Hindawi
Advances in Multimedia
Volume 2022, Article ID 8213637, 9 pages
https://doi.org/10.1155/2022/8213637

Research Article

Online English Teaching Quality Assessment Based on K-Means and Improved SSD Algorithm

Yuhua Dai

Foreign Language School, Huanghe Science and Technology College, Zhengzhou 450005, China

Correspondence should be addressed to Yuhua Dai; yhd369369@hhstu.edu.cn

Received 21 February 2022; Revised 26 March 2022; Accepted 13 April 2022; Published 10 May 2022

Academic Editor: Qiangyi Li

Classroom teaching quality is a key content to measure the teaching level, and the teaching effect can be intuitively reflected from the students' listening state. In order to improve the teaching quality, this paper proposes an online English teaching quality evaluation model based on K-means and an improved SSD algorithm. In the SSD algorithm, the backbone network is replaced by DenseNet with a dense connection to improve detection accuracy. The network structure of quadratic regression is designed to solve the problem of unbalance between positive and negative samples in the default box of the candidate region. A feature graph scaling method is used to fuse feature graphs without introducing additional parameters. The number of default boxes and the optimal aspect ratio were obtained by k-means clustering analysis. Finally, the state of students in the teaching process is predicted through the dual-mode recognition model of facial expression and posture, and the state of students in class is judged. Experimental comparison and analysis were conducted on the public data set and a self-built classroom teaching video data set. Experiments show that compared with other comparison algorithms, the algorithm presented in this paper performs better in the index of detection accuracy.

1. Introduction

With the development of information technology and the popularization of mobile Internet, the online education industry breaks through the limitation of time and space so that students can enjoy diversified educational resources. Online English education once became the mainstream mode of English teaching. The advantages of online English education are obvious.

In recent years, with the rapid development of Internet technology and the development of new technologies such as big data, cloud computing, and AI (Artificial Intelligence), intelligent education has become a new educational trend. As mentioned in literature [1], the word that has been neglected in intelligent education is emotion. However, as a non-intellectual factor, emotion can affect and regulate cognitive activities. Psychological studies show that positive emotions can stimulate learners' learning motivation, develop their interest in learning, and promote their cognitive process. On the other hand, negative emotions will affect the patience and attention of learners and hinder the cognitive process [2]. People's expression of emotion is complex and subtle. Similarly, people's recognition and interpretation of emotions are also completed through multichannel cooperation, including expression, posture, language, tone, and so on [3]. Currently, for emotion recognition, researchers mainly focus on physiological signals, psychological measurements, and explicit behaviours [4]. Among them, emotion recognition based on facial expression is the majority. Although facial expressions can express most people's emotions, there are also some inevitable problems. For example, facial occlusion, subtle expression, and posture change [5]. Therefore, the single modal emotion recognition method based on facial expression is not enough to accurately identify emotional states. In real life, people often judge a person's emotional state by integrating voice, facial expression, body movement, and other information. Using the complementarity of information, we can accurately identify the emotional state [6].

In the past decade, there have been more and more studies on learners' emotion recognition in class. In literature [7], a multimodal emotion feature fusion method based on genetic algorithm was proposed. Genetic algorithm is used to select, cross, and recombine the emotional features of a single mode. Literature [8] proposed a dual-mode emotion recognition system based on skin electrical signals and text information. In literature [9], a dual-mode emotion recognition method of expression and pose based on the bilateral sparse partial least square method was proposed. At the same time, foreign scholars have also carried out relevant studies. For example, literature [10] points out that emotion plays an important role in the human cognitive process. Therefore, a new emotion computing module is proposed. Biological, physical (heart rate, electrodermal, and blood volume pressure), and facial expression methods were used to extract learners' emotional states. By transforming the original physiological signal into the spectral image, the features are learned by using a bidirectional long- and short-term memory cyclic neural network (LSTM-RNNS). Finally, deep neural network (DNN) was used for prediction. It is found that the above methods are not completely suitable for Chinese classroom emotion recognition. The main reasons are as follows:

(1) Data set. At present, the data sets of emotion recognition research are collected by some foreign universities or research institutions. First, it does not conform to the classroom scene. Second, due to the influence of regional culture and skin colour, facial expressions collected in foreign data sets differ greatly from facial expressions in China.

