EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Research on facial expression recognition algorithm based on improved MobileNetV3

Bin Jiang[1*], Nanxing Li[1], Xiaomei Cui[1], Qiuwen Zhang[1], Huanlong Zhang[2], Zuhe Li[1] and Weihua Liu[1]

*Correspondence:
jiangbin@zzuli.edu.cn

[1] College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, China
[2] College of Electric and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, China

Abstract

Aiming at the problem that face images are easily interfered by occlusion factors in uncontrollable environments, and the complex structure of traditional convolutional neural networks leads to low expression recognition rates, slow network convergence speed, and long network training time, an improved lightweight convolutional neural network is proposed for facial expression recognition algorithm. First, the dilation convolution is introduced into the shortcut connection of the inverted residual structure in the MobileNetV3 network to expand the receptive field of the convolution kernel and reduce the loss of expression features. Then, the channel attention mechanism SENet in the network is replaced by the two-dimensional (channel and spatial) attention mechanism SimAM introduced without parameters to reduce the network parameters. Finally, in the normalization operation, the Batch Normalization of the backbone network is replaced with Group Normalization, which is stable at various batch sizes, to reduce errors caused by processing small batches of data. Experimental results on RaFD, FER2013, and FER2013Plus face expression data sets show that the network reduces the training times while maintaining network accuracy, improves network convergence speed, and has good convergence effects.

Keywords: Face expression recognition, MobileNetV3, Dilated convolution, SimAM, Group normalization

1 Introduction

With the development of artificial intelligence technology, facial expressions, as the most direct way to observe human emotions and convey human emotional states, have attracted widespread attention from researchers. As a result, facial expression recognition (FER) technology stands out among all information interaction methods and has become an important research direction for the implementation of artificial intelligence technology [1].

In recent years, researchers in the field of facial expression recognition have obtained a breakthrough research results. However, their models require complex computations and are not suitable for real-world scenarios such as robots and autonomous vehicles. In practical applications of facial expression recognition, such as in robots and autonomous driving vehicles, recognition tasks need to be performed promptly on platforms with limited computing power. Facial expressions involve rich details and dynamic features, requiring the processing and analysis of large amounts of data, which imposes high demands on device performance. Therefore, improving lightweight models has become one of the current research hotspots.

Reference [2] proposed a GAN network guided by geometric information. The network can not only generate new facial expression images, but also realize facial expression recognition through the expression classifier embedded in the network, effectively alleviating the problem of face occlusion caused by attitude deflection, but the recognition rate of negative expressions is not high and the number of model parameters is large. To deploy models more effectively on mobile devices while maintaining a balance between model size and recognition accuracy, Google's team introduced a lightweight convolutional neural network called MobileNetV1 [3] in 2017. MobileNetV1 introduced depthwise separable convolutions as an effective alternative to traditional convolutions. Depthwise separable convolutions separate spatial filtering from feature generation, effectively decomposing traditional convolutions into two separate layers: lightweight depthwise convolutions for spatial filtering and heavier 1×1 convolutions for feature generation. This approach significantly reduces model parameters and computational requirements compared to traditional convolutional neural networks, with only a slight decrease in accuracy. MobileNetV2 builds upon MobileNetV1 by introducing linear bottlenecks and inverted residuals to enhance the efficiency of layer structures. MobileNetV3 [4] further improves upon MobileNetV1 and MobileNetV2 by modifying the head convolutional kernel channel count, introducing an SE (squeeze and excitation) structure in the bottleneck, and using h-swish instead of the swish function. These improvements allow MobileNetV3 to maintain accuracy while significantly reducing parameter count and training time, making it more efficient and real time.

Considering the need for extensive experiments on diverse data sets and the high precision requirements for the facial expression recognition algorithm studied in this paper, an improved facial expression recognition algorithm is proposed based on the more comprehensive MobileNetV3 algorithm. The improvement steps are as follows:

1. Use dilated convolutions in the first convolution of the convolutional network and in the shortcut branch of the inverted residual structure in the network. Dilated convolutions can increase the receptive field without adding parameters and computational overhead, helping the model capture more features of key facial areas and thereby improving recognition accuracy.
2. Replace the original network's SENet with the SimAM (Simple Attention Module). SimAM, as a parameter-free attention mechanism, can automatically adjust attention allocation based on individual differences and facial expression changes, enhancing the model's adaptability and generalization. In addition, it can reduce the introduction of network parameters, further promoting the improvement of network detection speed.
3. Replace the original network's Batch Normalization with Group Normalization. Group Normalization divides facial expression samples into multiple groups without considering batch size, reducing coupling between samples, better preserving individual facial expression features, increasing the stability and applicability of the rec-

ognition model, and requiring fewer computational resources, suitable for resource-constrained devices.

2 Methods

2.1 Dilated convolution

After a standard convolution operation, pooling is usually performed. In addition to its role in dimensionality reduction and enhancing the robustness of local features, pooling also has the function of increasing the receptive field. However, pooling can cause loss of some detailed information in the feature maps, which may reduce the accuracy of image recognition. Without performing pooling operations, the receptive field may remain too small, hindering the extraction of global features. By incorporating pooling operations, the receptive field of the convolutional kernel is expanded, enabling the extraction of more detailed information. Dilated convolution was introduced to avoid the negative effects of pooling. Dilated convolution is a special convolution behavior that injects holes into the convolutional region of a standard convolution to increase the receptive field, and can ensure that the size of the input and output feature maps is the same as that of a standard convolution [5].

