future internet

MDPI

Article

Evaluation of Online Teaching Quality Based on Facial Expression Recognition

Changbo Hou, Jiajun Ai, Yun Lin *, Chenyang Guan, Jiawen Li and Wenyu Zhu

College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China; houchangbo@hrbeu.edu.cn (C.H.); aijiajun@hrbeu.edu.cn (J.A.); _victor_guan@hrbeu.edu.cn (C.G.); 1003453492@hrbeu.edu.cn (J.L.); zwy3743@hrbeu.edu.cn (W.Z.)

* Correspondence: linyun@hrbeu.edu.cn

Abstract: In 21st-century society, with the rapid development of information technology, the scientific and technological strength of all walks of life is increasing, and the field of education has also begun to introduce high and new technologies gradually. Affected by the epidemic, online teaching has been implemented all over the country, forming an education model of "dual integration" of online and offline teaching. However, the disadvantages of online teaching are also very obvious; that is, teachers cannot understand the students' listening status in real-time. Therefore, our study adopts automatic face detection and expression recognition based on a deep learning framework and other related technologies to solve this problem, and it designs an analysis system of students' class concentration based on expression recognition. The students' class concentration analysis system can help teachers detect students' class concentration and improve the efficiency of class evaluation. In this system, OpenCV is used to call the camera to collect the students' listening status in real-time, and the MTCNN algorithm is used to detect the face of the video to frame the location of the student's face image. Finally, the obtained face image is used for real-time expression recognition by using the VGG16 network added with ECANet, and the students' emotions in class are obtained. The experimental results show that the method in our study can more accurately identify students' emotions in class and carry out a teaching effect evaluation, which has certain application value in intelligent education fields, such as the smart classroom and distance learning. For example, a teaching evaluation module can be added to the teaching software, and teachers can know the listening emotions of each student in class while lecturing.

Keywords: artificial intelligence; online education; learning quality; face detection; expression recognition

check for updates

Citation: Hou, C.; Ai, J.; Lin, Y.; Guan, C.; Li, J.; Zhu, W. Evaluation of Online Teaching Quality Based on Facial Expression Recognition. *Future Internet* **2022**, *14*, 177. https://doi.org/10.3390/fi14060177

Academic Editor: Shuai Liu

Received: 18 May 2022
Accepted: 7 June 2022
Published: 8 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, information technologies, such as 5G and high-speed networks, have made significant progress, and information processing technologies represented by artificial intelligence and big data have been rapidly integrated into all aspects of human life. At the same time, people began to introduce artificial intelligence and computer vision into the field of education. Computer vision can be used to identify students' expressions and behaviors, thereby inferring students' participation in education. The emotions of learners in the learning process are precisely the key to improving the quality of teaching. Based on facial expression recognition technology, teachers can grasp the emotional state of students in time. According to the students' learning status, timely adjusted teaching strategies alleviate students' bad learning emotions and improve learning efficiency [1]. Affected by the epidemic, students in our country use the online classroom mode part of the time. In this mode, teachers cannot pay attention to each student's listening status in real-time. Therefore, this requires a technology that can collect students' listening status in real-time and give feedback to teachers in time so that teachers can customize personalized teaching by analyzing students' learning status data, stimulate students' enthusiasm for learning

and thinking, and return to the educational essence of "teaching students according to their aptitude".

Most of the teaching evaluation we have been using is the way of after-school investigation, which is not only cumbersome but also the information collected is not comprehensive and accurate. Moreover, we lack a complete and unified evaluation scheme to measure the teaching level of schools and teachers. In recent years, with the development of artificial intelligence, people have begun to apply face detection and expression analysis to classroom teaching evaluation. However, due to reasons, such as precision and accuracy, they have not been widely used in classrooms. Therefore, how to improve the accuracy of expression analysis is an urgent problem to be solved.

1.1. The Current State of Face Detection

With the breakthrough of Alexnet in the ImageNet competition, artificial intelligence technology represented by deep neural networks began to be applied in the field of image and object detection [2]. At present, the research on facial expression recognition (FER) technology mainly adopts the deep learning method because, compared with the traditional machine learning method [3], the convolutional neural network (CNN) can extract deeper and more abstract features. It has shown excellent performance in various image classification tasks [4]. Furthermore, the attention mechanism has been widely used in various computer vision tasks, such as saliency detection [5], crowd counting [6], and facial expression recognition. Benefiting from the extensive research and application of deep learning in the field of image and computer vision, as well as the disclosure of a large number of datasets, such as LFW [7] and Celeba [8], face detection and recognition algorithms based on deep learning are increasing day by day. For example, Deepface [9], proposed by Facebook at CVPR2014 in 2014, achieved 97.35% accuracy on LFW, and in 2015, the Face++ [10] adopted by Zhou Erjin et al. achieved 99.50% recognition accuracy on LFW, and, in 2015, FaceNet [11], launched by Google, achieved 99.63% accuracy on LFW and so on. Using deep learning for face detection can usually achieve higher accuracy. Introducing face detection and expression recognition based on deep learning into online teaching evaluation can provide teachers with more accurate classroom feedback in a timely manner.

