ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

A Light Weight Convolution Neural Network For Real time Facial Expression Detection

1. Prof.R.Yadagiri Rao,professor,Head,H&S,Sri Indu Institute of Engineering &Technology (SIIET), Sheriguda, Ibrahimpatnam, Hydarabad
Emailid:527karuna@gmail.com
2.Theja,,Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad
3.Pradeep Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad
4.Deepak,Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad
5.Saicharan,Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad
6.Raghava,Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad

ABSTRACT : In this paper our group proposes and designs a lightweight convolutional neural network (CNN) for detecting facial emotions in real-time and in bulk to achieve a better classification effect. We verify whether our model is effective by creating a real-time vision system. This system employs multi-task cascaded convolutional networks (MTCNN) to complete face detection and transmit the obtained face coordinates to the facial emotions classification model we designed firstly. Then it accomplishes the task of emotion classification. Multi-task cascaded convolutional networks have a cascade detection feature, one of which can be used alone, thereby reducing the occupation of memory resources. Our expression classification model employs Global Average Pooling to replace the fully connected layer in the traditional deep convolution neural network model. Each channel of the feature map is associated with the corresponding category, eliminating the black box characteristics of the fully connected layer to a certain extent. At the same time, our model marries the residual modules and depth-wise separable convolutions, reducing large quantities of parameters and making the model more portable. Finally, our model is tested on the FER-2013 dataset. It only takes 3.1% of the 16GB memory, that is, only 0.496GB memory is needed to complete the task of classifying facial expressions. Not only can our model be stored in an 872.9 kilobytes file, but also its accuracy has reached 67% on the FER-2013 dataset. And it has good detection and recognition effects on those figures which are out of the dataset.

I. INTRODUCTION

With the rapid development of human-computer interaction and pattern recognition, coupled with the rapid update of computer hardware, people can deliver complex work to computers to meet certain life and market needs. It brings great convenience to humanity. Facial expression recognition is an intelligent human-computer interaction method that has emerged in recent years. It has a wide range of applications, such as VR games, medical care, online education, driving, security, and so on. Nowadays, many cameras have added smile mode, that is, when a smile is detected on the camera, a photo is taken automatically without having to manually press the shutter, which makes the user experience better. In some European countries, people use facial expression recognition to capture the mood fluctuations of elementary school students in classes, so as to analyze their learning status and treat students as individuals. Some models of Toyota's high-end brand Lexus monitor the driver's eyes and facial expressions to detect fatigue driving so as to avoid some traffic accidents.

People's facial expression is one of the important ways to express their own emotions. Sometimes it is easy to find one's inner thoughts by his expressions. The main function of facial expression is to capture the emotional changes of the subject through facial emotions. Compared to other methods of communication, facial expressions are more diverse. It is easier to show someone's own true feelings inadvertently.

In 1971, Ekman [1] first divided expressions into six basic forms, including sadness, happiness, fear, disgust, surprise, and anger. A normal expression has been added to the FER-2013 dataset [2]. Fig. 1 shows the samples of the expressions from the FER-2013 dataset [2]. And as we can see, it is difficult to sort them out manually. Moreover, human beings can classify the images of faces with an accuracy of 63% ± 5% among the seven emotions.

The most advanced methods dealing with images, such as image classification and object detection, are based on

The associate editor coordinating the review of this manuscript and approving it for publication was Nizam Uddin Ahamed.

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

FIGURE 1. Samples of the FER-2013 emotion dataset.

convolution neural networks. Laerence and Giles [3] propose a hybrid neural-network which uses local image sampling, a self-organizing map (SOM) neural network, and a convolutional neural network in combination for human face recognition; Shin *et al.* [4] use deep convolutional neural networks to deal with the computer-aided detection problems, their model involves 5 thousand to 160 million parameters which has high requirements to computer hardware; Chang *et al.* [5] constructed a convolution neural network for extracting the features of the input images. The complexity-aware classification algorithm is used to divide the dataset into a simple classification sample subspace and a complex classification sample subspace, which reduces the complexity of facial expression recognition caused by environmental factors; Georgescu *et al.* [6] combined the automatic features learned by the convolution neural network with the manual features calculated by the bag of the visual word, and used support vector machines as classifiers to predict the class label. Du and Gao [7] present a method which realizes segmentation by the multiscale convolutional neural network, edits each input image in multi-scale analysis, obtains the feature mapping of the focus and defocused regions, and finally achieves the optimal fusion performance in both qualitative and quantitative aspects; Uddin *et al.* [8] Propose a new robust method for feature extraction, called local directional position pattern (LDPP), which can provide robustness for better facial features; A depth camera-based novel method put forward by Uddin *et al.* [9] can extract eigenvalues more robust.

Millions of parameters are required in the CNN architecture of these tasks [10], which makes it difficult to deploy on embedded devices. In GoogleNet [11] and AlexNet [12], the use of large convolution kernels on high-dimensional feature graphs to directly reduce dimensionality does not generate excessive calculations. And continuous large convolution kernels instead of small convolution kernels can reduce the complexity of the model and further compress the number of parameters. In this paper, we propose and design a convolution neural networks framework for detecting facial emotions in real-time and in bulk. Our model employs a Global Average

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

Pooling layer instead of a fully connected layer and marries the residual module and depth-wise separable convolution to subtly reduce a large number of parameters and make our network structure simpler. Moreover, the accuracy of the recognition rate achieves 67% on the FER-2013 dataset.

II. RELATED WORK

The current facial expression recognition methods are mainly divided into two categories, one is the traditional manual method, and the other is the network model using deep learn- ing. Although the traditional method is widely used, it is verylimited in practical applications [13], [14].

Using deep learning to classify facial expressions is usually learning how to use strong supervision methods [15]–[19] to represent the emotional features of great sample data. Thesedatasets mainly come from the 4 papers [20]–[23].

