IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 2, 2020, accepted December 14, 2020, date of publication December 21, 2020,
date of current version December 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3046225

Recognition of Teachers' Facial Expression Intensity Based on Convolutional Neural Network and Attention Mechanism

KUN ZHENG, DONG YANG, JUNHUA LIU, AND JINLING CUI
Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

Corresponding authors: Junhua Liu (liujunhua@bjut.edu.cn) and Jinling Cui (cuijnling@bjut.edu.cn)

This work was supported in part by the Beijing Education Science Planning Project, China, under Grant CADA18069.

ABSTRACT The evaluations of traditional teaching quality are mainly subjective, and there is a lack of fine-grained objective data to support the evaluation of teaching states in the classroom. In this paper, an intensity-based facial expression dataset is proposed and named EIDB-13, which contains 13 kinds and 10393 facial images collected from thousands of individuals and existing facial expression datasets. Convolutional neural network (CNN) and attention mechanism are combined to recognize facial expressions. Migration learning is used to solve over-fitting problem in the process of training deep network based on the small sample dataset. InceptionResNetV2 is employed as migration network. Furthermore, an InceptionResNetV2+CBAM network proposed extract similar feature information among facial expressions and it outperforms the network without attention mechanisms. Experiments show a classification accuracy rate of 78% on the intensity-based facial expression dataset EIDB-13 and of 88% on the public macro expression dataset RAF-DB. Combining facial expression recognition technology into teaching is a key foundation to study teaching quality on the intensity of teacher's expression.

INDEX TERMS Attention mechanism, convolutional neural network, expression recognition, intensity of facial expression.

I. INTRODUCTION

In the traditional classroom, students' knowledge acquisitions are inseparable from the teacher's performance. Relevant practitioners in the field of education began to reach a consensus that teacher evaluation is the key factor to improve the quality of teaching and professional development [1]. At present, colleges mainly use the method of anonymous evaluation to assess teachers' performance. The age, gender, skin color and even attractiveness of teachers have an impact on the evaluation of teaching, which has strong subjectivity and blindness [2], [3]. Due to the epidemic COVID-19, E-Learning has been more widely practiced than ever before. However, the teaching evaluation method has not been applied to E-Learning. Regardless of traditional or online teaching mode, the teaching evaluation methods are mainly subjective, and lack the support of fine-grained objective data.

Teacher is an emotional job like other caring professions. These emotions tend to emerge when teachers transact with students and have garnered the attention of a growing number of researchers [4]. However, the existing teaching mode is that teachers often pay more attention to the transmission of knowledge and ignore the expression of emotions in the teaching process. Psychologist Paul Ekman's research showed that the accuracy of mapping facial expressions to a single specific emotional state is 88% [5]. Teacher's facial expressions in the teaching process can reflect the teacher's emotions and affect students' emotions and concentration on teaching content.

Emotion is a multidimensional structure and it will make an impact on health and happiness. Emotional disharmony is a mainly stress factor for emotional work. The frequent occurrence of positive emotions is also helpful to enhance personal sense of achievement [6]. In 2014, S. PROSEN analyzed the most frequently expressed emotions in terms of teacher's verbal expression and students' reaction to happiness, anger, sorrow, and joy. They found that happy expressions have a better performance effect than angry expressions in teacher-student interaction [7]. Therefore, studying facial expression changes of teachers in the teaching process is extremely meaningful to teachers and students.

In addition, the change of facial expression intensity and frequency cannot be ignored when the expression is

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

transmitted as a signal. In 2019, Lindsey [8] found that when children have a higher frequency and intensity of happy expressions during the interaction process, the score of social skills will be higher. As the strong intensity of anger and sadness expressions appearing during the interaction process, students will have lower social skills in the future growth process. Gupta *et al.* [9] analyzed four different emotions included high positive emotions, low positive emotions, high negative emotions and low negative emotions of students. Expert evaluation and sentiment analysis are used as feedback to teachers to improve teaching strategies, thereby increasing the learning efficiency of students. Therefore, it is not difficult for us to find that studying the changes of intensity and frequency of expressions could play an indispensable role in improving the quality of teaching.

People have strong subjectivity about images or videos. Evaluating people's emotion with computer is able to avoid being subjective and one-sided. Facial expression recognition classifies expressions by extracting facial features in images or videos [10]. In 2018, Yang *et al.* [11] proposed an emotion recognition model in the field of learning, which proved that emotion recognition based on facial expressions is feasible in distance education. Pei and Shan [12] applied the micro-expression recognition algorithm of face detection in teaching and proved it is reasonable. According to these above researches, it can be inferred that the method of facial expression recognition can be applied to obtain the expression information of teachers when their faces are detected.