1.2. The Status Quo of Expression Analysis

Facial Expression Recognition (FER) refers to the ability to intelligently recognize and understand human emotions by extracting facial expression features [12]. Happy et al. [13] and Majumder et al. [14] proposed that changes in facial expressions usually occur in some prominent facial regions, such as near the mouth, nose, and eyes, and the research direction has also shifted from full-face feature extraction to focus on the regions related to facial expressions [15]. In 2013, Google realized the technology of unlocking the phone with specific expressions, such as blinking the eyes to unlock the phone. In 2014, Emotient installed a facial expression recognition system on the mobile phones of clients. Through this system, it can judge the user's preference for different products so that it can selectively recommend products to users. In 2015, a tool designed by Microsoft to recognize human emotions based on facial expressions defined eight common human emotions: anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise. By comparing the possible values of each emotion type, the tool determines that the emotion type with the largest possible value is the classification of the current expression. In the same year, Xiao Ying of Huazhong University of Science and Technology used the eyebrows, eyes, and mouth as monitoring objects to record the changes of the three in the learning process of learners and designed a facial feature monitoring system to provide effective support for the process evaluation of teaching.

2. Materials and Methods

The system consists of four parts: a video acquisition module, face detection module, expression analysis module, and teaching evaluation module. The system composition is shown in Figure 1.

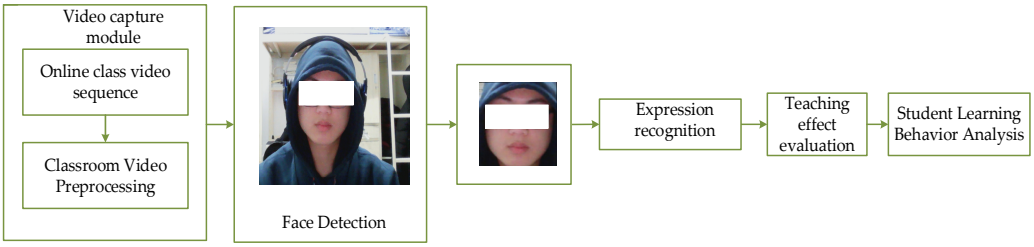


Figure 1. System composition.

The video acquisition module uses OpenCV to call the camera, which is responsible for collecting students' classroom videos in real-time for testing and preprocessing the collected videos. Face detection realizes the calibration of the face position of the image and extracts the facial feature area. The expression analysis module is responsible for performing expression analysis on the collected face images of the students to obtain the students' class emotions. The teaching evaluation module uses the students' class emotions to get the students' class situation and the teacher's teaching situation.

2.1. Data Set

Since the interpretation of students' expressions requires professionals to be able to interpret them correctly, the data set could not be self-made during the experiment in our study, and a public face data set was used for the experiment. The dataset of our study consists of two parts; one is the Kaggle public face dataset, Facial Expression Recognition 2013 (FER2013), for training the VGG16 [16,17] network, and the other is the CK+ dataset to verify the effectiveness of the algorithm. FER2013 consists of 35,886 facial expression pictures, of which 28,709 are in the training set, 3589 are in the public validation set, and 3589 are in the private validation set. There are a total of 7 kinds of expressions, namely anger, disgust, fear, happy, sad, surprise, and neutral. All kinds of expressions are shown in Figure 2, and each column is a type of expression. The CK+ dataset used in this study has a total of 123 experimenters, and the experiment uses a total of 981 labeled images for this experiment; the redundant background has been cropped out.

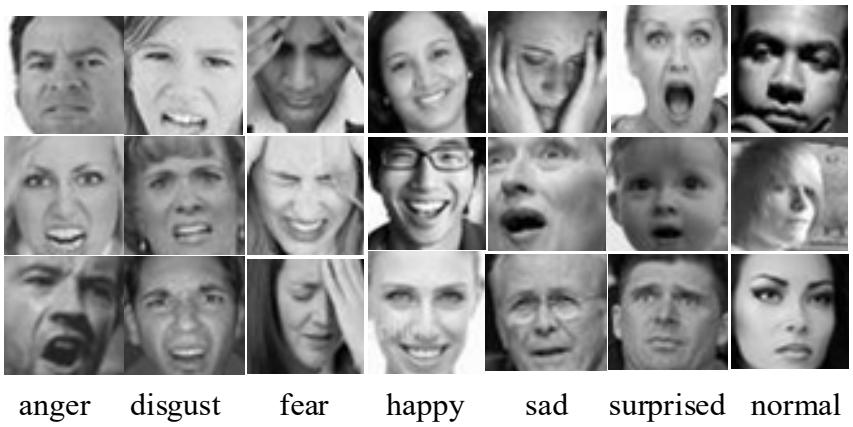


Figure 2. Schematic diagram of the FER2013 dataset.

2.2. Face Detection Algorithm

The face detection process is implemented using the MTCNN [18–20] algorithm. As shown in Figure 3, MTCNN consists of three cascaded convolutional neural network

architectures, namely P-Net, R-Net, and O-Net. P-Net is mainly responsible for quickly obtaining the regression vector of the candidate window and bounding box of the face region; it uses the bounding box for regression, calibrates the candidate window, and then merges the highly overlapping candidate boxes through a non-maximum suppression (NMS). R-Net removes misidentified candidate boxes through a bounding box regression and NMS. The network structure of R-Net has one more fully connected layer than P-Net, which has a better effect of suppressing a false positive. O-Net has more convolutional layers than R-Net, which makes its processing results more refined. The role of O-Net is the same as that of R-Net, but this layer performs more supervision on the face area and outputs 5 landmarks at the same time; then, it generates the final bounding box and outputs the final face feature key points [21].