Barros *et al.* proposed a network model based on thetopological structure of VGG-16 for the formalization ofthe Facial Channel neural network for Facial ExpressionRecognition (FER) [24]. Koujan *et al.* proposed a CNN thatrecognized human emotions from a single face image [25].Xiao *et al.* [26] combined the Region of Interest (ROI) andK-Nearest Neighbor algorithm for facial expression recog-nition and solved the problem of the poor generalizationability of deep neural networks in the case of small data.Liu *et al.* [27] proposed a deep learning method based onthe geometric model of the facial region for facial expressionrecognition. Zhao *et al.* [28] proposed a lightweight expres-sion detection model that can solve the delay problem under natural conditions. Abate *et al.* [29] proposed a neural net-work model for face attributes recognition based on transferlearning to group faces according to common facial features.In some common classic CNN models, the part used for feature extraction usually contains a set of fully connectedlayers at the end. And the number of parameters in the fullyconnected layers is often extremely large. For example, VGGNet [30] contains approximately 90% of all their param-eters in its last fully connected layers. The main work ofVGG16 is to prove that increasing the depth of the net-work can affect the final performance of the network toa certain extent. An improvement of VGG16 compared toAlexNet is to utilize several consecutive 3 3 convolutionkernels to replace the larger convolution kernels in AlexNet(11 11, 7 7, 5 5). For a given Region of Interest (ROI),the use of stacked small convolution kernels is better thanthe use of large convolution kernels, because multiple non-linear layers can increase the depth of the network to makethe learning method have more complex patterns and fewer parameters. The network structure of VGG16 is shown inFig. 2, which contains 13 convolution layers and 3 fullyconnected layers. The entire network utilizes the same size ofconvolution kernels (3 3) and maximum pooling size (2 2).Secondly, the combination of several small filter (3 3) convolution layers is better than one large filter (5 5or7 7)convolution layer. Thirdly, it is verified that the performancecan be improved by continuously deepening the network

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

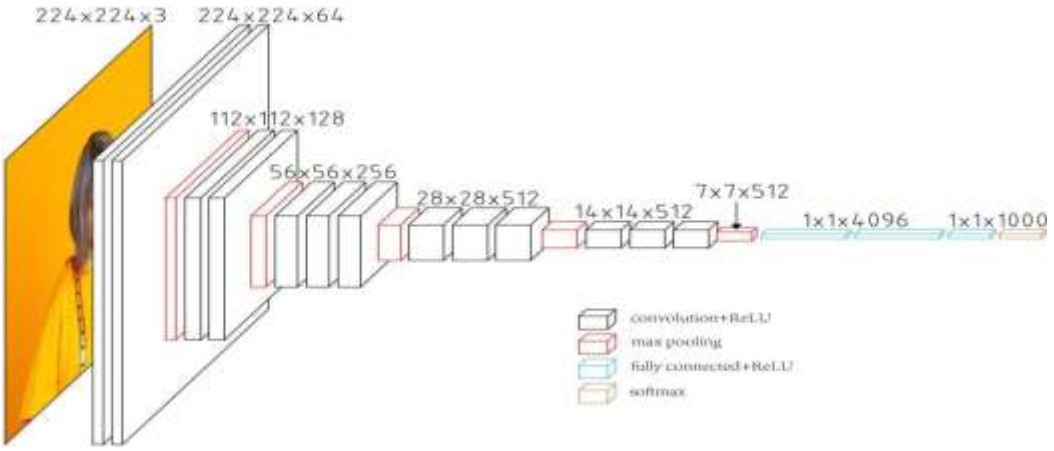

FIGURE 2. VGG16 architecture.

structure. However, VGG-Net also has shortcomings. It consumes more computing resources and uses more parameters, resulting in more memory consumption. Most of the parameters are from the first fully connected layer.

In recent years, an open source model, Inception V3 [31], reduces the number of parameters in the last layers by adding a Global Average Pooling operation. The fully connected layers integrate the feature representation and output it. This operation greatly reduces the impact of feature location on classification. But it has a few issues, such as, too many parameters, slowing down the training speed, and it is easy to overfit. Global Average Pooling reduces each feature image into a scalar value by taking the average over all elements in the feature image. The average operation forces the network to extract global features from the input image. Modern CNN architecture, Xception [32], further reduces the number of parameters by utilizing deep residual learning [33] and depth-wise separable convolutions [34]. Separating the processes of feature extraction and composition within the convolution layer can impel it better.

Whether it is VGG16 or Inception V3, they all improve their accuracy by increasing the depth of the network. However, the first problem with increasing the depth of the network is that these added layers are signals of parameter updates. Because the gradient is propagated backward, the gradient of the front layers will be small after increasing the depth of the network. This means that the learning of these layers is basically stalled, which is caused by gradient disappearance. The second problem of deep networks is training. When the network is deeper, it means that the parameter space is larger, and the optimization problem becomes more difficult. Therefore, simply increasing the depth of the network causes more training errors. The emergence of ResNet solves this problem. Fig. 3 shows the core idea of ResNet, Shortcut Connection. Similar to GoogLeNet, ResNet finally adopts a Global Average Pooling layer. A 152-layer residual network can be trained by using the residual module.

The model designed by us incorporates the idea of GoogLeNet, uses the Global Average Pooling layer in the

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

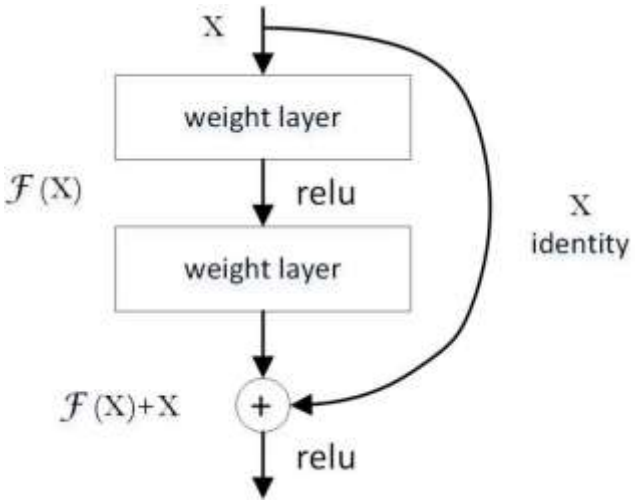

FIGURE 3. Shortcut connection.

end, and reduces the number of network layers. *4*2-norm is added to control the weight coefficient. With these improve- ments, the model will have a strong antidisturbance ability and a good recognition rate.

III. APPROACH

A. DATASET

This paper adopts the open source dataset FER-2013 [2]. Theoriginal dataset is in CSV format, so we need to exploit pan-das to parse and extract the images. After parsing, the datasetconsists of 35,887 facial expressions. Among them, the train set is 28709, the Public validation set and the Private valida- tion set are both 3589. Each figure is composed of a grayscale image with a fixed size of 48 48. There are 7 expres- sions, which correspond to digital labels 0-6 respectively: 0, anger; 1, disgust; 2, fear; 3, happy; 4, sad; 5, surprised; 6,
normal. In the train set, there are 3995, 436, 4097, 7215, 4830, 3171, 4965 figures of the seven kinds of expressions respectively.

The WIDER FACE dataset is a benchmark dataset for face detection which contains 32203 images and 393,703 faces.