The emergence of emotion recognition competitions such as FER2013 [13], Emotiw [14] and RAF-DB [15], [16] since 2013. These establishments of real scene datasets have been researched by a large number of scholars. In the task of facial expression recognition, the distribution of the expressions is usually unbalanced [17]. These expression datasets are great of diversity such as the change of head posture, the different illuminations of environment and the slight occlusion of the facial region. It is undoubtedly a challenge to extract features for the traditional methods such as Local Binary Pat-terns on Three Orthogonal Planes (LBP-TOP) [18], Pyramid Histogram of Oriented Gradients (PHOG) [19] and Local Quantized Patterns (LPQ) [20]. Pramerdorfer and Kampel [21] demonstrated that CNN outperforms the traditional methods in FER2013. Li *et al.* [15] designed the DLP-CNN algorithm based on RAF-DB dataset in 2017. By drawing on the method of manifold analysis, the neighbor relationship was constructed in the feature space, and the distance between classes was shortened to recognize expressions. Wang *et al.* [22] proposed to encode and assign weights to the overall features of the face in the study of facial expression features in 2020, and to characterize the importance of this overall feature. Wen *et al.* [23] proposed to add attention mechanism to ResNet named CBAM+ResNet, which improved the ability of feature extraction of the network. The above scholars solved the classification of macro facial expressions, but they ignored to divide the facial fine-grained expressions. And the accuracy of facial expression recognition of these methods could be improved by adding attention mechanism. So, in view of the above discussion, this article has made some work to deal with these problems. The main works of this paper are shown as follows:

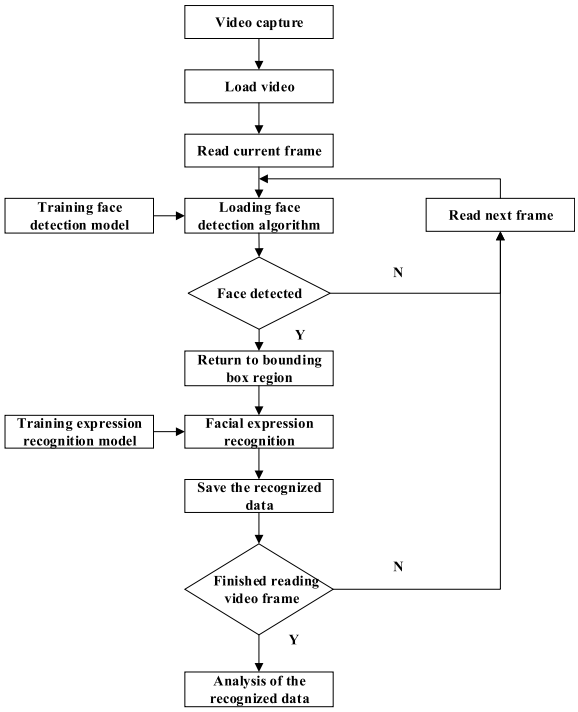

FIGURE 1. System design framework diagram.

1) In order to perform the fine-grained division of expression data, 13 types of expression intensity dataset EIDB-13 are proposed and established.

2) A new convolutional neural network model called InceptionResNetV2+CBAM that can reflect the facial expression information of characters is proposed. Using the transfer learning method, the InceptionResNetV2 network model is used to extract the deep features of the facial expression pictures, and the attention module CBAM [24] is inserted into the network to focus on details of the facial expression image.

3) Combine the face detection method in deep learning to detect the teacher's face and recognize his expression. The results can extract the teacher's transient expression strength information in the teaching video and provide data support for education researchers to study the influence of expression changes on the quality of teaching in the classroom.

II. MATERIALS AND METHODS

This paper designs an end-to-end facial expression intensity recognition system, which can detect and recognize the corresponding facial expressions of teachers in the teaching videos. The system framework is shown in FIGURE 1.

As shown in FIGURE 1, the system reads the collected teaching process video by frame and detects the character's face on the current frame. If the face is undetected, the system will read the next frame. If the face is detected, the face will be sent to the facial expression classifier that we have

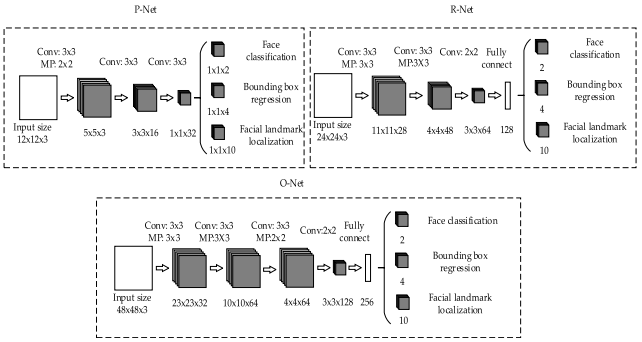

FIGURE 2. The architectures of P-Net, R-Net, and O-Net, where "MP" means max pooling and "Conv" means convolution. The step size in convolution and pooling is 1 and 2, respectively [25].