Regarding the specific advantages of dilated convolution: first, it introduces a dilation rate to expand the receptive field without sacrificing resolution, and the relative spatial position of pixels remains unchanged. Second, dilated convolution can obtain multi-scale contextual information. When multiple dilated convolutions with different dilation rates are stacked, different receptive fields can provide multi-scale information. Finally, dilated convolution can reduce computational complexity, because it does not require additional parameters:

$$z(\mathrm{p},q)=\sum_{h,j}f\left(p+d*h,q+d*j\right)*g\left(h,j\right) \tag{1}$$

In the equations, p represents the horizontal coordinate in the feature map, q represents the vertical coordinate in the feature map, h represents the horizontal coordinate in the convolution kernel, j represents the vertical coordinate in the convolution kernel, f represents the parameter value in the feature map, g represents the parameter value in the convolution kernel, and d represents the dilation rate.

The dilation rate represents the multiple at which each element of the original kernel is extended. Figure 1 shows the receptive fields of a 3×3 kernel at different dilation rates. By observing the changes in the central area, the expansion size can be clearly demonstrated. Indeed, dilated convolution expands the receptive field of features without introducing additional parameters, and a larger receptive field enables a more comprehensive acquisition of feature information. This ability to capture more context and global information makes dilated convolution a powerful tool for various tasks in computer vision and other fields.

2.2 SimAM attention mechanisms

Attention mechanism is a method that helps us focus on the most important areas in an image while ignoring irrelevant parts. Channel attention and spatial attention are the two main types of attention mechanisms currently used.

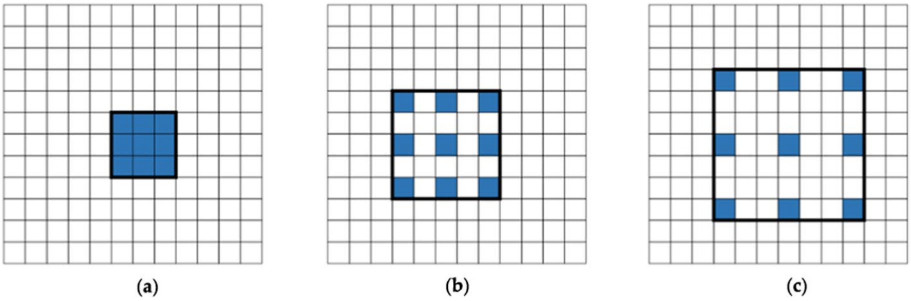


Fig. 1 Sensory wildness of different expansion rates. (**a**) Receptive field of a standard convolution kernel with a dilation rate of $d = 1$, which has a receptive field size of 3 × 3; (**b**) receptive field of a convolution kernel with a dilation rate of $d = 2$, which expands the receptive field size to 5 × 5; (**c**) receptive field of a convolution kernel with a dilation rate of $d = 3$, which expands the receptive field size to 7 × 7

In a neural network, different channels within various feature maps generally correspond to different objects or visual patterns. Channel attention can adaptively readjust the weights of each channel and select objects to focus on Hu [6] first proposed the concept of channel attention mechanism and developed the corresponding SENet [7] method, which explicitly models the interdependence between channels and adaptively recalibrates channel feature responses. Spatial attention is an adaptive mechanism for selecting spatial regions to focus on. Jaderberg [8] proposed the STN method, which relies on a subnetwork to predict relevant regions.

Certain researchers have explored the fusion of channel attention mechanism and spatial attention mechanism to harness the benefits of both approaches and achieve the ability to dynamically select crucial features and regions. The Residual Attention Network (RAN) proposed by Wang [9] is the first network that combines channel attention mechanism and spatial attention mechanism, emphasizing the importance of feature information in both spatial and channel dimensions. Woo [10] proposed the Convolutional Block Attention Module (CBAM). The proposed approach combines channel attention and spatial attention to enhance computational efficiency while introducing global pooling to capture spatial global feature information.

However, there are two issues with existing attention-based modules. First, they extract features using a single dimension (spatial or channel), which lacks flexibility when facing simultaneous changes in both dimensions. Second, combining spatial and channel attention in parallel or in series results in a complex structure and increased computational cost.

To address the aforementioned issues, Yang [11] proposed the SimAM attention mechanism based on a comprehensive neuroscience theory, which enables the collaboration between channel and spatial attention without the need for manual tuning of the network structure. The specific approach is to let the network learn more discriminative neurons and infer three-dimensional weights from these neurons, which are then optimized back. To accurately infer the three-dimensional weights, Reference [11] introduces a straightforward energy function that measures the linear separability between neurons. The energy function is shown in the following formula:

$$\mathrm{e}_t\left(\omega_t, b_t, y, x_i\right) = \left(y_t - \hat{t}\right)^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}\left(y_o - \hat{x}_i\right)^2 \tag{2}$$