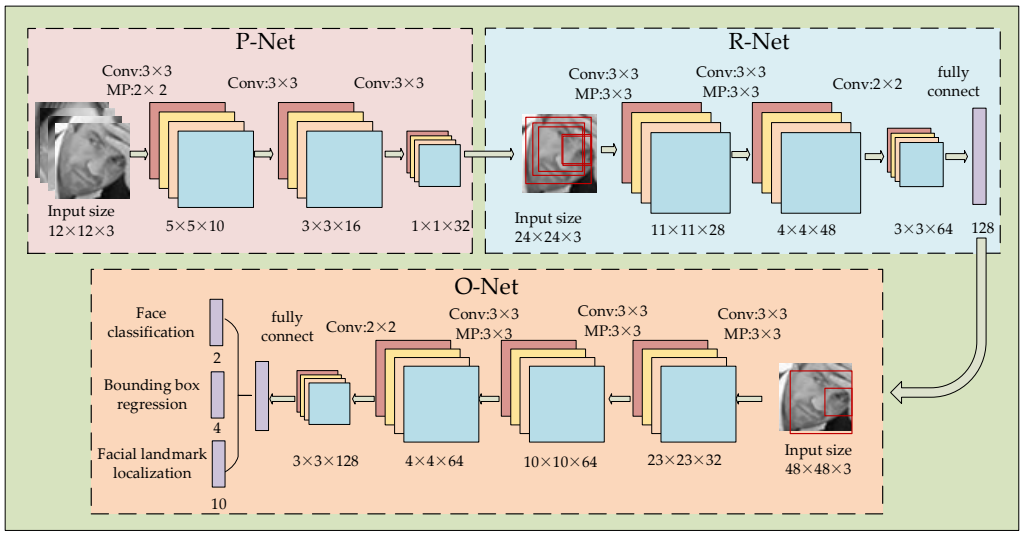


Figure 3. MTCNN network structure.

The specific process of face detection by MTCNN is as follows:

Step 1: Build an image pyramid. First, the image is scaled, rotated, and stretched at different scales, and an image pyramid is constructed to adapt to the detection of faces of different sizes.

Step 2: Input the image pyramid constructed in the first step into P-Net. P-Net is a proposal network for the face area. After inputting the features into $3 \times 3 \times 3$ convolutional layers, the face classifier determines whether the area is a human face and uses the bounding box regression and a facial key point locator to make a preliminary proposal of the face area. This part will eventually output many face areas that may have human faces, and these regions are fed into R-Net for further processing. The P-Net network will finally output a feature vector of size $1 \times 1 \times 32$.

Step 3: Input the features' output in the second step into the R-Net network; R-Net refines the input feature vector, removes the wrong box, and uses the border regression and facial key point locator again for face recognition. The bounding box regression and key point positioning of the region will finally output a more credible face region. R-Net uses a 128 fully connected layer at the end of the network to increase the accuracy of the network.

Step 4: Input the features' output in the third step into the O-Net network. O-Net uses a 256 fully connected layer at the end of the network, which retains more image features and further increases the accuracy of the network. At the same time, face discrimination, face region border regression, and face feature positioning are performed, and, finally, the coordinates of the upper left corner and the lower right corner of the face region and the five feature points of the face region are output.

2.3. Expression Recognition Algorithm

The expression recognition algorithm model is implemented using the VGG16 network with the addition of ECANet [22]. The authors of ECANet argue that SENet [23] brings a side effect to the prediction of the channel attention mechanism and that capturing all channel dependencies is inefficient and unnecessary. ECANet utilizes the good ability of convolution to obtain cross-channel information, removes the fully connected layer in the original SENet, and performs global average pooling on the feature map with an input size of H × W × C, and the feature map, after global average pooling, is learned directly through a 1D convolution. Figure 4 is a conventional SENet, and Figure 5 is ECANet. ECANet replaces two full connections with 1D convolutions.

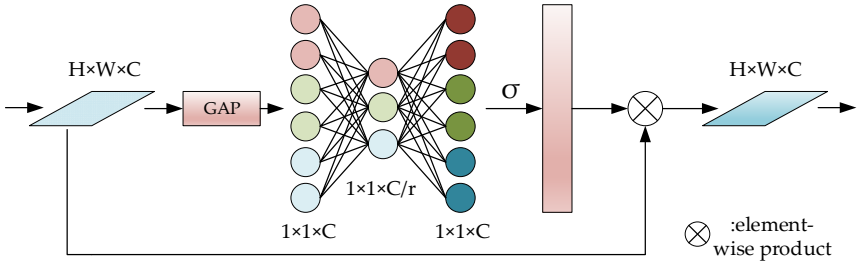


Figure 4. SENet network structure.

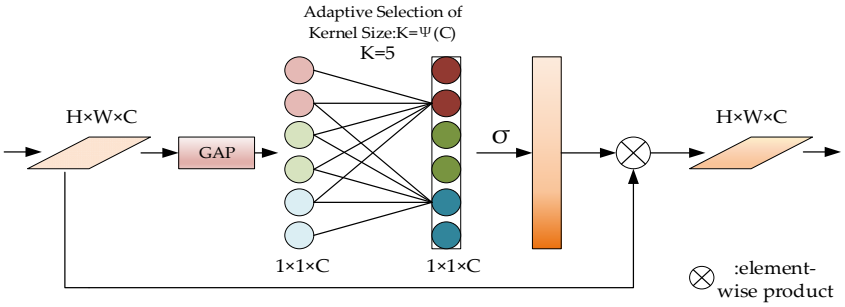


Figure 5. ECANet network structure.

Considering the aggregated features obtained by global average pooling (GAP), ECANet generates channel weights by rapidly performing a 1D convolution of size k, where k is adaptively determined by the channel dimension, C mapping.

The network structure of VGG16 is shown in Figure 6. VGG16 has a total of 16 layers, 13 convolution layers, and 3 fully connected layers. After the first two convolutions with 64 convolution kernels, one pooling is used, and the second after two convolutions with 128 convolution kernels, pooling is used, and after three convolutions with 512 convolution kernels are repeated twice, pooling is performed again, and, finally, three full connections are performed.