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

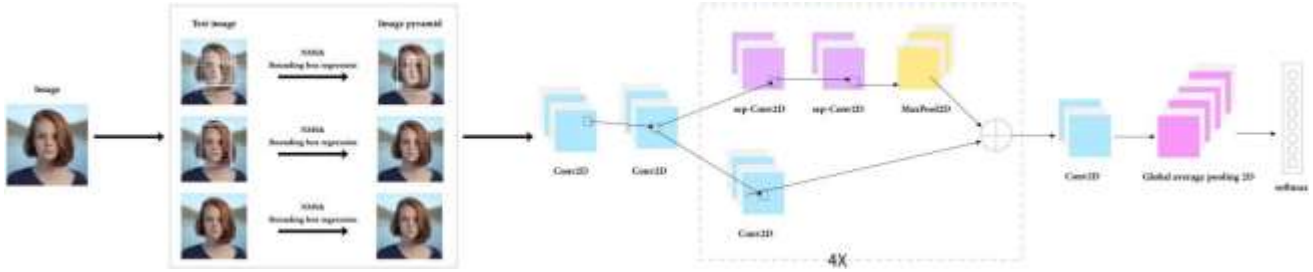

FIGURE 4. Samples of the WIDER FACE dataset.

These faces have a wide range of changes in scale, pose, and occlusion. The images selected by WIDER FACE are mainly derived from the public dataset WIDER. And the producers who come from the Chinese University of Hong Kong select 61 event categories of WIDER. For each category, one of 40%, 10%, and 50% is randomly selected as training, validation, and testing sets. Fig. 4 shows some samples of the WIDER FACE dataset.

The Karolinska Directed Emotional Faces (KDEF) is a set of totally 4900 pictures of human facial expressions, which belongs to a dataset of small data samples. This group of pictures contains 70 people, showing 7 different emotional expressions, including neutral, happy, angry, afraid, disgusted, sad, surprised. These 70 people include 35 women and 35 men. They did not wear makeup, beards, jewelry or glasses when taking pictures. Their expressions were pre-rehearsed. The image noise in the KDEF dataset was very small, which was very suitable as training data for expression recognition. Fig. 5 shows some samples of the KDEF dataset.

FIGURE 5. Samples of the KDEF dataset.

B. MODEL

This paper proposes and designs an emotion recognition model combining MTCNN [35] detection method. We abandon the traditional OpenCV face detection and replace it with MTCNN which uses cascade detection methods and has good

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

detection effects in recent years. We achieved good results in the final experimental test. We eliminate the interference factors of the multiple faces in the image, so that the effect of emotion recognition is greatly improved. In the expressionrecognition model, we learn from the idea of Xception [32], which fuses the use of deep residual learning and depth- wise separable convolutions. The main purpose of this design method is to achieve the best identification accuracy in mul-tiple parameter ratios.

Our initial model uses the Global Average Pooling to com- pletely remove the fully connected layer. This is achieved by placing a feature map in the final convolution layer consistent with the number of classes and using the SoftMax activation function to deal with the classification problem. Our model is trained with the ADAM optimizer [36]. The model structureof our expression classification is shown in Fig. 6.

FIGURE 6. The structure of our model.

The Network in Network model proposed by Lin *et al*. [37] uses the Global Average Pooling method to replace the fully connected layer in the traditional deep convolution neural network model and achieves good results on the CIFAR- 100 dataset. The model of using global average pooling gives the network output layer channels a clear meaning. It makes each channel of the feature map be associated withthe corresponding classification category. To a certain extent,

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

this method eliminates the black box characteristics [38] of the fully connected layer. Therefore, this paper draws on this idea, employs the Global Average Pooling Operation to average each feature map of the feature fusion, and uses it as a new feature map. Global Average Pooling can be linked to global information, strengthen the connection between spatial information, and learn more detailed and comprehensive facial expression features. At the same time, the pooling layer contains no parameters, reduces network parameters, and avoids overfitting.

In the convolution operation, the size of the output image is calculated by using (1), where W is the matrix width, H is the matrix height, F is the width and height of the convolution kernel, P is padding, and S is the step-size. The padding mode used in this paper is the same because the padding operation of the same mode can keep the size of the feature map unchanged after convolution. When the Pooling operation is used, the size of the output image is calculated by using (2), and the final result is rounded down.

$$\frac{(W - F + 2P)}{S} + 1 \quad \times \quad \frac{(H - F + 2P)}{S} + 1 \tag{1}$$

$$\left(\frac{(W - F)}{S} + 1\right) \times \left(\frac{(W - F)}{S} + 1\right) \tag{2}$$

As shown in Fig. 7, in Same mode, the orange part is the image, the blue part is the filter, and the white part is filled with 0. When the center of the filter (K) coincides with the corner of the image, the filter will make a convolution operation to the image. It can be seen that the range of motion is smaller than before. Same mode can keep the size of the feature map unchanged during the forward propagation.

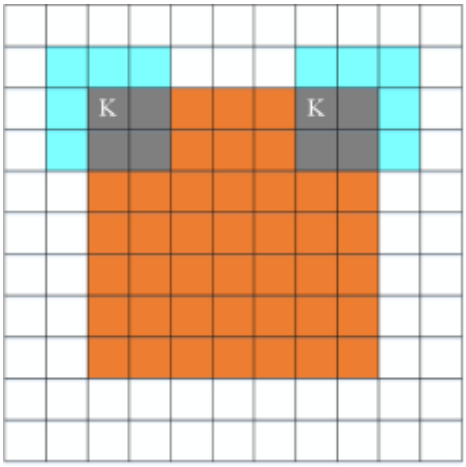

FIGURE 7. Same mode.

Although our group has expanded the dataset, the number of pictures is still relatively small and there are a lot of duplicate faces. At the same time, the pixels of the pictures are exceptionally low and even there is some noisy pictures in the FER-2013 dataset. In order to avoid the phenomenon of overfitting, we add 42-norm to the weight coefficient. We choose 42-norm rather than 41-norm because 42-norm can obtain parameters with small values. And 42-norm can not only prevent the overfitting, but also make our optimal solution stable and fast. The fitting process usually tends to keep the weights as small as possible, and finally constructs a model with all parameters relatively small. When the parameters are small enough, it can make the antidisturbance capability of the model strong. 42-norm loss function is (3), which can be written as (4). (5) means that the gradient descent method is used to update the parameters, no matter what w is, it tries to make it smaller. It is tantamount to multiplying each matrix by a coefficient (1 _ aλ/m), the coefficient is less than 1. So 42-norm is also called ''Weight Decay''.

$$\min_{w,b} J(w, b) = \frac{1}{m} \sum_{i=1} L(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2m} \|w\|_2^2 \tag{3}$$

$$J = J_0 + \frac{\lambda}{2m} \|w\|_2^2 \tag{4}$$

$$\frac{\partial J}{\partial w} = \frac{\partial J_0}{\partial w} + \frac{\lambda}{m} w$$

$$w^r = w - a \frac{\partial J}{\partial w}$$

$$= w - a \frac{\partial J_0}{\partial w} - a \frac{\lambda}{m} w$$

$$\frac{\lambda}{\partial J_0}$$

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

$$= (1 - a_m)w - a_{\partial w} \quad (5)$$

Adding 42-norm can avoid the weight matrix being too large. If the regularization λ is set large enough and the weight matrix W is set to a value close to 0, the intuitive understanding is to set the weight of multiple hidden units to 0, it can basically eliminate many impacts of these hidden units. In this case, the greatly simplified neural network will become a very small network, as small as a logistic regression unit, but it contains many layers of the network. And the network is closer to the state of ''High bias'' from the state of ''High variance'', when an intermediate value λ is given, the network can be in the intermediate state, ''Just right'',as shown in Fig. 8.