Here, t and x_i are the linear transformation matrices for channel $\hat{t} = \omega_t + b_t$ and spatial dimensions $\hat{x}_i = \omega_t \times b_t + b_t$, respectively, M is the number of neurons on that channel, and y is the output value of the neuron, defined as 1 or − 1. Minimizing the above formula is equivalent to training the linear separability between the neuron t within the same channel and other neurons. For the sake of simplicity, binary labels are employed, and a regularization term is incorporated. The ultimate energy function is defined in the following formula:

$$\mathrm{e}_t\left(\omega_t, b_t, y, x_i\right) = \frac{1}{M-1}\sum_{i=1}^{M-1}\left(-1 - \left(\omega_t x_i + b_t\right)\right)^2 + \left(1 - \left(\omega_t t + b_t\right)\right)^2 + \lambda\omega_t^2 \tag{3}$$

In theory, we have M energy functions for each channel. Each channel has $M = H \times W$ energy functions. ω_t and b_t represent the weight and bias transformations. The analytical solutions to the above formulas are given by Eqs. 4 and 5:

$$\omega_t = -\frac{2(t - \mu_t)}{(t - \mu_t)^2 + 2\sigma_t^2 + 2\lambda} \tag{4}$$

$$b_t = -\frac{1}{2}(t + \mu_t)\omega_t \tag{5}$$

where $\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i$ and $\sigma_t^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}(x_i - \mu_t)\omega_t$ are mean and variance calculated over all neurons except t in that channel. It can significantly reduce the computation costs to avoid iteratively calculating μ and σ for each position. Therefore, the minimum energy can be obtained by the following formula:

$$e_t^* = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{\left(t - \hat{\mu}\right)^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{6}$$

where $\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$ and $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu})^2$. Finally, according to the definition of the attention mechanism, the feature enhancement can be performed using the learned attention coefficients. The entire process is shown in the following formula:

$$\tilde{X} = \mathrm{sigmoid}\left(\frac{1}{E}\right) \odot X \tag{7}$$

In the context of the given statement, E groups all e_t^* across channel and spatial dimensions. *sigmoid* is added to limit excessively large values in E. This addition does not affect the relative importance of each neuron since *sigmoid* is a monotonous function, meaning it does not alter the ranking or order of importance among the neurons.

In summary, to better implement the attention mechanism, the SimAM mechanism needs to evaluate the significance of each neuron and measure the linear separability between neurons to find important ones. Formula 6 implies that the lower the energy, the greater the difference between the neuron and the surrounding neurons, and the higher the importance. Therefore, the importance of the neuron can be obtained by $1/e_t^*$. Comparing with the two-step method of CBAM that requires too much computation, the SimAM's calculation method using three-dimensional weights is relatively simple, which keeps the attention module lightweight.

In this paper, a comparative analysis of existing attention modules is presented in Table 1. In the tablet the acronym CAP stands for channel average pooling, GMP represents spatial max pooling, CMP means channel max pooling, GSP signifies spatial dimension standard deviation, C2D stands for 2D convolution, FC means fully connected network, and BN on behalf of batch normalization. *k* and *r* are convolutional filter numbers and reduction ratio, respectively. *C* is the current feature channels.

Based on Table 1, it can be seen that the SimAM attention module has a clear advantage in terms of parameter count compared to the other modules listed. Compared to the CBAM attention mechanism, SimAM is a simple parameter-free attention mechanism, so it does not add any extra parameters to the network.

2.3 Group normalization

Indeed, batch normalization (BN) is a fundamental and valuable technique in deep learning, particularly in computer vision tasks. However, the use of batch normalization also brings some problems. BN requires a sufficiently large batch size to work properly, and for small batch data processing, the estimation of statistical data becomes inaccurate. In addition, if the batch size is reduced, the model's error will quickly increase. However, in large network training tasks, memory constraints force researchers to use smaller batch normalization batch sizes. In response to these issues, Group Normalization (GN) was proposed in Reference [12]. This method divides the input feature map into multiple groups based on the number of channels. Within each group, it calculates the mean and variance of the samples and normalizes the corresponding channels using the obtained data. Finally, the normalized samples are reassembled in the original order to obtain the final output feature map. Unlike traditional batch normalization methods, group normalization's computations are independent of batch size, and it maintains stable accuracy across various batch sizes. Since group normalization does not require consideration of batch size, it can train neural networks with smaller batch sizes, which is useful for devices and applications with limited memory and computational resources.

2.4 Improved network structure

To meet the efficiency requirements of facial expression recognition, a lightweight network model needs to be used. MobileNetV3 is a lightweight network model developed by Google. The MobileNetV3 large version not only ensures accuracy but also improves

Table 1 Parameters of different attention modules

Attention Modules	Type	Operators	Parameter
SE	Channel attention	GAP, FC, ReLU	$2C^2/r$
CBAM	Channel & spatial attention	GAP, GMP, FC, ReLU, CAP, CMP, BN, C2D	$2C^2/r + 2k^2$
SimAM	Channel & spatial attention	CAP,/,$\odot$,$+$	0

speed compared to the previous MobileNetV2. This is of great significance for improving the efficiency of facial expression recognition. Consequently, utilizing the MobileNetV3 large version as the foundation, this paper introduces a facial expression recognition model. To enhance the training speed and minimize the number of training iterations, we have implemented modifications to the MobileNetV3 large network, driven by the following fundamental concept:

First, to extract facial expression features as comprehensively as possible, the improvement strategy incorporates dilated convolutions into the convolutional operations. Specifically, the improvement involves adding dilated convolutions to the initial convolution operation, introducing dilated convolutions and max-pooling operations to the shortcut connections of each inverted residual structure. Dilated convolutions allow the aggregation of context information from different scales without reducing the image resolution. Therefore, incorporating dilated convolutions in the initial convolution and inverted residual shortcut connections helps maintain the details and clarity of the image during feature extraction. In addition, since the kernel size of dilated convolutions remains unchanged, their parameter count is the same as that of regular convolutions. However, due to the larger receptive field of dilated convolutions, they can more effectively utilize parameters, improving the efficiency of feature extraction. By enhancing the network structure of MobileNetV3, more feature information is extracted, leading to faster network convergence. The improved network structure is illustrated in Table 2.