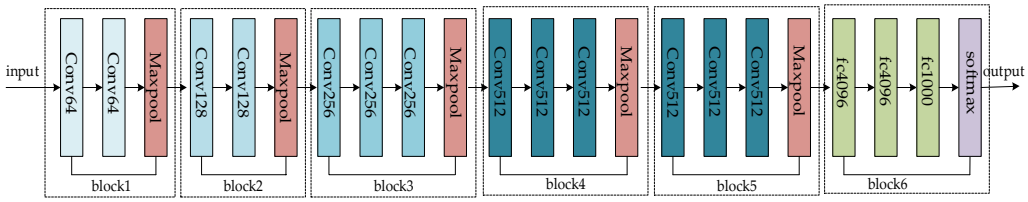


Figure 6. VGG16 network structure.

In this experiment, the method of migration learning is adopted, and the pre-training weights of VGG16 are added for training after the network is changed. The model structure of expression recognition is shown in Figure 7. The specific process is as follows: When

the facial expression image is input into the model, the feature map 1 is obtained through feature extraction through the VGG16 network, and then the feature map 2 output by the block4_conv3 layer of the VGG16 network is taken as the input of the ECANet network. In the ECANet network, feature map 2 is first subjected to Global Average Pooling, and then, through a 1D convolution of size K, the feature map output by 1D convolution is stacked with feature map 2 to obtain feature map 3. Finally, the feature map 3 output by the ECANet network is stacked with the feature map 1 to obtain the final feature map and inputs it into the classification module. The classification module of the network performs Global Average Pooling on the extracted features, converts the feature map output by the last convolutional layer into a 1024-dimensional vector, and, finally, outputs the probability values of 7 expression categories through SoftMax regression.

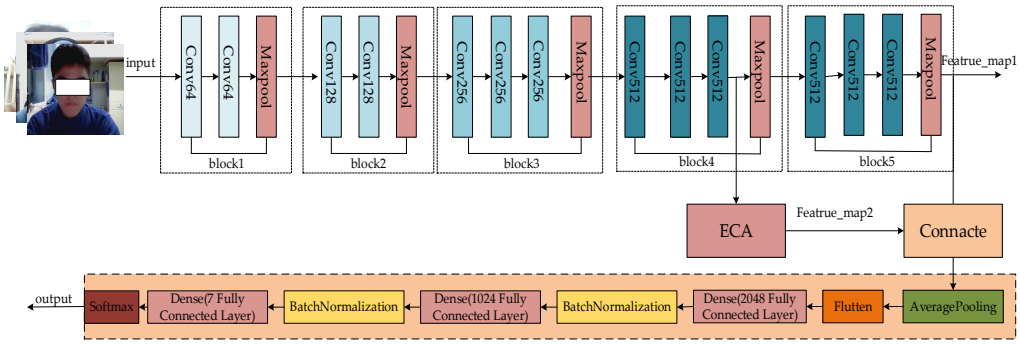


Figure 7. The overall structure of the model.

2.4. Classroom Teaching Evaluation Module

According to the Fer2013 dataset, the facial expressions are divided into seven types: namely angry, disgusted, fearful, happy, sad, surprised, and neutral. Among them, anger, disgust, fear, and sadness are negative learning behaviors, while happiness, surprise, and neutrality are positive and focused learning behaviors. After the recognition result of the facial expression recognition module is sent to the teaching evaluation module, the students' learning situation is obtained according to the classification of emotion.

There are 7 kinds of classroom learning states set in our study. The confidence level of each learning state can be obtained through facial expression recognition. The confidence level reflects the possibility of each state and can be used as a scoring standard. Take fear and sadness as the emotions that students are dissatisfied with the classroom and set their weights to -1; take disgust and anger as more dissatisfied emotions, and set their weights to -2; take neutrality as the students' satisfaction with the classroom, set its weight as 1; take happiness as the emotion of students who are more satisfied with the classroom, and set the weight as 2; take surprise as the emotion of students who are most satisfied with the classroom situation, and set its weight as 3. The weighted summation can obtain the emotional score of the sample as Formula (1). All scores are normalized to between [0,1] to obtain the final score. The calculation method of score normalization is as Formula (2).

$$score = p_0 + 2p_1 + 3p_2 - (p_3 + p_4 + 2p_5 + 3p_6) \tag{1}$$

$$norm(score) = \frac{score - \min(X)}{\max(X) - \min(X)}, score \in X \tag{2}$$

Among them, $P_i(i = 1, \ldots, 7)$ is the confidence level of the expression of i class.

The higher the score, the higher the students' concentration at this moment, and the score is relative to the whole class. When the score is higher than 0.5, it can be regarded as the classroom mood is higher than the average level of the whole class. Statistically average the scoring results obtained in a certain period of time to obtain the learning situation of the target students in a certain period of time. Finally, taking the average of the scores of all

students in the whole class can produce the overall student emotional score of the whole class so as to evaluate the quality of classroom teaching.