(2) Emotional classification. In the field of emotion recognition, there are many categories of emotion. Among them, the most basic are the six basic emotions proposed by Ekman et al. They are happiness, anger, boredom, fear, sadness, and surprise. Studies have found that not all these six basic emotions play a key role in learning [11].

In the process of learning, emotion can affect students' cognitive behaviour. Therefore, grasping students' emotional state in the classroom is particularly important to improve classroom efficiency and promote the development of students' personalized education. Students' facial expressions are an important representation of class status. At present, there are many related studies that evaluate students' listening status according to their facial expressions. But there are still some potential problems. For example, the facial expression of a student who is unaware of a certain behaviour state. It is not rigorous and comprehensive to evaluate the students' listening state only based on their facial expressions. Similarly, there are certain limitations in evaluating the status of students' listening purely according to their classroom behaviours. The reason lies in that students' classroom behaviour can be passive under the guidance of the teacher's requirements, or it can be actively occurred by students according to their own status in the class. Therefore, without knowing the motivation of the behaviour, it is not comprehensive to evaluate the status of listening only by students' behaviour.

In order to solve the above problems, this paper proposes an online English teaching quality assessment model that integrates student expression recognition and behaviour recognition in classroom videos. VGG16 was replaced by DenseNet as the basic network, and feature extraction capability and computing speed were improved by optimizing the DenseNet structure [12].

This paper mainly has the following innovations:

(1) The target and background are simply distinguished

(2) Classified and location regression is carried out to obtain accurate default box information

(3) The feature information is extracted by designing a feature graph fusion module

(4) Feature image scale transformation is used to fuse feature images

(5) K-means clustering method is used to obtain the optimal length to width ratio of the initial default box

2. SSD Algorithm

2.1. Network Structure. SSD network is a typical network of target detection algorithms based on regression. It uses a single deep convolutional neural network combined with six feature maps of different scales to predict the classification and location information of the target. SSD network consists of a basic network and an additional network. Its structure is shown in Figure 1.

The vgg16 network is formed by removing the last two full connection (FC) layers and adding two convolution (conv) layers. By adding an additional network, the conv4-3 feature map is continuously downsampled to form a feature map with a gradually smaller scale.

2.2. Existing Problems. SSD algorithm has high accuracy and fast detection speed for conventional size targets. Although it can extract feature maps of different scales for target detection through a feature pyramid network, the effect of small-scale target detection is poor.

2.2.1. Insufficient Feature Extraction. For small-scale targets, the SSD algorithm mainly uses a Conv4_3 shallow feature map with high resolution. However, the Conv4_3 shallow feature map was located earlier in the model, and its feature extraction ability was insufficient, and the contextual semantic information was not rich enough. The semantic information of the deep feature map is abundant. But its scale will become small after multiple convolutions, pooling, and downsampling, and some location information and important detail features will be lost. And its default box size is large, which is not suitable for small target detection.

2.2.2. Imbalance of Positive and Negative Samples. The SSD algorithm generated 8,735 default boxes on the feature graphs of 6 different scales. The background accounts for a

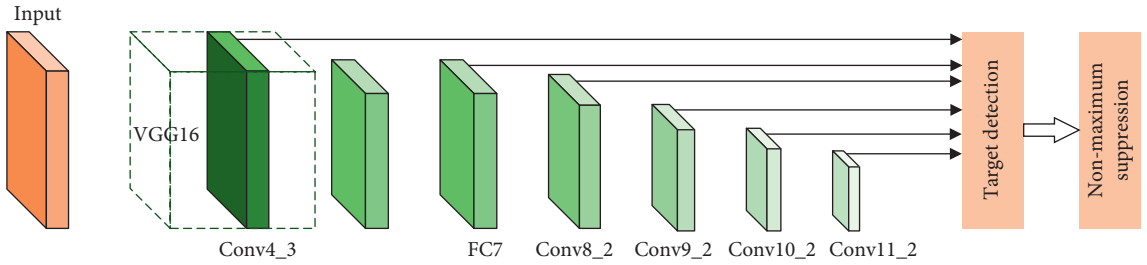

Figure 1: Structure of SSD network.