Second, in pursuit of reducing network parameters and enhancing network operation efficiency, the original SENet attention mechanism of MobileNetV3 is replaced with

Table 2 Improved MobileNetV3 network architecture

Input	Operator	Exp size	#out	SimAM	NL	s
$224^2 \times 3$	Conv2d, dilation = 2, GN		16	N	HS	2
$112^2 \times 16$	Bneck, 3 × 3,dilation = 2, GN	16	16	N	RE	1
$112^2 \times 16$	Bneck, 3 × 3, GN	64	24	N	RE	2
$56^2 \times 24$	Bneck, 3 × 3, dilation = 2, GN	72	24	N	RE	1
$56^2 \times 24$	Bneck, 5 × 5,GN	72	40	Y	RE	2
$28^2 \times 40$	Bneck, 5 × 5, dilation = 2, GN	120	40	Y	RE	1
$28^2 \times 40$	Bneck, dilation = 2, 5 × 5, GN	120	40	Y	RE	1
$28^2 \times 40$	Bneck, 3 × 3, GN	240	80	N	HS	2
$14^2 \times 80$	Bneck, dilation = 2, 3 × 3, GN	200	80	N	HS	1
$14^2 \times 80$	Bneck, dilation = 2, 3 × 3, GN	184	80	N	HS	1
$14^2 \times 80$	Bneck, dilation = 2, 3 × 3, GN	184	80	N	HS	1
$14^2 \times 80$	Bneck, dilation = 2, 3 × 3, GN	480	112	Y	HS	1
$14^2 \times 112$	Bneck, dilation = 2, 3 × 3, GN	672	112	Y	HS	1
$14^2 \times 112$	Bneck, dilation = 2, 5 × 5, GN	672	160	Y	HS	1
$7^2 \times 160$	Bneck, 5 × 5, GN	960	160	Y	HS	2
$7^2 \times 160$	Bneck, dilation = 2, 5 × 5, GN	960	160	Y	HS	1
$7^2 \times 160$	Conv2d, 1 × 1, GN	–	960	–	HS	1
$7^2 \times 960$	Pool, 7 × 7, GN	–	–	–	–	1
$1^2 \times 960$	conv2d, 1 × 1	–	1280	–	HS	1
$1^2 \times 1280$	Conv2d, 1 × 1	–	8	–	–	1

SimAM attention mechanism. Compared with the original single-dimensional (channel) attention mechanism, the replaced attention mechanism not only realizes parameter-free introduction, but also infers from two independent dimensions (space and channel). The improved structure is shown in the red box in Fig. 2;

Third, to reduce the errors caused by processing small batches of data, the original BN in the network is replaced with GN, which is independent of batch size and maintains stable accuracy under various batch sizes.

The overall improved network architecture is shown in Table 2; "bneck" represents the basic structure of the network; "#out" represents the number of output channels; "SimAM" indicates whether the channel attention mechanism (Simple Attention Module) is used; "NL" represents the type of activation function, including "HS" (h-swish) and "RE" (ReLU); and "s" stands for the stride, indicating the step size used in convolution stride operations for down sampling. The network employs convolution stride operations for down sampling.

3 Experiments and results

This section presents the experiments and results achieved by using our proposed architecture for facial expression recognition on RaFD, FER2013, and FER2013Plus data sets. We introduce the data sets, implementation details for our network, and the experiments in the following subsections.

3.1 Data sets

The experiments used three classic facial expression data sets. The Radbound Faces Data Set (RaFD) [13] contains 8040 static facial images of 67 models displaying eight facial emotions.

The FER2013 [14] data set as a benchmark for facial expression recognition algorithms. It contains 35,886 images for training and testing facial expression recognition models. Each image is a grayscale image of size 48 × 48 and is labeled with one of seven emotion categories.

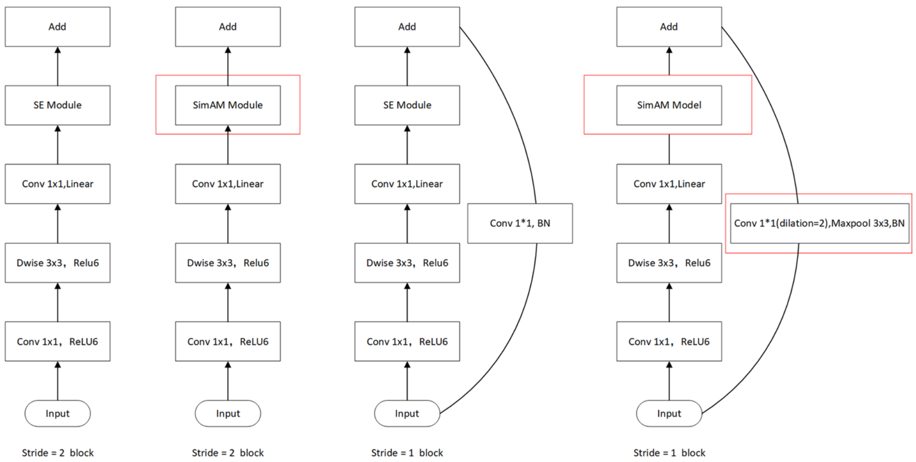


Fig. 2 Improved network structure

FER2013Plus [15] has precisely the same images as FER2013, only minus some cartoons, but after a process of curating labels. FER2013Plus includes several optimizations and improvements, such as more accurate labels, better data cleaning, and more facial expression images. It contains eight emotion categories, including the addition of the contempt emotion category not found in FER2013.