3. Results

3.1. Expression Recognition Model Training Results

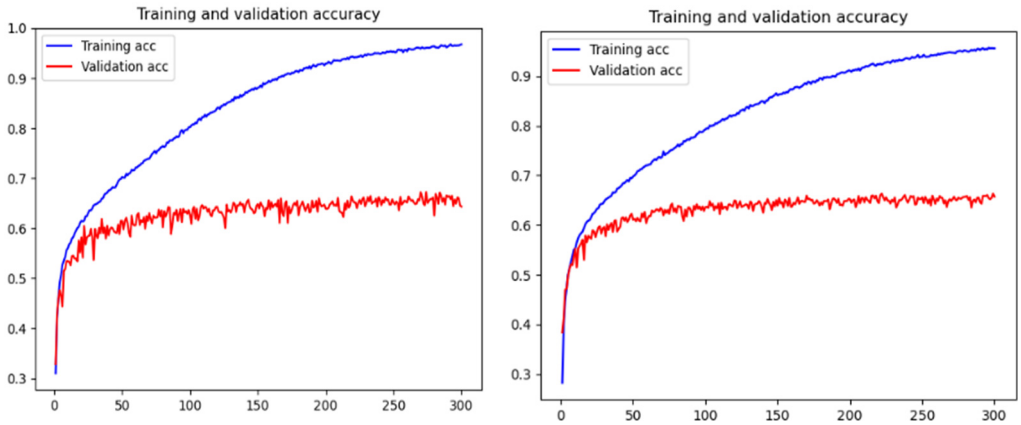
When training on the Fer2013 dataset, the network hyperparameters are as follows: The experiment is iterated 300 times. The batch size of the training set is set as 128, the batch size of the verification set is set as 128, and the initial learning rate is set to 0.0001. We optimize the weights of the network through the public validation set, and, finally, the model is tested on the private validation set to evaluate the model performance. The Stochastic Gradient Descent (SGD), shown in Formula (3), is used for optimization, and the categorical crossentropy loss function, shown in Formula (4), is used for model training.

$$g_t = \nabla_\theta f(\theta; x_i, y_i) + \nabla_\theta \phi(\theta) \tag{3}$$

where $f(\theta; x_i, y_i)$ represents the loss function in each sample, g_t is the gradient, and $\phi(\theta)$ is the regular term.

$$Loss = -\sum_{i=1}^{n} \hat{y_{i1}} \log y_{i1} + \hat{y_{i2}} \log y_{i2} + \ldots + \hat{y_{im}} \log y_{im} \tag{4}$$

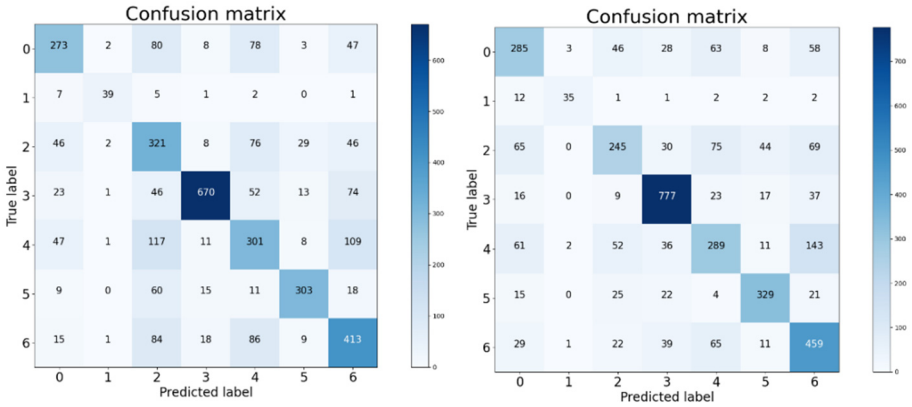
Figure 8 shows the accuracy curves of the two methods of VGG16 and VGG16 with ECANet. The accuracy of the two figures will gradually decrease with the increase of epoch times and, finally, become stable. When using VGG16 for model training, the accuracy of the training set will tend to be 0.94, and the accuracy of the validation set tends to be 0.64, while the accuracy of the training set of the VGG16 model with ECANet tends to be 0.95, and the accuracy of the validation set tends to be 0.66.



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

Figure 8. Acc curve. (**a**) VGG16 training acc curve; (**b**) Add ECANet's VGG16 model training acc curve.

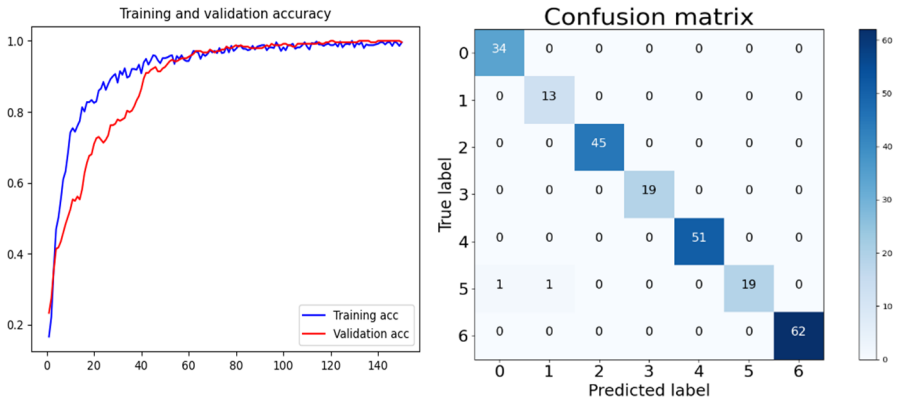
The confusion matrix obtained by testing this experimental model is shown in Figure 9. The abscissa is the predicted expression category, and the ordinate is the real expression category. The VGG16 model obtained an accuracy of 64.64% in the FER2013 test set, while the accuracy of the algorithm in this study was 67.40%, a relative improvement of about 2.76%. The accuracy of each category is: anger (58.04%), disgust (63.64%), fear (46.4%), happy (88.40%), sad (48.65%), surprised (79.09%), and neutral (73.32%). Among them, the recognition rate of happy and surprised expressions is relatively high, while the misjudgment rate between fear and sad, two expressions, is relatively high. Since the test set of this dataset has many label errors, the test accuracy is not very high.