TABLE 1. Training accuracy rate of each network module of MTCNN.

Cascade CNN	Validation Accuracy
P-Net	94.6%
R-Net	95.2%
O-Net	95.2%

been trained for classification. Repeat the above steps until all frames have been read.

A. DESIGN OF FACE DETECTION PART

The facial expression information collection system of teachers in the classroom should extract the teacher's face information in real time and generate the detection regression box clearly and effectively. Therefore, the face detection model must be able to perform well in the teaching video with teacher's face. Based on the above considerations, MTCNN (Multi-Task Convolutional Neural Network) [25] is selected as the face detection algorithm.

1) DATA COLLECTION

The dataset for face detection comes from the Wider face database [26], which covers faces with different lighting, postures, and slight occlusion. The face key point dataset LFW [27] covers the coordinate information of 5 key point positions under different face poses.

2) MODEL TRAINING

Firstly, to adapt to the face of different sizes, the network uses the image pyramid pooling. Secondly, it initially generates the face through the CNN model called P-Net to candidate boxes and boundary regression vectors and use the method of non-maximum suppression to remove duplicate candidate boxes. Thirdly, it will improve the candidate boxes through the CNN model called R-Net model. Finally, it uses the CNN model called O-Net model to output the position of the face and 5 feature points. The functional block diagram of MTCNN is shown in FIGURE 2.

The training accuracy of each network module of the algorithm is shown in TABLE 1.

As shown in TABLE 1, MTCNN has a high accuracy rate in all stages of CNN model, which can well meet the needs of face detection in teaching videos.

B. DESIGN OF EXPRESSION RECOGNITION PART

In the stage of expression recognition, we first collect and preprocess the dataset to be trained. Secondly, the convolution neural network is constructed as our feature extraction network, and finally a classifier is established to classify the expression.

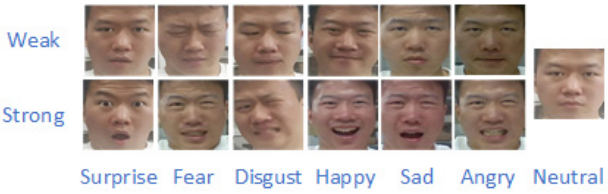

FIGURE 3. Thirteen kinds of expressions based on intensity (EIDB-13).

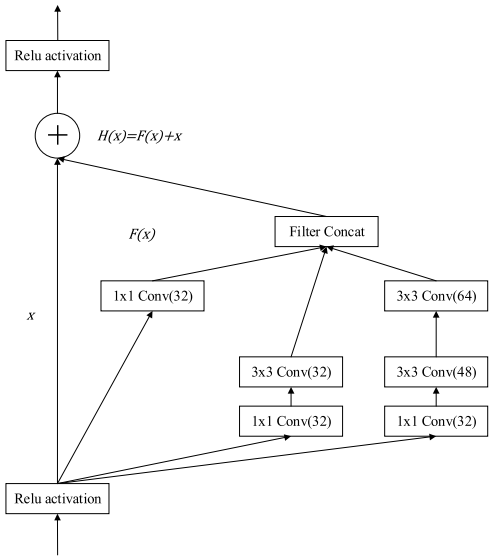

FIGURE 4. The InceptionResNet block's architecture.

1) DATA COLLECTION

Due to the lack of intensity-based expression datasets in real scenes. This experiment uses the recognized 7 types of basic expressions including surprise, fear, disgust, happy, sad, angry expressions and neutral expression as macro expressions. The dataset is collected from the public datasets such as FER-13 [13], RAF-DB [16], CelebA [28]. The facial muscle changes of non-neutral expressions are divided into strength and weakness that labeled by our Lab team, thereby generating 13 types of expression strength dataset. The dataset contains 10393 images in the training set and 1164 images in the test set. As shown in FIGURE 3.

As can be seen in FIGURE 3, Each column from left to right includes Surprise, Fear, Disgust, Happy, Sad, Angry, and Neutral expressions in sequence; in addition to Neutral expressions, the two rows from top to bottom correspond to expressions of weak and strong intensity respectively.