large proportion, while the detection target accounts for a small proportion. So most of the default boxes will be marked as negative samples after matching. Many negative samples losses account for the vast majority of the total loss of the model, weakening the impact of positive sample losses on the total loss, resulting in a serious decline in the training efficiency of the detection model. The model optimization direction will also be affected by different degrees of interference, resulting in the failure to update the model parameters to the optimal value.

3. Improved Algorithm Proposed in This Paper

3.1. Basic Network. The target detection algorithm usually selects the network that performs well in the classification task as the base network. The basic network is the network obtained by removing the whole connection layer from the classification network model. It is responsible for extracting image features and has a great influence on the performance of the target detection algorithm. SSD algorithm uses VggNet as the base network. YOLO algorithm uses GoogleNet as the basic network. The DSSD algorithm uses ResNet as the basic network [13].

The size of different classification networks is different, which will affect the speed of the target detection algorithm. Therefore, it is necessary to select an appropriate classification network to improve the detection accuracy and speed of the algorithm. The performance of the basic network can be improved by increasing the number of classified network layers or broadening the network structure. However, it will increase the number of network parameters and reduce the detection speed. From the perspective of features, DenseNet significantly reduces the number of network parameters and alleviates the phenomenon of gradient disappearance by reusing features and setting bypass, which can achieve a better detection effect [14].

DenseNet's first convolution kernel has a size of 7×7 and a step size of 2. After the first layer of convolution and pooled downsampling, the feature information of the input image has been partially lost before it has been fully extracted, which affects the subsequent feature extraction. Therefore, DenseNet is improved in this paper. Three consecutive 3×3 convolution kernels were used to replace the 7×7 convolution kernels in the original DenseNet. Three consecutive 3×3 convolution can reduce the number of network parameters more effectively than 7×7 convolution under the same scale receptive field. Moreover, the loss of input image feature information can be reduced, and the target details can be retained to the maximum extent, to extract feature information effectively. Parameters of the improved DenseNet network layer are shown in Table 1. After the feature images with the scale of 19 pixels × 19 pixels were obtained by improving DenseNet, feature images with the scale of 10 pixels × 10 pixels and 5 pixels × 5 pixels were obtained by downsampling for further detection.

3.2. Quadratic Regression. Target detection algorithms based on candidate regions such as FTP-RCNN need to preprocess candidate regions. Although this kind of algorithm has high detection accuracy. However, due to the existence of a full connection layer and many network parameters, the detection speed is slow, and real-time detection cannot be carried out. Target detection algorithms based on regression, such as SSD and YOLO, sacrifice detection accuracy to improve detection speed to a certain extent. At the same time, the network is trained based on the relationship between default box, prediction box, and object real box, and the default box is regressive.

The classification imbalance is the main reason that the accuracy of the target detection algorithm based on the candidate region is lower than that based on regression. In the first-stage end-to-end detection algorithm such as SSD, nearly 10,000 default boxes will be generated after the original image passes through the convolutional neural network. However, the proportion of target default boxes is very small, and the proportion of negative samples to positive samples is as high as 1,000:1, resulting in a serious imbalance of positive and negative samples. In order to solve the problem of category imbalance, literature [15] proposed online hard case mining. Bootstrapping technology is used to suppress simple negative samples to improve the training efficiency of the model. However, this method is only applicable to models with a small number of batches. Literature [16] redefined the cross-entropy loss function. By adding control weight to the standard cross-entropy loss function, the model pays more attention to the less difficult positive samples in training. But it does not solve the problem of category imbalance in essence.