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

Figure 9. Confusion Matrix. (**a**) VGG16 confusion matrix; (**b**) confusion matrix for the VGG16 model with ECANet added.

When using the CK+ data set for model verification, the CK+ data set is divided into a training set, validation set, and test set according to the ratio of 2:1:1. The results obtained by training 150 are shown in Figure 10. The batch size of the training set is 128, the batch size of the verification set is 128, and the initial learning rate is set to 0.0001.



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

Figure 10. Validation results of the CK+ dataset. (**a**) acc curve; (**b**) confusion matrix.

It can be seen from Figure 10 that the experimental model produced two false positives for the test set made with 245 pictures in the CK+ data set, and the recognition accuracy can reach 99.18%, which can achieve a high recognition effect.

The classification performance of this our study is measured through multiple measurements on the FER2013 test dataset, and we also compare the classification performance of the model used in this study with other models. For example, the Simple CNN, Simpler CNN, and Tiny XCEPTION models mentioned in [24], and the models used in [25–28], were compared, and the results are shown in Table 1 below.

3.2. Model Test Results

In the experiment, the online classroom was simulated for the system test, the weight files obtained by training the experimental model 300 times with the FER2013 data set were used to recognize students' faces and expressions, and the possibility of expression classification of each classmate was obtained, as shown in Table 2.

Table 1. Model comparison results.

Model	Accuracy
Simple CNN	62.91%
Simpler CNN	63.56%
Tiny XCEPTION	62.36%
[25]	66%
[26]	66%
[27]	67%
[28]	66.8%

Table 2. Expression recognition confidence.

Student	Anger	Disgust	Fear	Happy	Sad	Surprised	Neutral
Student0	$1.01 \times 10^{-0.2}$	$7.87 \times 10^{-0.5}$	$7.15 \times 10^{-0.6}$	$1.26 \times 10^{-0.6}$	$1.58 \times 10^{-0.3}$	$2.60 \times 10^{-0.6}$	$9.88 \times 10^{-0.1}$
Student1	$4.50 \times 10^{-0.5}$	$5.04 \times 10^{-0.5}$	$1.35 \times 10^{-0.4}$	$9.96 \times 10^{-0.1}$	$2.81 \times 10^{-0.6}$	$3.27 \times 10^{-0.4}$	$3.62 \times 10^{-0.3}$
Student2	$7.09 \times 10^{-0.5}$	$6.84 \times 10^{-0.6}$	$4.28 \times 10^{-0.6}$	$9.85 \times 10^{-0.1}$	$1.23 \times 10^{-0.7}$	$3.29 \times 10^{-0.5}$	$1.49 \times 10^{-0.2}$
Student3	$9.81 \times 10^{-0.5}$	$2.14 \times 10^{-0.6}$	$1.10 \times 10^{-0.6}$	$1.96 \times 10^{-0.5}$	$9.35 \times 10^{-0.7}$	$8.32 \times 10^{-0.4}$	$9.99 \times 10^{-0.1}$
Student4	$9.26 \times 10^{-0.1}$	$4.24 \times 10^{-0.4}$	$7.55 \times 10^{-0.3}$	$1.68 \times 10^{-0.4}$	$4.53 \times 10^{-0.2}$	$7.21 \times 10^{-0.3}$	$1.32 \times 10^{-0.2}$
Student5	$9.90 \times 10^{-0.2}$	$4.08 \times 10^{-0.3}$	$2.75 \times 10^{-0.2}$	$8.30 \times 10^{-0.3}$	$2.86 \times 10^{-0.4}$	$8.58 \times 10^{-0.1}$	$3.15 \times 10^{-0.3}$
Student6	$4.46 \times 10^{-0.8}$	$5.52 \times 10^{-0.7}$	$1.58 \times 10^{-0.7}$	$9.99 \times 10^{-0.1}$	$6.07 \times 10^{-0.9}$	$1.63 \times 10^{-0.7}$	$5.22 \times 10^{-0.5}$
Student7	$4.71 \times 10^{-0.3}$	$8.21 \times 10^{-0.4}$	$5.39 \times 10^{-0.1}$	$4.72 \times 10^{-0.7}$	$1.59 \times 10^{-0.3}$	$1.83 \times 10^{-0.1}$	$2.71 \times 10^{-0.1}$
Student8	$1.93 \times 10^{-0.2}$	$2.34 \times 10^{-0.3}$	$2.58 \times 10^{-0.1}$	$6.56 \times 10^{-0.6}$	$4.80 \times 10^{-0.3}$	$1.44 \times 10^{-0.1}$	$5.71 \times 10^{-0.1}$

Then, through the confidence in Table 2, the final expression recognition results of the students were obtained, as shown in Figure 11.

After the facial expression recognition was completed, the students' learning scores could be output. Table 3 was obtained by scoring the expressions in Figure 11 according to Formulas (1) and (2). It can be seen from Table 2 that the emotional scores of Student 1, Student 2, Student 5, and Student 6 at the time were above 0.7, which are relatively positive classroom emotions; the emotions scores of Student 0, Student 3, and Student 8 at the time were between 0.5 and 0.7, and they belong to the normal listening state, while the emotional scores of Student 4 and Student 7 at the time were below 0.5; the emotional score of Student 4 is especially too low, indicating that the listening state of these two students is not ideal; the teacher can ask them whether they understand the knowledge in time.

Based on the emotion score analysis of facial expression recognition, a relatively accurate score can be made for a specific student's emotional state in class at a certain moment, and the score reflects the enthusiasm and concentration of the student's listening state. Through the evaluation method in this study, the quality of the whole classroom teaching and the students' listening state can be evaluated, or it can be used as a reference index.