2) DATA PREPROCESSING

Because of the relatively small number of expression intensity dataset samples, we designed a very deep network with a total of 204 convolution layers that is calculated by accumulating the convolution layers from each block of InceptionResNetV2+CBAM.The architecture of network is shown in Figure 6. In order to avoid over fitting during

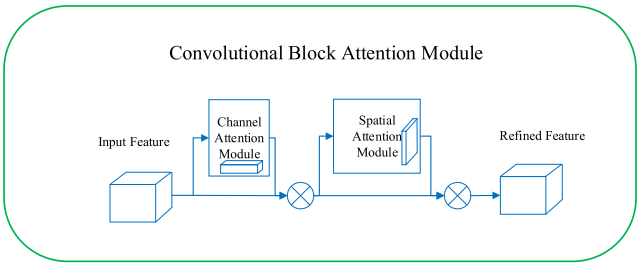

FIGURE 5. CBAM attention mechanism module [24].

training, we use horizontal flip, vertical flip, translation transformation, rotation transformation and other methods to expand the number of samples.

3) MODEL DESIGN AND TRAINING

Szegedy *et al.* [32] recommended combining the Inception architecture with the residual connections. It was verified by experiments that training with residual connections could significantly accelerate the training speed of the Inception Network. The block of Inception architecture with residual connections was called InceptionResNet. The block's architecture is shown in Figure 5. The distributed convolution layer convolution operation $F(x)$ is used to reduce the number of parameters and speed up the calculation. At the same time, fast connection is performed for feature x. The output of the residual module is $H(x) = F(x) + x$. The learning goal is transformed into residual learning $F(x) = H(x) - x$. By learning to minimize the residual error $F(x)$ to train the weights. The accuracy does not decrease when deepening network.

We adopt the idea that using migration learning for reference on small sample datasets and migrate the network InceptionResNetV2 to EIDB-13. In the network layers, we insert the CBAM (Convolutional Block Attention Module) attention mechanism module [24], which has better performance and better interpretability than its baseline network, pays more attention to the important features of the target and suppresses unnecessary features. The network structure diagram of the attention CBAM module is shown in FIGURE 5.

As shown in FIGURE 5, Firstly, the baseline network is given an intermediate feature map: $\boldsymbol{F} \in \boldsymbol{R}^{\boldsymbol{C}\times\boldsymbol{H}\times\boldsymbol{W}}$ as input. Secondly, this module sequentially connects the one-dimensional channel attention map $\boldsymbol{M_c} \in \boldsymbol{R}^{\boldsymbol{C}\times\boldsymbol{1}\times\boldsymbol{1}}$ and two-dimensional spatial attention map $\boldsymbol{M_s} \in \boldsymbol{R}^{\boldsymbol{1}\times\boldsymbol{H}\times\boldsymbol{w}}$. Finally, the refined feature is outputted, so the whole attention process can be summarized as (1).

$$\begin{aligned} F' &= M_c(F) \otimes F, \\ F'' &= M_s\left(F'\right) \otimes F', \end{aligned} \tag{1}$$

The special character $\otimes$ denotes multiplication by elements. In the process of multiplication, attention value is propagated, and channel attention value propagates along spatial dimension and vice versa. $\boldsymbol{F''}$ is the final optimized output. $\boldsymbol{M_c}(\boldsymbol{F})$ means that the average pooling and maximum pooling operations are used to aggregate the spatial information of feature maps to generate two different spatial context descriptors: $\boldsymbol{F^c_{avg}}$ and $\boldsymbol{F^c_{max}}$, which represent the average pooling characteristics and the maximum pooling characteristics respectively. Then the two descriptors are forwarded to a shared network to generate channel attention map $\boldsymbol{M_c} \in \boldsymbol{R}^{\boldsymbol{C}\times\boldsymbol{1}\times\boldsymbol{1}}$. The shared network is composed of a hidden layer multi-layer perceptron $\boldsymbol{MLP}$, which is the feature vector combined by the element level summation. In short, the formula of channel attention is calculated as (2).

$$\begin{aligned} M_c(F) &= \sigma\left(MLP\left(\text{AvgPool}\left(\text{F}\right)\right) + MLP\left(\text{MaxPool}\left(\text{F}\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{max}\right)\right)\right) \end{aligned} \tag{2}$$

The character $\boldsymbol{\sigma}$ is the sigmoid activation function. $\boldsymbol{W_0} \in \boldsymbol{R}^{\boldsymbol{C/r}\times\boldsymbol{C}}$, $\boldsymbol{W_1} \in \boldsymbol{R}^{\boldsymbol{C}\times\boldsymbol{C/r}}$ denote the two shared input weights of MLP respectively. In the spatial attention mapping $\boldsymbol{M_s}(\boldsymbol{F})$, two pooling operations are used to generate two two-dimensional maps to aggregate the channel information of the function map: $\boldsymbol{F^s_{avg}} \in \boldsymbol{R}^{\boldsymbol{1}\times\boldsymbol{H}\times\boldsymbol{W}}$ and $\boldsymbol{F^s_{max}} \in \boldsymbol{R}^{\boldsymbol{1}\times\boldsymbol{H}\times\boldsymbol{W}}$ denotes the average and maximum pooling characteristics of the whole channel, respectively. Then, a standard convolution layer is used to connect and convolute the two-dimensional spatial attention map. In short, the formula of spatial attention mapping is as (3).

