–5403, Mar. 2023, doi: 10.1007/s11042-021-11885-x.

[4] R. R. Adyapady and B. Annappa, "An ensemble approach using a frequency-based and stacking classifiers for effective facial expression recognition," *Multimedia Tools Appl.*, vol. 82, no. 10, pp. 14689–14712, Oct. 2022, doi: 10.1007/s11042-022-13940-7.

[5] Z. He, B. Meng, L. Wang, G. Jeon, Z. Liu, and X. Yang, "Global and local fusion ensemble network for facial expression recognition," *Multimedia Tools Appl.*, vol. 82, no. 4, pp. 5473–5494, Apr. 2022, doi: 10.1007/s11042-022-12321-4.

[6] Y. Yang and X. Song, "Research on face intelligent perception technology integrating deep learning under different illumination intensities," *J. Comput. Cogn. Eng.*, vol. 1, no. 1, pp. 32–36, Jan. 2022, doi: 10.47852/bonviewjcce19919.

[7] R. Ni, B. Yang, X. Zhou, A. Cangelosi, and X. Liu, "Facial expression recognition through cross-modality attention fusion," *IEEE Trans. Cogn. Develop. Syst.*, vol. 15, no. 1, pp. 175–185, Mar. 2023, doi: 10.1109/TCDS.2022.3150019.

[8] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," *Neurocomputing*, vol. 411, pp. 340–350, Oct. 2020, doi: 10.1016/j.neucom.2020.06.014.

[9] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.

[10] R. A. Borgalli, and S. Surve, "Deep convolution neural networks for CrossDataset facial expression recognition system," *Int. J. Eng. Manuf.*, vol. 12, no. 6, pp. 40–51, Dec. 2022, doi: 10.5815/ijem.2022.06.05.

[11] M. Jiang and S. Yin, "Facial expression recognition based on convolutional block attention module and multi-feature fusion," *Int. J. Comput. Vis. Robot.*, vol. 13, no. 1, pp. 21–37, Jun. 2023, doi: 10.1504/IJCVR.2022.10044018.

[12] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *Vis. Comput.*, vol. 36, no. 2, pp. 391–404, Feb. 2020, doi: 10.1007/s00371-019-01627-4.

[13] H. Zhao, J. Wu, Z. Li, W. Chen, and Z. Zheng, "Double sparse deep reinforcement learning via multilayer sparse coding and nonconvex regularized pruning," *IEEE Trans. Cybern.*, vol. 53, no. 2, pp. 765–778, Feb. 2023, doi: 10.1109/TCYB.2022.3157892.

[14] Y. Shen, L. Jing, T. Gao, Z. Song, and J. Ma, "A multiple classifiers time-serial ensemble pruning algorithm based on the mechanism of forward supplement," *Appl. Intell.*, vol. 53, no. 5, pp. 5620–5634, Jun. 2022, doi: 10.1007/S10489-022-03855-Z.

[15] Y. Chen, R. Zhou, B. Guo, Y. Shen, W. Wang, X. Wen, and X. Suo, "Discrete cosine transform for filter pruning," *Appl. Intell.*, vol. 53, no. 3, pp. 3398–3414, May 2022, doi: 10.1007/s10489-022-03604-2.

[16] K. Wang, J. Lu, A. Liu, G. Zhang, and L. Xiong, "Evolving gradient boost: A pruning scheme based on loss improvement ratio for learning under concept drift," *IEEE Trans. Cybern.*, vol. 53, no. 4, pp. 2110–2123, Apr. 2023, doi: 10.1109/TCYB.2021.3109796.

[17] X. Zheng, C. Yang, S. Zhang, Y. Wang, B. Zhang, Y. Wu, Y. Wu, L. Shao, and R. Ji, "DDPNAS: Efficient neural architecture search via dynamic distribution pruning," *Int. J. Comput. Vis.*, vol. 131, no. 5, pp. 1234–1249, Feb. 2023, doi: 10.1007/s11263-023-01753-6.

[18] S. Han, H. Shao, J. Cheng, X. Yang, and B. Cai, "Convformer-NSE: A novel end-to-end gearbox fault diagnosis framework under heavy noise using joint global and local information," *IEEE/ASME Trans. Mechatronics*, vol. 28, no. 1, pp. 340–349, Feb. 2023, doi: 10.1109/TMECH.2022.3199985.

[19] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul./Sep. 2022, doi: 10.1109/TAFFC.2020.2981446.

[20] S. Saurav, R. Saini, and S. Singh, "Fast facial expression recognition using boosted histogram of oriented gradient (BHOG) features," *Pattern Anal. Appl.*, vol. 26, no. 1, pp. 381–402, Sep. 2022, doi: 10.1007/s10044-022-01112-0.

[21] P. Barra, L. De Maio, and S. Barra, "Emotion recognition by web-shaped model," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 11321–11336, Mar. 2023, doi: 10.1007/s11042-022-13361-6.

[22] M. D. Putro, D.-L. Nguyen, and K.-H. Jo, "A fast CPU real-time facial expression detector using sequential attention network for human–robot interaction," *IEEE Trans. Ind. Informat.*, vol. 18, no. 11, pp. 7665–7674, Nov. 2022, doi: 10.1109/TII.2022.3145862.

JING PU was born in Nanchong, Sichuan, China, in 1985. She received the bachelor's degree in animation and the master's degree in design and art from Sichuan University, China, in 2008 and 2011, respectively, and the Ph.D. degree in design from Sangmyung University, South Korea, in 2023. Since 2011, she has been a Lecturer of product design and digital media art with Sichuan Agricultural University, and has completed 14 related papers. Her research interests mainly include digital media art design and interaction design.

XINXIN NIE was born in Shandong, China, in 1980. She received the B.A. and M.A. degrees from the Shaanxi University of Science and Technology, Shaanxi, China, in 2004 and 2007, respectively. From 2007 to 2014, she was a University Teacher with the Engineering and Technical College, Chengdu University of Technology, Sichuan, China. Currently, she is with the Chengdu Jincheng College, Chengdu, China.