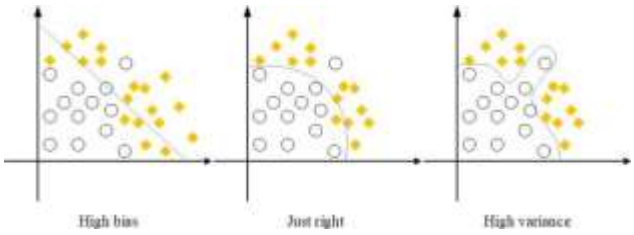

FIGURE 8. Influence of $\boldsymbol{I}$ 2-norm on network.

In other words, if λ is increased enough, W will be close to 0, but in fact this will not happen. We try to eliminate or reduce some impacts of hidden units to make the network simpler and get closer and closer to logistic regression even- tually. We intuitively think that a large number of hidden units are completely eliminated. Actually, this is not so, all hidden units of the neural network will still exist, but their impacts

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

become smaller and the neural network becomes simpler so that overfitting is less likely to occur.

In addition, this model also marries the residual modules and depth-wise separable convolutions. Depth-wise separable convolutions resolve the traditional convolutions into a depth-wise convolution plus a 1×1 convolution. Fig. 9 (a) shows the standard convolution. Let us assume that the size of the input feature map is $D_F \times D_F \times M$, the size of the convolution kernel is $D_K \times D_K \times M$ the size of the output feature map is $D_F \times D_F \times N$ and the parameter of the standard convolution layer is $D_K \times D_K \times M \times N$. Fig. 9 (b) shows depth-wise convolution, and Fig. 9 (c) shows point-by-point convolution.

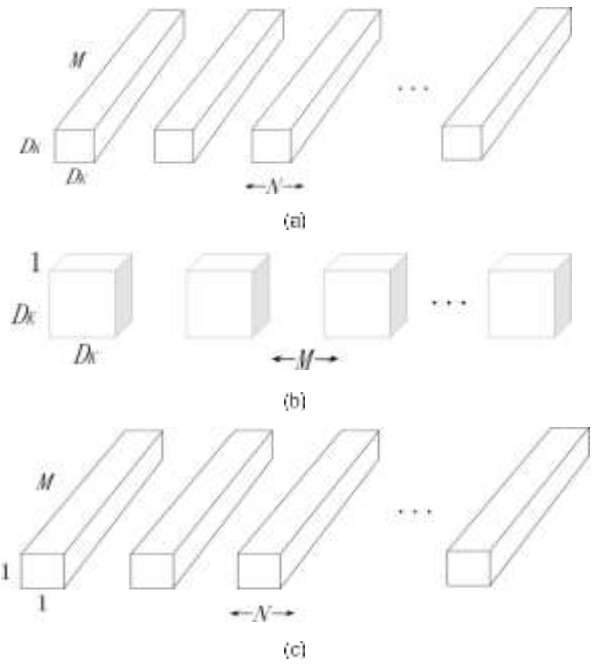

FIGURE 9. (a)Standard convolution filters (b)Depth-wise convolution filters (c)Convolution filters called point wise convolution in the context of depth-wise separable convolution.

The combination of these two convolutions is depth-wise separable convolution. The depth-wise convolution is responsible for filtering, the size is (D_K , D_K , 1), the number is M, acting on each input channel. The point-by-point convolution is responsible for converting the channel, the size is (1, 1, M), the number is N, acting on the output feature mapping of depth-wise convolution. The amount of parameters of depth-wise convolution is $D_K \times D_K \times 1 \times M$, and the amount of parameters of point-by-point convolution is $1 \times 1 \times M \times N$, so the number of parameters of depth-wise separable convolution is $(1/N + 1/D^2_K)$ of the standard convolution.

Our final model is a neural network containing 4 residual depth-wise separable convolutions and combined with an MTCNN detection to achieve the facial expression recognition. Each of these four convolutions is followed by batches of normalized operation and a ReLU [39], [40] activation function. The last layer adopts the Global Average Pooling layer and a soft-max activation function for classification. This architecture has a total of 58423 parameters, of which there are 56951 trainable parameters.

We test on the FER-2013 dataset. The accuracy of sentiment classification achieves 67%. At the same time, the weight of our final recognition architecture can be saved in an 872.9 kilobytes file.

IV. RESULTS

The experiment of our research group is run on an Intel (R) Core (TM) i5-8400 CPU @ 2.80GHz processor, using 16G memory, NVIDIA GeForce GTX 1060 GPU, and Ubuntu 16.04 operation system.

We retrain the MTCNN model using the WIDER FACE dataset and save the obtained weight parameters in the graph file for face detection. Fig. 10 is the true positive rate of the MTCNN we have trained, which can reach about 95%. We have achieved good results in face recognition by using the MTCNN model instead of the traditional Application Programming Interface (API), OpenCV. As can be seen in Fig. 11, (a) is the effect of using OpenCV to recognize faces, (b) is the effect of using MTCNN to recognize faces. Obviously, the detection effect of (b) has been improved a lot, not only to remove the noise in the picture, but also to recognize and classify the unrecognized faces in (a).