$$\begin{aligned} M_s(F) &= \sigma\left(f^{2\times2}\left(\left[AvgPool\left(F\right); MaxPool\left(F\right)\right]\right)\right) \\ &= \sigma(f^{2\times2}([F^s_{avg}; F^s_{max}])) \end{aligned} \tag{3}$$

The character $\boldsymbol{f^{2\times2}}$ is the convolution operation with the convolution kernel size of $2 * 2$. After inserting the attention mechanism module into the convolution layer of the network, we can pay attention to the salient parts of the feature graph so as to achieve our intention of adding it. The network module diagram of InceptionResNetV2 + CBAM after adding CBAM module is shown in FIGURE 6.

The model's training network diagram is shown in FIGURE 6. The CBAM module is inserted after the InceptionResNet-C volume build-up layer. And then we add L2 regularization layer, Dropout layer, Flatten layer and Softmax layer as the network and classifier. In the model training stage, we use the Adam optimizer to update the weights and use batch data for training. In the model training stage, we adopt the Early Stopping mechanism, which can alleviate the over-fitting problem to a certain extent and reduce the complexity of the model. The accuracy on the validation set did not improve after 50 epochs, and the learning rate became 0.1 times. After the accuracy did not improve on the validation set or the loss function value no longer decreased in 200 epochs, the training network saved the model with the best accuracy automatically. In the weight update phase of back propagation, we choose the cross-entropy loss function as the optimization function of the model.

III. RESULTS

The comparison between the training results of InceptionResNetV2 network model on RAF-DB and other network models are shown in TABLE 2.

TABLE 2. Comparison of the basic model on the RAF-DB validation set.

Model	Validation Accuracy
VGG16 [29]	78.19%
ResNet50 [30]	82.56%
InceptionV3 [31]	84.45%
Xception [34]	85.38%
InceptionResNetV2 [32]	86.32%

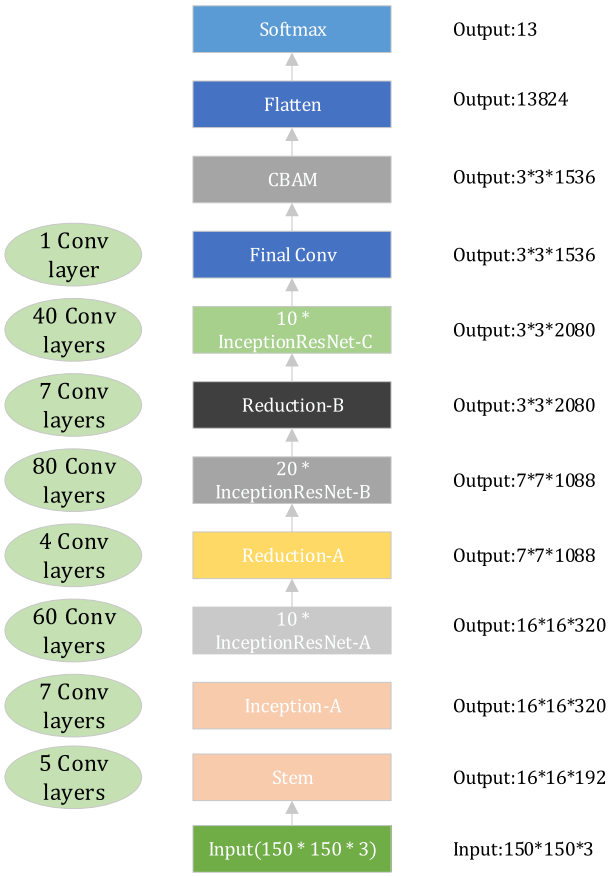

FIGURE 6. InceptionResNetV2+CBAM network diagram.

TABLE 3. Comparison between this MODEL and other methods.

Model	Validation Accuracy
VGG16+CBAM [24]	80.28%
DLP-CNN [16]	84.13%
CBAM+ResNet50 [23]	84.53%
gACNN [35]	85.07%
InceptionV3+CBAM	86.26%
RAN+ResNet18 [33]	86.90%
Xception+CBAM	87.23%
InceptionResNetV2+CBAM(Ours)	88.18%

Compared with other expression recognition methods, the result of training on RAF-DB dataset with CBAM attention mechanism added to the InceptionResNetV2 network is better. The results are shown in TABLE 3.