This paper proposes our-SSD algorithm (hereinafter referred to as the algorithm in this paper). The network structure is shown in Figure 2. The hierarchical SSD target detection network structure is designed based on the regression idea of selecting the default box from coarse to fine

Table 1: Network layer parameters of improved DenseNet.

Network layer name	Output size per pixel	Image manipulation
Convolution 1	64 × 150 × 150	3 × 3 conv, stride 2
Convolution 2	64 × 150 × 150	3 × 3 conv, stride 1
Convolution 3	128 × 150 × 150	3 × 3 conv, stride 1
Pooling	128 × 75 × 75	2 × 2 max pool, stride 2
Dense block 1	416 × 75 × 75	(1 × 1 conv, 3 × 3 conv) × 6
Dense block 2	800 × 38 × 38	(1 × 1 conv, 3 × 3 conv) × 8
Dense block 3	1,184 × 19 × 19	(1 × 1 conv, 3 × 3 conv) × 8
Dense block 4	1,586 × 19 × 19	(1 × 1 conv, 3 × 3 conv) × 8
Transition layer 1	416 × 75 × 75	1 × 1 conv
	416 × 38 × 38	2 × 2 max pool, stride 2
Transition layer 2	800 × 38 × 38	1 × 1 conv
	800 × 19 × 19	2 × 2 max pool, stride 2
Transition layer 3	1,184 × 19 × 19	1 × 1 conv
Transition layer 4	1,586 × 19 × 19	1 × 1 conv

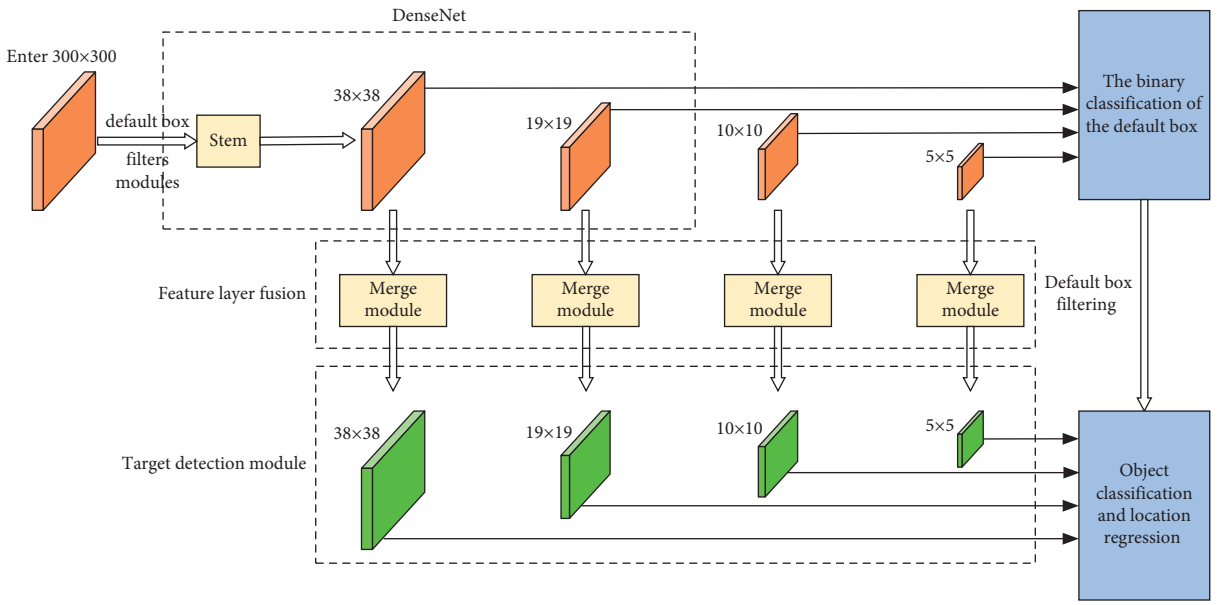

Figure 2: Network structure of the proposed algorithm.