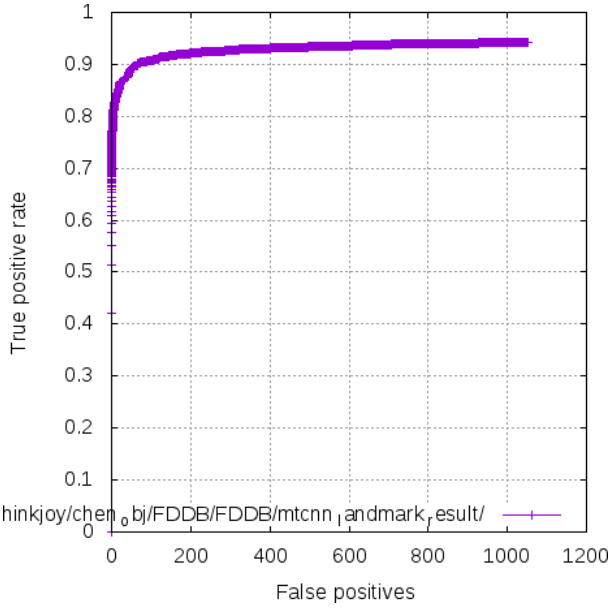

FIGURE 10. True positive rate of the MTCNN.

Table 1 is the recognition rate of the seven expressions based on the standardized confusion matrix of our network architecture. As can be seen, when it comes to recognizing the fear category, the accuracy is still flawed. The main reason for this result is that the diversity of the number of fears in the dataset is low. It means most of them are European faces and lack of data samples of other types.

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

(a)

(b)

FIGURE 11. Comparison of the effects of two different detection modules.

TABLE 1. The test result of our model on the FER-2013 dataset.

Our model	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	0.62	0.01	0.10	0.03	0.12	0.13	0.10
Disgust	0.26	0.55	0.05	0.04	0.06	0.02	0.03
Fear	0.12	0.01	0.40	0.05	0.20	0.11	0.11
Happy	0.02	0.00	0.02	0.88	0.02	0.02	0.05
Sad	0.10	0.01	0.10	0.09	0.53	0.01	0.15
Surprise	0.03	0.00	0.10	0.04	0.02	0.75	0.03
Neutral	0.04	0.00	0.05	0.08	0.11	0.02	0.65

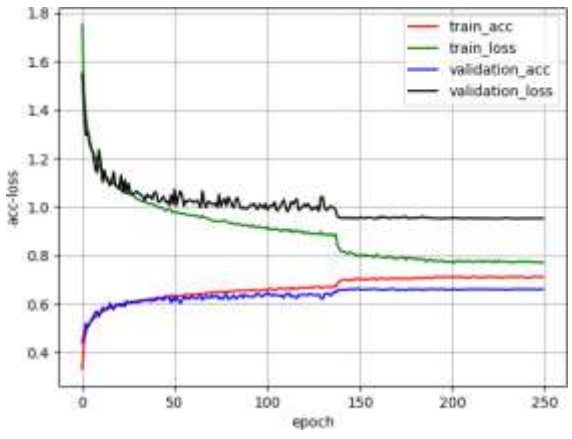

FIGURE 12. The acc-loss curve of our model on FER-2013 dataset.

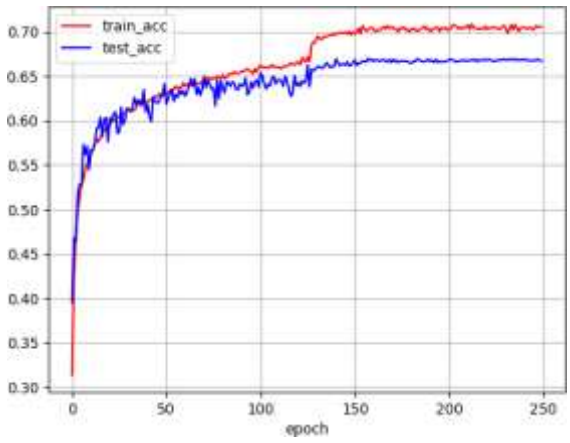

FIGURE 13. Learning curve of our model on train set and test set.

Fig. 12 is the change curves of Loss value and Accuracy value after 250 epochs of our model on the FER-2013 dataset. As can be seen from Fig. 12, the accuracy rate on the train set can reach about 71%, and the accuracy rate on the validation set can achieve 67%. At the same time, the result of our final model on the test set is shown in Fig. 13, and the accuracy converges to a range between 66.8% and 67.0%. As can be seen, after merging deep residual learning and depth-wise separable convolutions, our model has a relatively high degree of accuracy. Because we use Global Average Pooling rather than the fully connected layer and add $l2$-norm, the number of parameters in the model is reduced, which ultimately impels our model more portable.

Table 2 is a comparison of the experimental results of our model and several other models. Since the code is not published in some literature, we cannot retrain their models. We can only get the accuracy of their models on the test set from their literature. We plot Fig. 14 through Table 2 to show the difference of accuracy of these models more intuitively.

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

TABLE 2. Classification results of different models on the train set and test set.

Model	Train set accuracy /%	Test set accuracy /%	Validation set loss
VGG-Net	98.98	59.32	2.39
ResNet-50	98.87	57.48	2.10
CNN	99.70	58.90	2.11
HOG+CNN [41]	-	61.86	-
Improved Inception [42]	-	66.41	-
Network from [43]	-	66.40	-
Network from [44]	-	64.98	-
CNN-based+Softmax [45]	-	65.03	-
Net EXP_DAL_MSE [46]	-	61.59	-
ShallowNet [47]	-	63.49	-
SN(D&BN) [47]	-	64.78	-
Network from [48]	-	65.60	-
ResNet (ReLU-Maxout) [49]	-	66.51	-
Ryank [50]	-	65.08	-
DenseNet-1 [28]	-	70.91	-
Sub-network1 [50]	-	67.00	-
Our model	71.00	67.00	0.98