It can be seen that the performance of the InceptionResNetV2 + CBAM network in RAF-DB training process is better than that of the common baseline network and the methods proposed by other scholars in recent years. The confusion matrix of our method on RAF-DB test set is shown

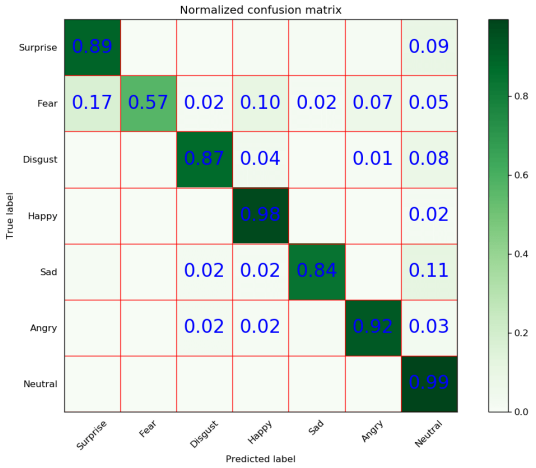

FIGURE 7. Confusion matrix of InceptionResNetV2 + CBAM on RAF-DB test set.

in FIGURE 7. The average accuracy of our method is as high as 86.42%.

As shown in FIGURE 7, the recognition accuracy of happy, sad, and neutral on the test set is significantly higher than others, but the accuracy of disgust is lower than the average. We think the reason for this problem is that the number of different expressions in the dataset is unbalanced. For example, there are 4342 happy pictures but only 260 fear pictures. The average number of pictures in each category is 1753, so common features cannot be extracted. There are few pictures in the test set, which is easy to be confused with other categories, leading to errors in classification. When we apply this method to EIDB-13, the feature difference between different expressions with weak intensity is small. It is difficult for the loss of the validation set in the network to converge to the ideal value. Therefore, this article uses the above seven types of macro expressions for multi-classification, and then classifies the non-neutral six types of expressions into strong and weak categories. The experimental results are shown in FIGURE 8, the average accuracy rate is as high as 78%.

As shown in FIGURE 8, the average correct rate of our proposed network on the EIDB-13 test set could reach about 78%. And when we train the benchmark network in the EIDB-13 test set, we find that the accuracy rate of our proposed method is much higher than that of the ordinary model. The comparison table of each model is shown in TABLE 4.

Based on the above results, which can meet the task of teacher's expression recognition based on intensity in normal classroom scenarios. However, the weak fear expression samples are less, it is easy to be confused with other types.

IV. DISCUSSION

We selected a number of online teaching videos published on the MOOC website and performed corresponding face detection and expression recognition for teachers among them. And save the teacher's identity, facial expressions and time information in the video to a csv file. And generate a

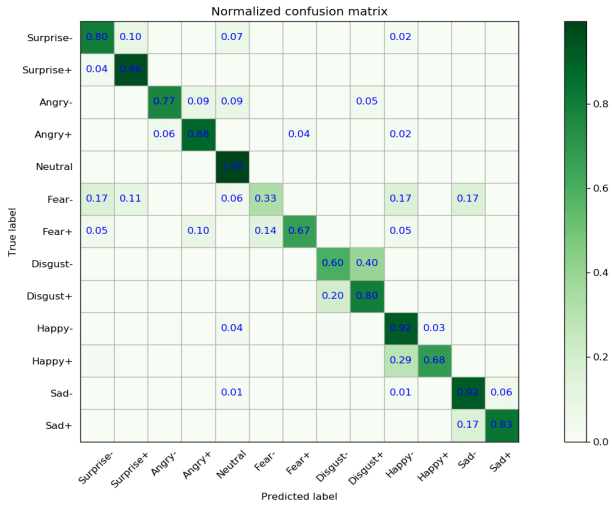

FIGURE 8. Confusion matrix of InceptionResNetV2+CBAM on EIDB-13 test set.

TABLE 4. Comparison on EIDB-13 between the network and other methods.

Model	Validation Accuracy
VGG16+CBAM [24]	71.44%
DLP-CNN [16]	73.50%
CBAM+ResNet50 [28]	74.97%
gACNN [35]	75.88%
InceptionV3+CBAM	76.07%
RAN+ResNet18 [33]	76.52%
Xception+CBAM	77.39%
InceptionResNetV2+CBAM (ours)	78.15%

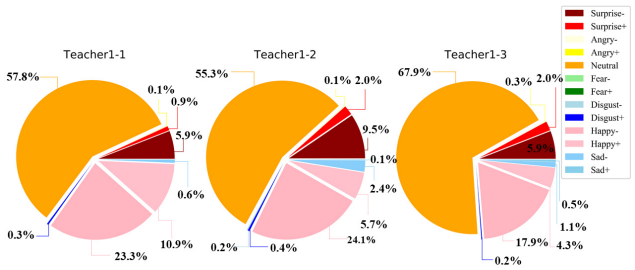

FIGURE 9. The distribution of expression intensity of the same teacher under the same subject.

teacher's facial expression ratio map based on intensity. When the teacher's identity and the subjects taught do not change, the expression intensity distribution under different courses taught is shown in FIGURE 9.