Table 3. Scores of students' learning status.

Student	Score
Student0	0.593
Student1	0.799
Student2	0.797
Student3	0.600
Student4	0.026
Student5	0.872
Student6	0.800
Student7	0.454
Student8	0.540
mean	0.609

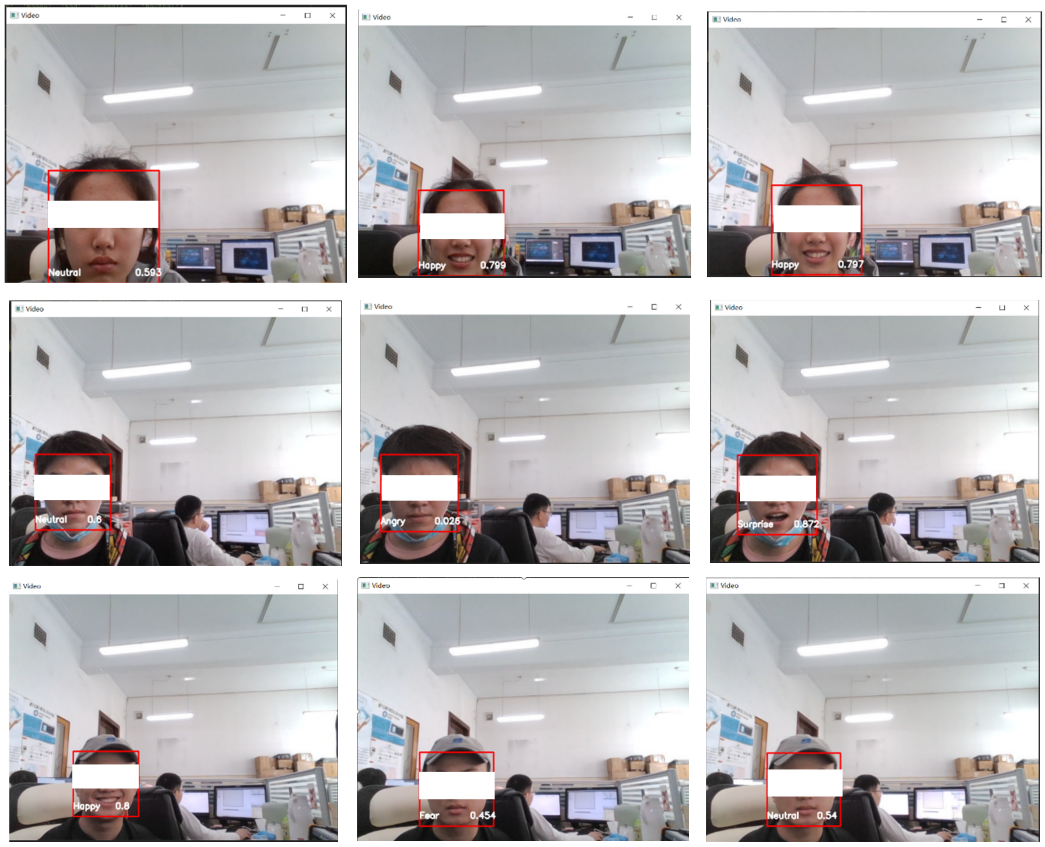


Figure 11. Expression recognition results.

4. Discussion

With the development of modern information technology, the concept of educational evaluation is also gradually changing. It advocates paying more attention to students' emotions and attaching importance to the means and effects of teaching quality evaluation. In such an era, this our study applies face detection and expression recognition technology to online teaching evaluation, and it obtains students' listening status so as to provide teachers with students' listening effects in a timely manner. The MTCNN face detection technology used in our study cascades three shallow networks together; the accuracy of each level is gradually improved, the coordinate information of the face frame is determined, and the face image is cut out from the picture. Using the VGG16 algorithm with ECANet for expression recognition, the accuracy of the test on the FER2013 data set can reach 67.4%, and the accuracy on the CK+ data set can reach 99.18%, which can more accurately identify the emotions of students. Using the model in our study to conduct classroom evaluations on nine students who took online classes, the emotional scores of the nine students were obtained as 0.593, 0.799, 0.797, 0.600, 0.026, 0.872, 0.800, 0.454, and 0.540. It can be concluded that the students with a score between 0.5 and 0.6 were in a normal state of listening, the students with a score above 0.7 were in an active learning state, and the two students with a score below 0.5 were in a negative learning situation. Of course, there are still some problems in the model of our study. Although the MTCNN algorithm we adopted has certain advantages for small target detection, the running speed of the algorithm is relatively slow. Moreover, our study lacks the corresponding student classroom face data set to train the expression recognition network, which leads to a decline in the accuracy of the model. In the following research, the corresponding data set can be produced and the MTCNN algorithm can be optimized to improve the running speed of the algorithm. Additionally, a new computer vision-based student attention evaluation criteria can be added, making classroom teaching evaluation more objective and robust.

5. Conclusions

Aiming at the problem that teachers cannot communicate with students face-to-face in real-time to understand the students' listening status in the online teaching process, this study uses the VGG16 network added with ECANet to perform real-time facial expression recognition on the students in online classes, and it obtains the students' listening status and gives timely feedback to the teacher. Using the VGG16 network with ECANet to predict the FER2013 dataset can obtain an accuracy of 67.4%, which is about 2.76% higher than when VGG16 itself predicts. However, due to the lack of conditions for making student expression datasets in the experiment, the use of open-source expression datasets to train the model leads to a decrease in the accuracy of the model. In subsequent research, a corresponding dataset can be created to improve the accuracy of the model. In addition, the students' listening state can be reflected by arm movements and head-up rate, which can be used as part of the students' attention evaluation criteria in the follow-up research, making classroom teaching evaluation more objective and comprehensive.