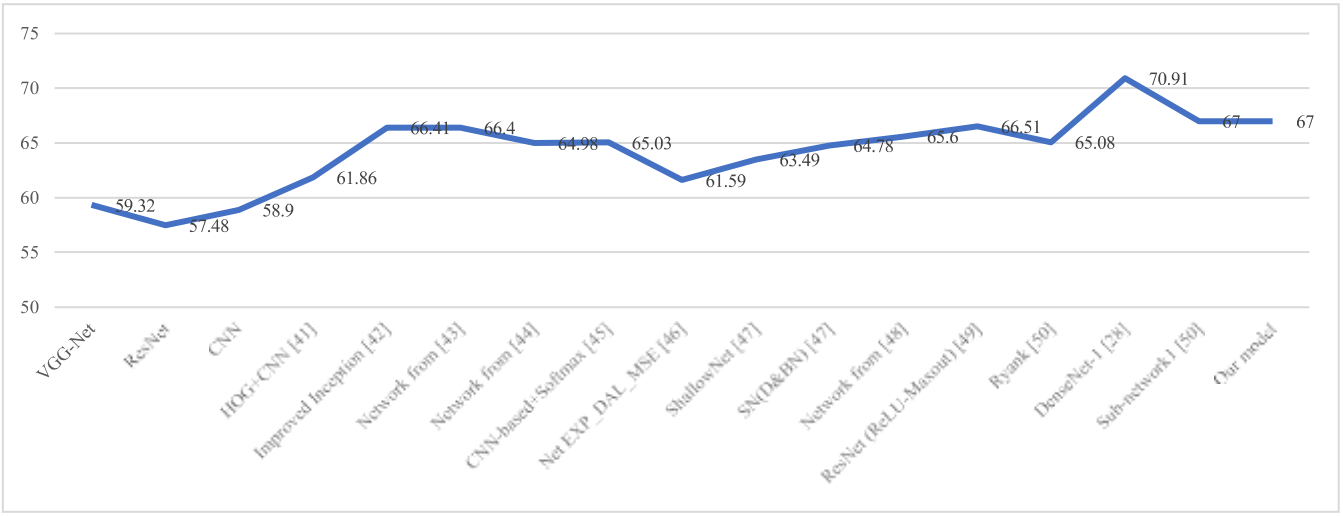

FIGURE 14. Accuracy line chart of each model.

It can be seen from Fig. 14 that the accuracy of our model is higher than that of other models except DenseNet-1. Table 3 shows the data comparison of parameters as well as the complexity between our model and other models. From Table 3, we can see that the parameters of DenseNet-1 are nearly twice of our model's parameters, but our accuracy is only about 3% lower. So, our model still has certain advantages in lightweight. We also train the FER-2013 dataset with three models, VGG-Net, ResNet-50, and CNN, and we control their number of layers so that the number of layers is basically similar to our model. The training results are shown in Fig. 15, Fig. 16, and Fig. 17. It can be seen from these figures that the accuracy of VGG-Net is 59.32%, the accuracy of ResNet-50 is 57.48%, and the accuracy of CNN is 58.90%. The accuracy of these three sets is basically the same. However, the loss value of ResNet-50 suddenly increases during training, which causes the gradient to be too large, then overfitting may occur later. To compare the relationship between the parameters and accuracy of each model, we plot Fig. 18. From Fig. 18, we can see that the accuracy of our model is the second with the least parameters, behind to DenseNet-1. It can be seen from Table 3 that our model has more pooling layers than other models. This is because the pooling layer does not contain parameters. We use the global average pooling layer to replace the fully connected layer, which significantly reduces the number of parameters.

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

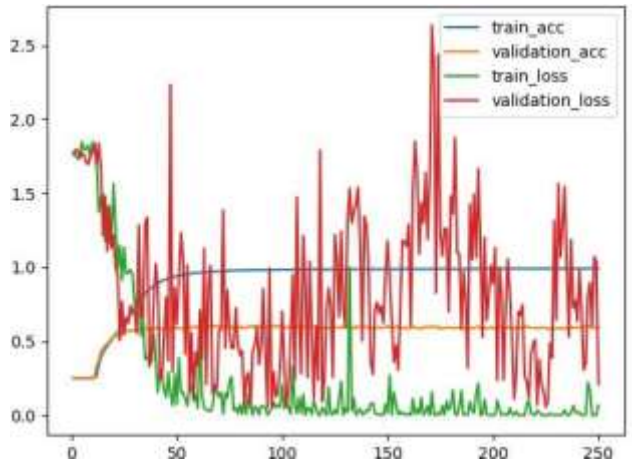

FIGURE 15. The acc-loss curve of VGG-Net on FER-2013 dataset.

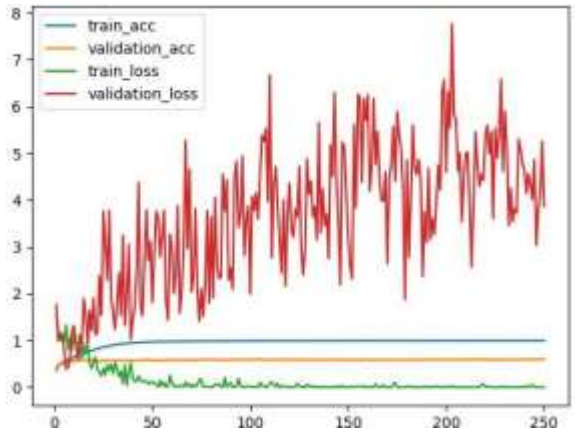

FIGURE 16. The acc-loss curve of ResNet-50 on FER-2013 dataset.

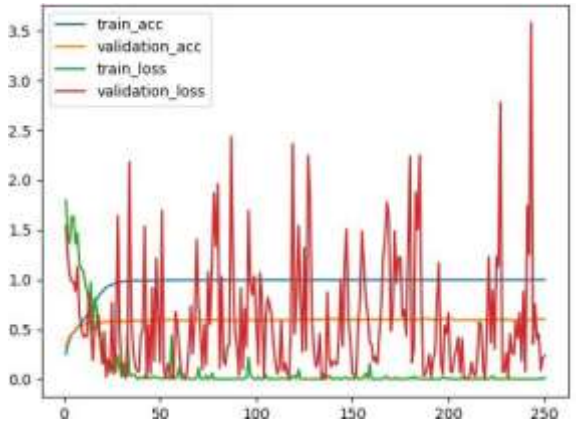

FIGURE 17. The acc-loss curve of CNN on FER-2013 dataset.

Experiments have shown that although we cancel the fully connected layer, it does not affect the accuracy of our model to a large extent.

At the same time, we also monitor the hardware resource occupation of the four models in the training and testing stages. We record the CPU usage, the percentage of CPU occupied by user space, the percentage of CPU occupied by

TABLE 3. Parameters and complexity of models.

Model	Conv	Pooling	FC	Parameters
VGG-Net	10	4	3	87566680
ResNet-50	49	2	1	25500000
DenseNet-128	36	4	1	95263
Our model	15	9	0	58423

TABLE 4. Hardware resource occupation of the four models on the training stages.

Model	CPU [a] / %	Us [b] /%	Sy [c] /%	MEM [d] /%
VGG-Net	18.50	22.05	3.11	4.6
ResNet-50	25.00	29.50	5.92	5.9
CNN	8.67	16.00	4.62	12.5
Our model	16.67	18.97	1.38	4.3

[a] CPU usage of the process; [b] percentage of CPU occupied by user space; [c] time running kernel process; [d] Percentage of physical and total memory used by the process.

TABLE 5. Hardware resource occupation of the four models on the testing stages.