In the subsequent experiments, we partitioned the data sets. The RAFD data set and FER2013Plus data set were split into training and test sets in an 8:2 ratio. For the FER2013 data set, the training and test sets were divided in a 9:1 ratio.

3.2 Experimental parameters

The experiments were conducted on the deep learning framework Pytorch on a Windows system. The hardware configuration consisted of an Intel(R) Core(TM) i5-11400 CPU, an NVIDIA GeForce RTX 3060Ti GPU, and 16 GB of RAM. The stochastic gradient descent method was used to update the parameters, with an initial learning rate of 0.0001, and the total number of training batches was set to 16. The experiments employed stochastic gradient descent (SGD) to update parameters, with an initial learning rate set to 0.0001. For better visualization of the training curve, experiments were conducted on the RaFD data set with 30 epochs. To further validate the algorithm's recognition accuracy and reduce data errors, we conducted five experiments for each data set. In each experiment, the training epochs was set to 100, and obtained the mean as well as the standard deviation for each data set. The purpose of setting these parameters was to prevent the occurrence of the gradient vanishing or exploding problem during the model training process.

3.3 Results and analysis

3.3.1 Results

To validate the effectiveness of the improved network in terms of recognition accuracy, the study compared the recognition accuracy between the MobileNetV3 baseline network and the enhanced network. Using the RaFD data set as an example, five experiments were conducted, each consisting of 100 training epochs. To minimize errors, the average value of each experimental run was calculated as the final training result. The experimental results are displayed in Fig. 3.

Through the analysis of experimental data, the conclusion is drawn that the network proposed in this paper has achieved significant improvements in both accuracy and convergence speed compared to the baseline network. The baseline network ultimately achieved a recognition rate of 90%, while the network proposed in this paper achieved a recognition rate of 95.8%. This confirms the effectiveness of the proposed network in this study. To further validate the performance of the improved network across different data sets, experiments were also conducted on the FER2013 and FER2013Plus data sets. The mean and standard deviation for each data set were obtained, and the final accuracy and error intervals are presented in Table 3.

3.3.2 Ablation experiments

To assess the effectiveness of various optimization strategies in facial expression recognition, we conducted a step-by-step experimental comparison on the RaFD data

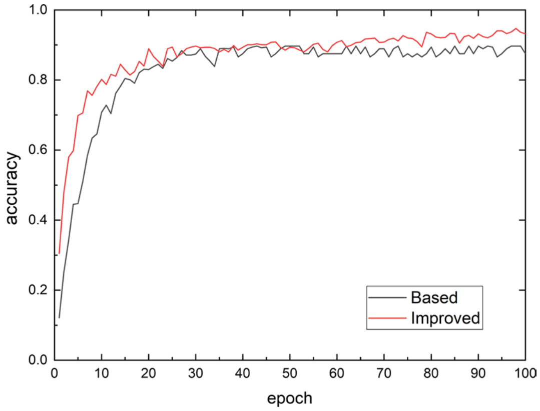


Fig. 3 Recognition accuracy of different networks on the RaFD data set

Table 3 Accuracy of data sets

Data set	RaFD	FER2013	FER2013Plus
Accuracy (%)	95.80 ± 0.20	67.65 ± 0.18	83.47 ± 0.23

set. We trained each improved network for 30 epochs and tested the loss and accuracy curves. The steps of the improved networks are as follows:

1. Using only MobileNet-Large as the basic framework without any fine-tuning, this method is referred to as "base".
2. Implementing dilated convolution solely on the first convolutional layer of the base network, without making any other modifications. This method is referred to as "first-conv."
3. Excluding dilated convolution from the first convolutional layer of the base network, but solely applying it to the shortcut connection. This approach is referred to as the "shortcut" method.
4. Applying dilated convolution to both the first convolutional layer and shortcut connection of the base network. This method is referred to as "first-conv + shortcut."
5. Adding max-pooling to the shortcut connection of 4. This method is referred to as "first-conv + shortcut + maxpool."