Author Contributions: Conceptualization, C.H.; methodology, C.H. and J.A.; investigation, C.G. and J.L.; data curation, W.Z.; writing—original draft preparation, C.H.; writing—review and editing, C.H., Y.L. and J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Heilongjiang Province Higher Education Teaching Reform Research Project (SJGY20200146);the National Natural Science Foundation of China (62001137); the Natural Science Foundation of Heilongjiang Province (JJ2019LH2398); the Fundamental Research Funds for the Central Universities (3072021CF0805).

Data Availability Statement: Expression dataset FER2013 https://pan.baidu.com/s/1i6p40jb accessed on 17 May 2022; expression dataset CK+ https://pan.baidu.com/s/1LFu52XTMBdsTSQjMIPYWnw accessed on 17 May 2022; the extraction password is 2pmd.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, Q.; Liu, X.; Gong, X.; Jing, S. INDReview on Facial Expression Analysis and Its Application in Education. In Proceedings of the 2019 Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 4526–4530. [CrossRef]
2. Krizhevsky, A.; Sutskever, I.; Hinton, G. Image Net Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Processing Syst.* **2012**, *25*. Available online: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (accessed on 17 May 2022).
3. Wu, T.; Bartlett, M.S.; Movellan, J.R. Facial expression recognition using gabor motion energy filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 42–47.
4. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
5. Zhao, T.; Wu, X. Pyramid Feature Attention Network for Saliency Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3080–3089. [CrossRef]
6. Varior, R.R.; Shuai, B.; Tighe, J.; Modolo, D. Scale-aware attention network for crowd counting. *arXiv* **2019**, arXiv:1901.06026.
7. Learned-Miller, E.; Huang, G.B.; Roychowdhury, A. Labeled Faces in the Wild: A Survey. In *Advances in Face Detection and Facial Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2016.
8. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. *Int. J. Comput. Vis.* **2014**, *1*, 3730–3738.
9. Taigman, Y.; Yang, M.; Ranzato, M.A.; Wolf, L. Deep Face: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
10. Zhou, E.; Cao, Z.; Yin, Q. Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not? *arXiv* **2015**, arXiv:1501.04690. Available online: https://arxiv.org/abs/1501.04690 (accessed on 17 May 2022).
11. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823. [CrossRef]
12. Rogier, A.M. Communication without words. *Tijdschr. Voor Ziekenverpl.* **1971**, *24*, 1084.
13. Happy, S.L.; Routray, A. Automatic facial expression recognition using features of salient facial patches. *IEEE Trans. Affect. Comput.* **2014**, *6*, 1–12. [CrossRef]

14. Majumder, A.; Behera, L.; Subramanian, V.K. Automatic facial expression recognition system using deep network-based data fusion. *IEEE Trans. Cybern.* **2016**, *48*, 103–114. [CrossRef] [PubMed]
15. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046. [CrossRef] [PubMed]
16. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Comput. Sci.* 2014, pp. 1–14. Available online: https://arxiv.org/abs/1409.1556 (accessed on 17 May 2022).
17. Huang, Y.; Dong, C.; Luo, X.; Dai, Q. Facial Expression Recognition Algorithm Based on Improved VGG16 Network. In Proceedings of the 2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT), Changsha, China, 11–13 June 2021; pp. 480–485. [CrossRef]
18. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
19. Xiang, J.; Zhu, G. Joint Face Detection and Facial Expression Recognition with MTCNN. In Proceedings of the 2017 4th International Conference on Information Science and Control Engineering (ICISCE), Changsha, China, 21–23 July 2017; pp. 424–427. [CrossRef]
20. Ghofrani, A.; Toroghi, R.M.; Ghanbari, S. Realtime Face-Detection and Emotion Recognition Using MTCNN and miniShuffleNet V2. In Proceedings of the 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI), Tehran, Iran, 28 February–1 March 2019; pp. 817–821. [CrossRef]
21. Gyawali, D.; Pokharel, P.; Chauhan, A.; Shakya, S.C. Age Range Estimation Using MTCNN and VGG-Face Model. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 1–3 July 2020.
22. Wang, Q.; Wu, B.; Zhu, P. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 85–102.
23. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
24. Raksarikorn, T.; Kangkachit, T. Facial Expression Classification Using Deep Extreme Inception Networks. In Proceedings of the 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), Nakhonpathom Thailand, 11–13 July 2018; pp. 1–5. [CrossRef]
25. Arriaga, O.; Ploger, P.G. *Real-Time Convolutional Neural Networks for Emotion and Gender Classification*; Hochschule Bonn-Rhcin-Sieg Department of Computer Science: Sankt Augustin, Germany, 2017.
26. Jaiswal, R. Facial Expression Classification Using Convolutional Neural Networking and Its Applications. In Proceedings of the 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), Rupnagar, India, 26–28 November 2020; pp. 437–442. [CrossRef]
27. Lalitha, S.K.; Aishwarya, J.; Shivakumar, N.; Srilekha, T.; Kartheek, G.C.R. A Deep Learning Model for Face Expression Detection. In Proceedings of the 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, 27–28 August 2021; pp. 647–650. [CrossRef]
28. Nie, H. Face Expression Classification Using Squeeze-Excitation Based VGG16 Network. In Proceedings of the 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 14–16 January 2022; pp. 482–485. [CrossRef]