Model	CPU / %	us /%	sy /%	MEM /%
VGG-Net	21.50	24.67	5.90	3.2
ResNet-50	49.78	53.67	15.73	2.9
CNN	10.83	15.12	3.95	12.0
Our model	17.5	21.12	2.65	3.1

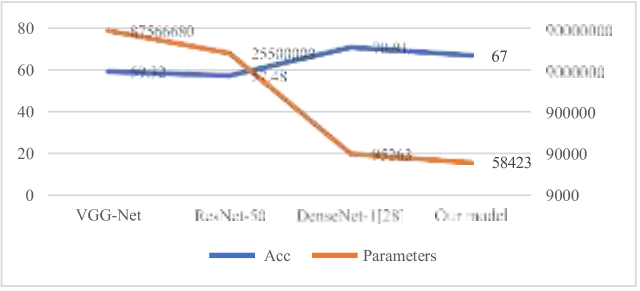

FIGURE 18. Models' parameters and accuracy.

kernel space, the percentage of physical memory occupied, and total memory of the 4 models when they are trained. The results are shown in Table 4.

Table 5 records the hardware resources occupied by the 4 models during testing. Although there is no data in the papers to show the occupation of hardware resources of the models which are shown in Table 2, it can be seen that their models are more complicated than the models we designed by analyzing their model structure. So, it can be deduced that their models' occupation of hardware resources must be higher than ours. As can be seen, when we test the 4 models, ResNet-50 has the highest CPU usage and CNN has

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

TABLE 6. Classification results of the four models on the train set and test set.

Model	Train set accuracy /%	Test set accuracy /%	Validation set loss
ExpNet [51]	-	71.00	-
Network from [52]	-	72.55	-
Network from [53]	-	78.00	-
Our model	96.93	87.71	0.6224

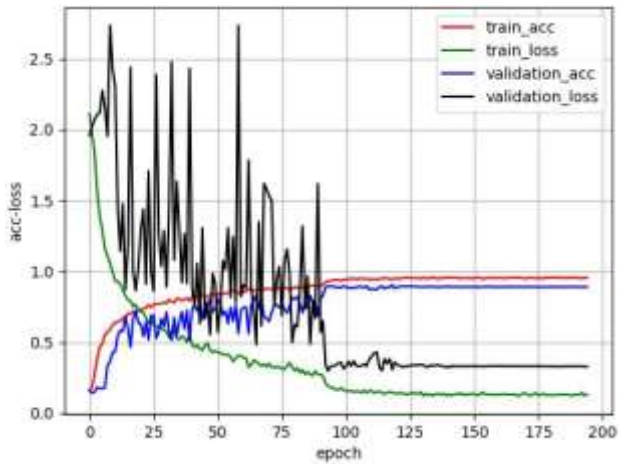

FIGURE 19. The acc-loss curve of our model on KDEF dataset.

FIGURE 20. Emotion classification for specific face.

the largest memory usage. In our model, the usage of CPU is depressed, but the occupation of memory is extremely low, only consuming 3.1% of the memory. It can also be seen in Table 2 that the accuracy of our model is almost the highest, which allows us to deploy our model on embedded devices. At the same time, we can get a relatively accurate recognition rate. In contrast, for lightweight models, the performance of ResNet-50 is not so prominent, the recognition rate is below the average, and the dependence on CPU is also immense. So, it is not desirable to use ResNet-50 as a lightweight detection model.

We also trained our model on the KDEF dataset and compared it with other recent models, such as Table 6. Fig. 19 shows the change curves of Loss value and Accuracy value of our model on the KDEF dataset. At the same time, the result of our final model on the test set can reach 87.71%.

Fig. 20 shows some representative faces searched by us, such as, the face with covering on the left or right side, the face covered by two hands on the cheek, the face covered by teacup on the chin, the face with cap, and the face cocked to one side. For these kinds of faces, this model also achieves a good detection effect.

V. CONCLUSION

In this paper, our group proposes and designs a lightweight convolutional neural network for recognizing facial expressions. Our network model reduces the number of parameters in the convolutional layer by eliminating the fully connected layer, combining the residual depth-wise separable convolution, and adding the $\mathcal{4}2$-norm regularization term. And our model has no obviously adverse effect on detection and classification. Our model obtains the good detection results by identifying images outside the dataset, which proves that the model designed in this paper is suitable for multi-classification of facial expressions. In general, we have realized a visual system that can be integrated on devices with low computing power to achieve facial expression classification and reduce a large number of parameters. After comparing with the models in recent years, the accuracy of our model is higher than theirs, and it has achieved good detection results in images outside the dataset from the experimental results.

Although our model has achieved some results, there may be a lot of noise in the facial expressions captured in real life, such as the images with too strong or too dark lights, blurred images, most of the face is blocked, and other factors that are not conducive to detection. In order to solve this kind of problem, we need to continue our efforts.

REFERENCES

[1] P. Ekman and W. V. Friesen, ''Constants across cultures in the face and emotion.,'' *J. Personality Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[2] I. J. Goodfellow *et al.*, ''Challenges in representation learning: A report on three machine learning contests,'' *Neural Netw.*, vol. 64, pp. 59–63, Apr. 2015.

[3] S. Lawrence, C. L. Giles, A. Chung Tsoi, and A. D. Back, ''Face recognition: A convolutional neural-network approach,'' *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.

[4] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, ''Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,'' *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[5] T. Chang, G. Wen, Y. Hu, and J. Ma, ''Facial expression recognition based on complexity perception classification algorithm,'' 2018, *arXiv:1803.00185*. [Online]. Available: http://arxiv.org/abs/1803.00185

[6] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, ''Local learning with deep and handcrafted features for facial expression recognition,'' *IEEE Access*,

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020
vol. 7, pp. 64827–64836, 2019.