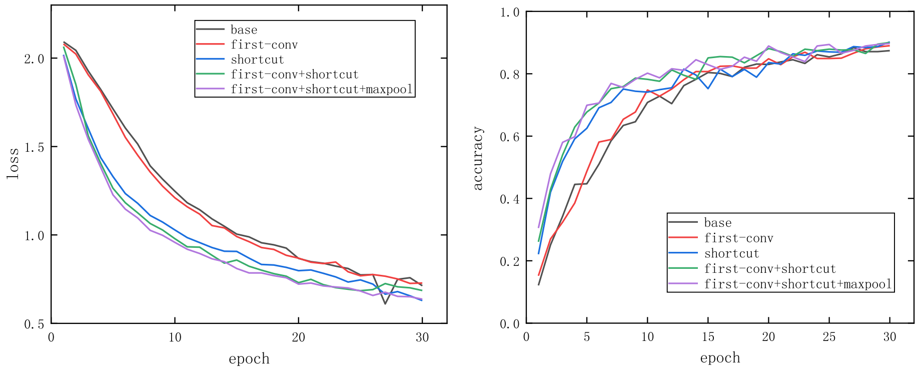
In the RaFD data set experiments, the improved networks outperformed the baseline network in terms of recognition accuracy. The training loss and curves of each improved network are shown in Fig. 3.

The performance metrics for this experiment include the recognition rate of the model at each epoch during the training process. To analyze the feature extraction effectiveness of the network more accurately, we conducted a performance comparison of the algorithm on the RaFD data set. The experimental results are shown in Table 4.

In this experiment, the measurement indicator is the recognition rate of the model in each epoch during the training process. For purpose of more accurately analyze the

Table 4 Training recognition rates on RaFD data set

Epochs	1	2	3	4	5	10	20	30
Base	12.1	25.1	34.5	44.5	44.7	70.8	83.0	87.4
First-conv	15.2	27.0	34.7	38.5	58.1	74.8	84.8	89.0
Shortcut	22.1	42.0	51.9	59.1	62.5	74.1	83.6	90.2
First-conv + shortcut	26.1	42.9	54.0	62.9	67.7	78.2	89.9	91.9
First-conv + shortcut + maxpool (The Proposed Method)	30.5	47.9	58.0	59.8	69.9	80.2	90.3	92.1



(**a**) (**b**)

Fig. 4 RaFD training curve. (**a**) Training loss function in RaFD data set; (**b**) training curve of accuracy in RaFD data set

feature extraction effect of the network, we compared the performance of the algorithms on the RaFD data set. The experimental results are shown in Table 4.

As shown in Fig. 4b, the improved network has the fastest convergence speed, the fastest decrease in the loss function, and the recognition accuracy reaches the optimal value quickly. The accuracy shown in Table 4a for different iterations of the network shows that as the network is continuously improved, by employing these modifications, the convergence speed of the network accelerates, and the accuracy continues to improve. This validates that the network proposed in this paper effectively reduces the number of network iterations and enhances the network training speed.

3.3.3 Confusion matrices

Tables 5, 6, and 7 show the confusion matrices of the improved MobileNetV3 algorithm on three data sets. The confusion matrix can show which categories the model will confuse when making predictions. By setting the horizontal axis to represent the predicted labels and the vertical axis to represent the true labels, the performance of the algorithm can be better visualized and presented.

According to Table 5, in the RaFD data set, the confusion matrix of the improved algorithm shows that the recognition rates of disgust and happiness expressions are the highest, both reaching 100%. However, the recognition rate of the fear expression is only 83%, of which 10% of the expressions are incorrectly recognized as sad expressions and 7% of the expressions are incorrectly recognized as surprise expressions.

Table 5 Improved MobileNetV3's confusion matrix on the RaFD data set

True/Predicted	Angry	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.94	0	0.06	0	0	0	0	0
Contempt	0	0.93	0	0	0	0.07	0	0
Disgust	0	0	1.00	0	0	0	0	0
Fear	0	0	0	0.83	0	0	0.10	0.07
Happy	0	0	0	0	1.00	0	0	0
Neutral	0	0.1	0	0	0	0.90	0	0
Sad	0.05	0	0	0.11	0	0	0.84	0
Surprise	0	0	0	0.06	0	0	0	0.94

Table 6 Improved MobileNetV3's confusion matrix on the FER2013 data set

True/Predicted	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.60	0	0.09	0.05	0.15	0	0.11
Disgust	0.15	0.72	0.02	0.02	0.07	0.01	0.01
Fear	0.12	0	0.50	0.04	0.20	0.05	0.09
Happy	0.01	0	0.01	0.91	0.02	0.03	0.05
Neutral	0.09	0	0.08	0.06	0.55	0.01	0.21
Sad	0.03	0	0.07	0.05	0.02	0.83	0.04
Surprise	0.04	0	0.03	0.06	0.15	0.01	0.71

Table 7 Improved MobileNetV3's confusion matrix on the FER2013Plus data set

True/Predicted	Angry	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.83	0	0	0	0.04	0.09	0.03	0.01
Contempt	0.06	0.33	0	0	0.06	0.34	0.15	0.06
Disgust	0.40	0	0.47	0	0	0.02	0	0.11
Fear	0.01	0	0	0.55	0	0.09	0.10	0.25
Happy	0	0	0	0	0.93	0.04	0	0.03
Neutral	0	0	0	0	0.01	0.93	0.04	0.02
Sad	0.02	0	0	0.01	0.02	0.22	0.73	0
Surprise	0.01	0	0	0.01	0.02	0.07	0.01	0.88

The recognition rate of the sad expression is only 84%, of which 11% of the expressions are incorrectly recognized as fear expressions and 5% of the expressions are incorrectly recognized as anger expressions. From the observation of facial expression images, the possible reason is that the model learns similar features of fear and sadness expressions, such as both expressions may show the feature of an open mouth with lips facing down, which leads to angry expressions being incorrectly classified as sad expressions and sad expressions being incorrectly classified as angry expressions when predicting facial expressions.