We find that in FIGURE 9, the proportions of expressions of the same teacher in teaching are almost similar. When the teacher's identity changes, the expression intensity distribution under different teachers' courses is shown in FIGURE 10.

We found that in FIGURE 10, the expression ratios of different teachers in the teaching process are often very different. This is of exploratory significance for us to further study the changes of teachers' expressions in teaching scenarios. When we use the same model to extract information from one of

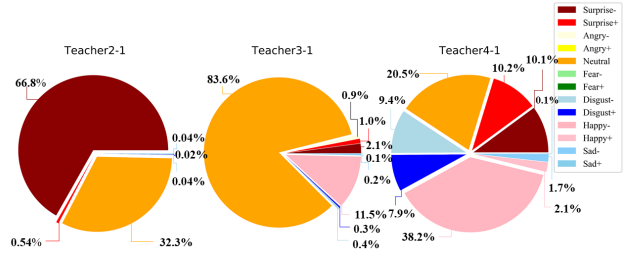

FIGURE 10. The distribution of facial expression intensity of different teachers in the same course.

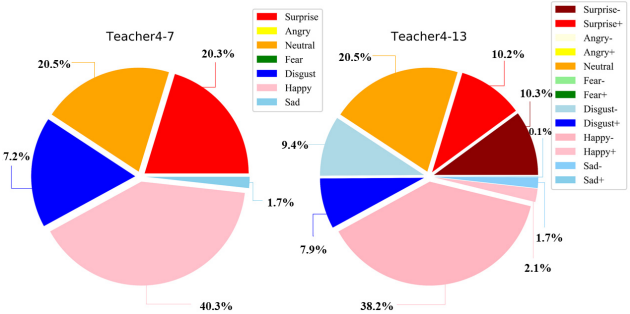

FIGURE 11. Distribution of 7 types of macro expressions and 13 types of expressions with intensity.

the teachers in the video with 7 types of macro expressions and 13 types of expressions based on intensity, their intensity distribution is shown in FIGURE 11.

Through the fine-grained intensity division of teacher's expressions, it can be found that expression pie charts with intensity display more information than expression pie charts without intensity. The results can provide data support for educational researchers to study the impact of different facial expressions in the classroom. Optimizing the large model parameters should be the key goal in our next step, so as to accelerate the training speed. At the same time, teacher's body movements and voice could also convey emotion in the teaching scene. Accordingly, the teaching emotion should be evaluated objectively through more aspects instead of single facial expression.

V. CONCLUSION

This paper takes teachers' expression intensity in real scenes as the research object based on deep learning. Firstly, a dataset of 13 kinds of expression based on intensity is proposed and constructed. And we apply it as the training set to classify the teachers' facial expression in the classroom. Secondly, a recognition algorithm based on convolutional neural network InceptionResNetV2 and attention mechanism module CBAM is proposed. Through the detection of the frequency and intensity of teachers' facial expression in the classroom, we can understand the positive degree of teachers' emotion in the teaching situation. It can not only provide an objective reference for teaching assessment, but also be used to analyze students' interest in teaching content. On the verification set of EIDB-13, the recognition accuracy can reach 78%. The result shows that it can divide teachers' facial

expression more precisely. Finally, the system combining with the face detection algorithm MTCNN can detect the face of the character in real time and recognize the corresponding facial expression. However, more fine-grained segmentation of emotional expression in real scenes needs further research. The robustness of deep learning method still needs to be further improved in extracting feature when it is applied to facial expression recognition with finer granularity. Meanwhile, small sample and large task intelligent recognition based on facial expression is also one of the research directions in the future.

REFERENCES

[1] D. F. Warring, "Teacher evaluations: Use or misuse," *Universal J. Educ. Res.*, vol. 3, no. 10, pp. 703–709, 2015.

[2] B. P. Smith and B. Hawkins, "Examining student evaluations of black college faculty: Does race matter," *J. Negro Edu.*, vol. 80, no. 2, pp. 149–162, 2011.

[3] S. L. Sohr-Preston, S. S. Boswell, K. McCaleb, and D. Robertson, "Professor gender, age, and 'hotness' in influencing college students' generation and interpretation of professor ratings," *Higher Learn. Res. Commun.*, vol. 6, no. 3, pp. 1–21, 2016.