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

[7] C. Du and S. Gao, ''Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network,'' *IEEE Access*, vol. 5, pp. 15750–15761, 2017.
[8] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, ''Facial expression recognition utilizing local direction-based robust features and deep belief network,'' *IEEE Access*, vol. 5, pp. 4525–4536, 2017.
[9] M. Z. Uddin, W. Khaksar, and J. Torresen, ''Facial expression recognition using salient features and convolutional neural network,'' *IEEE Access*, vol. 5, pp. 26146–26161, 2017.
[10] D. Amodei *et al.*, ''Deep speech 2: End-to-end speech recognition in english and mandarin,'' in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 173–182.
[11] C. Szegedy, S. Ioffe, and V. Vanhoucke, ''Inception-v4, inception-ResNet and the impact of residual connections on learning,'' in *Proc. 31st AAAI Conf. Artif. Intell.*, 2016, p. 1.
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[13] M. R. Koujan, A. Akram, P. McCool, J. Westerfeld, D. Wilson, K. Dhaliwal, S. McLaughlin, and A. Perperidis, ''Multi-class classification of pulmonary endomicroscopic images,'' in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1574–1577.
[14] O. Leonovych, M. R. Koujan, A. Akram, J. Westerfeld, D. Wilson, K. Dhaliwal, S. McLaughlin, and A. Perperidis, ''Texture descriptors for classifying sparse, irregularly sampled optical endomicroscopy images,'' in *Proc. Annu. Conf. Med. Image Understand. Anal.* Cham, Switzerland: Springer, 2018, pp. 165–176.
[15] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, ''Self-attentive feature-level fusion for multimodal emotion detection,'' in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 196–201.
[16] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, ''Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5866–5870.
[17] M. E. Kret, K. Roelofs, J. J. Stekelenburg, and B. de Gelder, ''Emotional signals from faces, bodies and scenes influence observers' face expressions, fixations and pupil-size,'' *Frontiers Human Neurosci.*, vol. 7, p. 810, Dec. 2013.
[18] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, ''Deep affect prediction inthe-wild: Aff-wild database and challenge, deep architectures, and beyond,'' *Int. J. Comput. Vis.*, vol. 127, pp. 1–23, Jun. 2019.
[19] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, ''Analysing affective behavior in the first ABAW 2020 competition,'' 2020, *arXiv:2001.11409*. [Online]. Available: http://arxiv.org/abs/2001.11409
[20] W. Y. Choi, K. Y. Song, and C. W. Lee, ''Convolutional attention networks for multimodal emotion recognition from speech and text data,'' in *Proc. Grand Challenge Workshop Hum. Multimodal Lang. (Challenge-HML)*, 2018, pp. 28–34.
[21] E. Marinoiu, M. Zanfir, V. Olaru, and C. Sminchisescu, ''3D human sensing, action and emotion recognition in robot assisted therapy of children with autism,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2158–2167.
[22] Z. Du, S. Wu, D. Huang, W. Li, and Y. Wang, ''Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition,'' *IEEE Trans. Affect. Comput.*, early access, Sep. 10, 2019, doi: 10.1109/TAFFC.2019.2940224.
[23] J. Yang, K. Wang, X. Peng, and Y. Qiao, ''Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction,'' in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 594–598.
[24] P. Barros, N. Churamani, and A. Sciutti, ''The FaceChannel: A light-weight deep neural network for facial expression recognition,'' in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Buenos Aires, AR, USA, Apr. 2020 pp. 449–453.
[25] M. Koujan, L. Alharbawee, G. Giannakakis, N. Pugeault, and A. Roussos, ''Real-time facial expression recognition 'in the wild' by disentangling 3D expression from identity,'' in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Buenos Aires, AR, USA, May 2020 pp. 539–546.
[26] S. Xiao, P. Ting, and R. Fu-Ji, ''Facial expression recognition usingROI-KNN deep convolutional neural networks,'' *Acta Automatica Sinica*, vol. 42, no. 6, pp. 883–891, 2016.

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

[27] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, ''Deeply learning deformable facial action parts model for dynamic expression analysis,'' in*Proc. ACCV*. Cham, Switzerland: Springer, 2014, pp. 143–157.

[28] G. Zhao, H. Yang, and M. Yu, ''Expression recognition method basedon a lightweight convolutional neural network,'' *IEEE Access*, vol. 8,pp. 38528–38537, 2020.

[29] A. F. Abate, P. Barra, S. Barra, C. Molinari, M. Nappi, and F. Narducci, ''Clustering facial attributes: Narrowing the path from soft to hard biometrics,'' *IEEE Access*, vol. 8, pp. 9037–9045, 2020, doi: 10.1109/ACCESS.2019.2962010.

[30] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ''Rethinking the inception architecture for computer vision,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[32] F. Chollet, ''Xception: Deep learning with depthwise separable convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016,pp. 1251–1258.

[33] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*,Jun. 2016, pp. 770–778.

[34] G. A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, ''Mobilenets: Efficient convolutional neural networks for mobile vision applications,'' *CoRR*, vol. abs/1704.04861,pp. 1–9, Apr. 2017.

[35] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, ''Joint face detection and alignment using multitask cascaded convolutional networks,'' *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[36] D. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[37] M. Lin, Q. Chen, and S. Yan, ''Network in network,'' 2019, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400

[38] J. Z. Ji, *Application Research on Weakly-Supervised Learning in Computer Vision*. Chengdu, China: Univ. of Electronic Science and Technology of China, 2011.

[39] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, ''Striving for simplicity: The all convolutional net,'' 2014, *arXiv:1412.6806*. [Online]. Available: http://arxiv.org/abs/1412.6806

[40] X. Glorot, A. Bordes, and Y. Bengio, ''Deep sparse rectifier neural networks,'' in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[41] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, ''Hand-crafted feature guided deep learning for facial expression recognition,'' in *Proc.13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018,pp. 423–430.

[42] H. Wang and S. Hou, ''Facial expression recognition based on the fusion of CNN and SIFT features,'' in *Proc. IEEE 10th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2020, pp. 190–194, doi: 10.1109/ICEIEC49280.2020.9152361.

[43] A. Mollahosseini, D. Chan, and M. H. Mahoor, ''Going deeper in facial expression recognition using deep neural networks,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Placid, NY, USA, Mar. 2016,pp. 1–10, doi: 10.1109/WACV.2016.7477450.

[44] S. Miao, H. Xu, Z. Han, and Y. Zhu, ''Recognizing facial expressions using a shallow convolutional neural network,'' *IEEE Access*, vol. 7, pp. 78000–78011, 2019, doi: 10.1109/ACCESS.2019.2921220.

[45] K. Liu, M. Zhang, and Z. Pan, ''Facial expression recognition with CNN ensemble,'' in *Proc. Int. Conf. Cyberworlds (CW)*, Chongqing, China, Sep. 2016, pp. 163–166, doi: 10.1109/CW.2016.34.

[46] Z. Yi-Kui and L. Jian, ''Facial expression recognition based on transferring convolutional neural network,'' *J. Signal Process.*, vol. 34, no. 6,pp. 729–738, 2018.

[47] Y. Fang, ''Research of facial expression recognition based on convolutional neural network,'' M.S. thesis, Dept. Comput. Softw. Comput. Appl., XidianUniv., Xi'an, China, 2017.

[48] L. L. Xu, S. M. Zhang, and J. L. Zhao, ''Expression recognition algorithm for parallel convolutional neural networks,'' *J. Image Graph.*, vol. 24, no. 2,pp. 0227–0236, 2019.

[49] H. Ma and T. Celik, ''FER-Net: Facial expression recognition using densely connected convolutional network,'' *Electron. Lett.*, vol. 55, no. 4, pp. 184–186, Feb. 2019, doi: 10.1049/el.2018.7871.

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020