Table 6 shows the confusion matrix of the improved algorithm on the FER2013 data set. As observed, the fear expression exhibits the highest misclassification rate among all the facial expressions, mainly being misclassified as sad and angry expressions.

Table 8 Comparison of recognition rates of different algorithms under RaFD data set

Method	Accuracy (%)
Chen [16]	90.13
CNN [17]	93.10
LBP + CNN [18]	93.28
UCFEAN(GAN) [19]	92.15
VGG [20]	93.33
Cascade CNN [21]	93.43
Four-stage CNN [22]	94.44
The proposed method	95.80

Table 9 Comparison of recognition rates of different algorithms under FER2013 data set

Method	Accuracy (%)
Xu [23]	65.60
Xception [24]	66.00
Inception [25]	66.40
MobileNetV2 [26]	66.99
F-Xception [27]	67.19
SentiNet [28]	67.50
SVM + RNN [29]	67.65
The proposed method	67.65

This may be because when people feel afraid, their facial expressions may show features of sadness or anger, causing these three expressions to be confused with each other. In contrast, the recognition rate of other expressions is relatively high, especially the happy expression, because there are more samples of happy expressions in the FER2013 data set, and the facial expression changes when people feel happy are more obvious, making the classification of the happy expression relatively easy.

According to the results shown in Table 7, the model achieved a recognition rate of over 93% for "happy" and "neutral" expressions in the FER2013Plus data set, but a relatively lower recognition rate for "contempt" and "disgust" expressions, as these two types of expressions had fewer samples in the training set of FER2013Plus. Some expressions were easy to misclassify, such as 40% of "disgust" expressions being mistakenly recognized as "anger" expressions and 34% of "contempt" expressions being mistakenly recognized as "neutral" expressions. The recognition rates of "happy", "neutral", and "surprise" expressions were relatively high, with "happy" expressions having the highest recognition rate among all expressions, partly due to the largest number of samples for "happy" expressions and also indicating that the features of "happy" expressions are more easily recognizable than those of other expressions. In contrast, the recognition rates of "contempt", "disgust", and "fear" expressions were lower, with "contempt" expressions having the lowest recognition rate due to the smallest number of samples for "contempt" expressions.

Table 10 Comparison of recognition rates of different algorithms under FER2013Plus data set

Methods	Accuracy (%)
LPL [30]	78.66
DETN [31]	75.82
SAFN [32]	79.56
NGO–BILSTM [33]	78.75
The proposed method	83.47

To validate the effectiveness of the enhanced algorithm, this chapter compared it with other advanced algorithms, and the accuracy of different algorithms on RaFD, FER2013, and FER2013Plus data sets are shown in Tables 8, 9, and 10, respectively.

According to Tables 8, the accuracy of the proposed network in the RaFD data set is improved compared to VGGNet [24] and GAN [27], indicating that the proposed network has a good ability to express facial expression image features. However, compared with other CNN methods, the recognition rate is only slightly reduced, indicating that the network did not affect the recognition performance while improving the training speed. Table 9 shows that the proposed network achieved the highest accuracy of 67.65% in the FER2013 data set, fully demonstrating the excellent performance of the network. Table 10 shows that the proposed network achieved the highest recognition rate while improving the network speed in the FER2013Plus natural condition data set.

4 Conclusion

This article utilizes the lightweight convolutional neural network MobileNetV3 as the fundamental network framework. First, dilated convolution operation is added to the shortcut branch of the first convolution and bottleneck layer to expand the receptive field, obtain more facial expression feature information during convolution, suppress non-feature information, and improve the convergence speed of the network, thereby reducing the number of training times. Second, the SimAM attention mechanism without parameters is used to replace the original network's attention mechanism. Finally, the GN, which is stable under various batch sizes, is used to replace the original network's BN, reducing the error caused by processing small batch data. Experimental results on different data sets show that the improved method improves the convergence speed of network training, reduces the number of network parameters, and demonstrates that our method has good advanced and robust properties.

Author contributions

All authors read and approved the final manuscript.

Funding

The authors acknowledge the financial support from National Natural Science Foundation of China under Grant 61702464, 62272423, 61702462, the Henan Provincial Science and Technology Research Project under Grant 222102210103, 232102211014, 222102210010, 232102211006, 232102210044, Henan Province Key R&D Project (241111210400), the Basic Research Projects of Education Department of Henan No. 21zx003, and the Key projects Natural Science Foundation of Henan No.232300421150.