[4] P. A. Schutz, "Inquiry on teachers' emotion," *Educ. Psychologist*, vol. 49, no. 1, pp. 1–12, 2014.

[5] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[6] D. Zapf and M. Holz, "On the positive and negative effects of emotion work in organizations," *Eur. J. Work Organizational Psychol.*, vol. 15, no. 1, pp. 1–28, 2006.

[7] S. Prosen, H. S. Vitulic, And O. P. Skraban, "Teachers' emotional expression in the classroom," *Hacettepe Üniversitesi Egitim Fakültesi Dergisi*, vol. 29, no. 1, pp. 226–237, 2014.

[8] E. W. Lindsey, "Frequency and intensity of emotional expressiveness and preschool children's peer competence," *J. Genetic Psychol.*, vol. 180, no. 1, pp. 45–61, 2019.

[9] S. K. Gupta, T. S. Ashwin, and R. M. R. Guddeti, "Students' affective content analysis in smart classroom environment using deep learning techniques," *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 25321–25348, 2019.

[10] Y. Sun and Q. Li, "The application of deep learning in mathematical education," in *Proc. 1st IEEE Int. Conf. Knowl. Innov. Invention (ICKII)*, Jul. 2018, pp. 130–133.

[11] D. Yang, A. Alsadoon, P. C. Prasad, A. K. Singh, and A. Elchouemi, "An emotion recognition model based on facial recognition in virtual learning environment," *Procedia Comput. Sci.*, vol. 125, pp. 2–10, Jan. 2018.

[12] J. Y. Pei and P. A. Shan, "Micro-expression recognition algorithm for students in classroom learning based on convolutional neural network," *Traitement Du Signal*, vol. 36, no. 6, pp. 557–563, 2019.

[13] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. H. Lee, and Y. Zhou, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.* Daegu, South Korea: Springer, 2013, pp. 117–124.

[14] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2016, pp. 427–432.

[15] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR*, Jul. 2017, pp. 2584–2593.

[16] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[17] Z. Luo, J. Hu, and W. Deng, "Local subclass constraint for facial expression recognition in the wild," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3132–3137.

[18] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[19] V. Ojansivu and J. Heikkila, "Blur insensitive texture classification using local phase quantization," in *Proc. Image Signal Process. 3rd Int. Conf., ICISP*, Cherbourg-Octeville, France, 2008, pp. 236–243.

[20] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. 6th ACM Int. Conf. Image Video Retr.*, 2007, pp. 401–408,

[21] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," 2016, *arXiv:1612.02903*. [Online]. Available: http://arxiv.org/abs/1612.02903

[22] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6897–6906.

[23] P. Wen, Y. Ding, Y. Wen, Z. Deng, and Z. Xu, "Facial expression recognition method based on convolution neural network combining attention mechanism," in *Proc. Int. Conf. Artif. Intell. Secur. ICAIS*, Hohhot, China, 2020, pp. 136–147.

[24] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[25] 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.

[26] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5525–5533.

[27] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, Marseille, France, 2008, pp. 1–11.

[28] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3730–3738.

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

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

[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.*, Jun. 2016, pp. 2818–2826.

[32] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: http://arxiv.org/abs/1602.07261

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

[34] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jul. 2017, pp. 1251–1258.

[35] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

KUN ZHENG was born in Hebei, China, in 1977. He received the B.Sc. degree in electronic engineering from Hebei University, Baoding, China, in 2001, and the M.Sc. degree in software engineering and the Ph.D. degree in electronic engineering from the Beijing University of Technology, Beijing, China, in 2006 and 2018, respectively. He is currently an Associate Professor with the Beijing University of Technology. His current research interests include neural networks and applications, image processing, and smart classroom.

DONG YANG was born in Jiangsu, China, in 1996. He received the B.Sc. degree from the Huaiyin Institute of Technology, Huai'an, China, in 2018. He is currently pursuing the M.Sc. degree with the Beijing University of Technology, China.

JUNHUA LIU was born in Shandong, China, in 1974. He received the M.S. degree in electronic engineering from the Beijing Institute of Fashion Technology, Beijing, in 2004, and the Ph.D. degree in microelectronics and solid state electronics from the University of Chinese Academy of Sciences, Beijing, in 2009.

Since 2009, he has been an Assistant Professor with the Electronic Engineering Department, Beijing University of Technology, Beijing. His research interests include high-performance and high-efficiency compute units and compiler optimizations for the hardware description language.

JINLING CUI was born in Shandong, China, in 1985. He received the B.S. degree in measurement and control technology and instrument and the M.S. and Ph.D. degrees in geodetection and information technology from the China University of Geosciences, in 2007, 2010, and 2014, respectively.

He is currently a Lecturer with the Beijing University of Technology. His current research interests include neural networks and image processing.