Availability of data and materials

The data sets supporting the conclusions of this article are included within the article.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 21 August 2023 Accepted: 2 August 2024
Published online: 22 August 2024

References

1. B. Jiang, N. Li, R. Zhong et al., New research advances in facial expression recognition under partial occlusion. J. Comput. Eng. Appl. **58**(12), 12–24 (2022)
2. F. Zhang, T. Zhang, Q. Mao et al., Geometry guided pose-invariant facial expression recognition. IEEE Trans. Image Process. **29**, 4445–4460 (2020)
3. A.G. Howard, M. Zhu, B. Chen et al., Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint (2017). https://doi.org/10.4855/arXiv.1704.04861
4. Andrew Howard, Mark Sandler, Grace Chu, et al. "Searching for Mobilenetv3", Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1314–1324.
5. Yuanzhu Liu, Zhiming Ding, Yang Cao, et al. "Multi-scale Feature Fusion UAV Image Object Detection Method Based on Dilated Convolution and Attention Mechanism", Proceedings of the 2020 8th International Conference on Information Technology: IoT and Smart City, 2020: 125–132.
6. Long Chen, Hanwang Zhang, Jun Xiao, et al. "SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5659–5667.
7. Hu. Jie, Li. Shen, S. Albanie et al., Squeeze-and-excitation networks. IEEE Trans. Patt. Anal. Mach. Intell. **42**(8), 2011–2023 (2020)
8. Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. "Spatial Transformer Networks", Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015: 2017–2025.
9. Fei Wang, Mengqing Jiang, Chen Qian, et al. "Residual Attention Network for Image Classification", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3156–3164.
10. Sanghyun Woo, Jongchan Park, Joon-Young Lee, et al. "CBAM: Convolutional Block Attention Module", Proceedings of the European Conference on Computer Vision, 2018: 3–19.
11. Lingxiao Yang,·Ru-Yuan Zhang, Lida Li, et al. "SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks", Proceedings of the 38th International Conference on Machine Learning, 2021: 11863–11874.
12. Yuxin Wu, Kaiming He. "Group Normalization", Proceedings of the European Conference on Computer Vision, 2018: 3–19.
13. O. Langner, R. Dotsch, G. Bijlstra et al., Presentation and validation of the radboud faces dataset. Cogn. Emot. **24**(8), 1377–1388 (2010)
14. I.J. Goodfellow, D. Erhan, P.L. Carrier et al., Challenges in representation learning: a report on three machine learning contests, in *Neural information processing*. ed. by M. Lee, A. Hirose, Z.-G. Hou, R.M. Kil (Springer, Berlin, 2013)
15. Y. Yan, Z. Zhang, S. Chen et al., Low-resolution facial expression recognition: a filter learning perspective. Signal Proc. (2020). https://doi.org/10.1016/j.sigpro.2019.107370
16. S. Chen, *Multi-angle facial expression recognition and its application on based improved VGGNet"* (Shenyang University of Technology, Shenyang, 2020)
17. W. Sun, *Facial expression recognition methods based on deep learning* (Nanjing University of Science & Technology, Nanjing, 2018)
18. Z. Xue, Z. Shao, X. Jiang et al., Expression recognition based on quaternion local coding and convolutional network. Comput. Eng. Des. **41**(2), 507–512 (2020)
19. N. Sun, Lu. Qingyi, W. Zheng et al., Unsupervised cross-view facial expression image generation and recognition. IEEE Trans. Affect. Comput. **14**(1), 718–731 (2020)
20. Abir Fathallah, Lotfi Abdi, Ali Douik, et al. Facial Expression Recognition via Deep Learning. IEEE/ACS 14th International Conference on Computer Systems and Applications, 2017: 745–750.
21. Gozde Yolcu, Ismail Oztel, Serap Kazan, et al. Deep Learning-Based Facial Expression Recognition for Monitoring Neurological Disorders. IEEE International Conference on Bioinformatics and Biomedicine, 2017: 1652–1657.
22. G. Yolcu, I. Oztel, S. Kazan et al., Facial expression recognition for monitoring neurological disorders based on convolutional neural network. Multimed. Tools Appl. **78**, 31581–31603 (2019)
23. Xu. Linlin, S. Zhang, J. Zhao, Expression recognition algorithm for parallel convolutional neural networks. J. Image Gr. **24**(2), 227–236 (2019)
24. O. Arriaga, M. Valdenegro-Toro, P. Plöger, Real-time convolutional neural networks for emotion and gender classification. arXiv preprint (2017). https://doi.org/10.4855/arXiv.1710.07557
25. A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks. IEEE Winter Conf. Appl. Comput. Vision **2016**, 1–10 (2016)
26. Hu. Zhibin, *Face expression recognition based on convolutional neural network combined with attention mechanism* (Northwest Normal University, Lanzhou, 2021)
27. Y. Zhou, S. Chen, Y. Wang et al., *Review of research on lightweight convolutional neural networks* (IEEE, Piscataway, 2020)
28. M.A.R. Refat, B.C. Singh, M.M. Rahman, SentiNet: a nonverbal facial sentiment analysis using convolutional neural network. Int. J. Patt. Recogn. Artif. Intell. **36**(04), 2256007 (2022)
29. L. Ji, Wu. Shilong, Gu. Xiaoqing, A facial expression recognition algorithm incorporating SVM and explainable residual neural network. SIViP **17**(8), 4245–4254 (2023)

30. T. Chen, Pu. Tao, Wu. Hefeng et al., Cross-domain facial expression recognition: a unified evaluation benchmark and adversarial graph learning. IEEE Trans. Patt. Anal. Mach. Intell. **44**(12), 9887–9903 (2021)
31. S. Li, W. Deng, A deeper look at facial expression dataset bias. IEEE Trans. Affect. Comput. **13**(2), 881–893 (2020)
32. Ruijia Xu, Guanbin Li, Jihan Yang, et al. "Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervised Domain Adaptation", Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1426–1435.
33. J. Zhong, T. Chen, L. Yi, Face expression recognition based on NGO-BILSTM model. Front. Neurorobot. **17**, 1155038 (2023)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.