in the two-stage non-end-to-end target detection algorithm. In the first part, SSD (ARM) performs simple binary classification and rough positioning of objects and backgrounds. In the second part, SSD (ODM) filters most of the simple negative samples according to the binary classification results of the first part and then conducts the judgment and position regression of the target category. The network structure of cascade multiple regression has higher detection accuracy. In order to increase the semantic features of the shallow feature graph and the detailed information of the deep feature graph, a feature fusion module is added between two cascade parts.

3.3. Scale Transformation of Feature Map. After many times of convolution, pooling, and downsampling, the original input image will get a feature map with a gradually decreasing size. In order to increase the semantic features of shallow feature maps and detailed information of deep feature maps, it is necessary to fuse feature maps of different scales. Before the feature graph fusion, the feature graph with a high score is generated from the bottom feature graph with a low score. The specific steps are as follows: (1) in the DSSD algorithm, the deconvolution method is used to fill the feature graph and its surroundings with 0 and then carry out convolution and pruning. After removing the last column on the right and the last row on the bottom, the high-resolution feature graph is obtained. (2) The bilinear interpolation upsampling method is used in the FSSD algorithm to interpolate the gaps without pixel values in the feature graph and enlarge the feature graph to a preset size. However, both the deconvolution method and the bilinear interpolation upsampling method will increase the number of network parameters, prolong the computing time, and reduce the real-time performance of this algorithm. In order to avoid reducing algorithm detection speed, a feature graph scaling method is proposed. Enlarge the size of the feature map without increasing the number of parameters. The scaling

process of feature maps is shown in Figure 3. First, the input feature map was divided into C feature maps with channel length r^2 in channel dimension, and then each feature map with channel number r^2 and size $B \times M$ was converted into a feature map with channel number 1 and size $r_B \times r_M$.

3.4. Cluster Analysis of Default Boxes. The detection accuracy and speed of the SSD algorithm are affected by the number of default boxes in the network. A small number of default boxes can improve detection speed but reduce detection accuracy. Many default boxes can improve the detection accuracy but reduce the detection speed. In addition, the SSD algorithm default box aspect ratio is manually set according to the experience of the detection personnel. Although its aspect ratio can be adjusted automatically during model training, if the initial number and aspect ratio of default boxes are more consistent with the characteristics of labelled objects in the data set, the model convergence can be accelerated, and the detection accuracy and speed can be improved.

In this paper, the optimal length-width ratio of the default frame is obtained through the k-means clustering calculation of the sizes of all marked target frames in the training set. In the K-means clustering algorithm, distance is selected as the evaluation index of target similarity. If the target distance is smaller, the similarity is larger. Close and independent cluster results can be obtained by k-means clustering calculation. The specific steps are as follows:

(1) Determine a k value as the number of sets obtained after clustering analysis of the algorithm

(2) Randomly select K data points in the training set as the initial centroid

(3) Calculate the distance between each point and k centroids in the training set and divide them into the set where the nearest centroid is located

(4) All the targets in the training set form K sets and recalculate the centroid of each set

(5) If the distance between the newly calculated centroid and the original centroid is less than the preset standard, the algorithm is completed

(6) If the distance between the newly calculated centroid and the original centroid is greater than the preset standard, repeat steps 3 to 5

In order to achieve a balance between detection accuracy and detection speed, the number of prior frames is selected as 5 ($k = 5$) to ensure that the detection speed of the algorithm is less affected. Then, the k-means algorithm is used for clustering analysis of all annotation boxes in the data set.

3.5. Two-Mode Recognition Model. In the intelligent teaching evaluation algorithm integrating classroom students' expression and behaviour, a represents the weight value assigned to expression recognition results, and β represents the weight value assigned to behaviour recognition results. The weight alpha and the weight beta add up to 1.

First, the count is used to represent the number of all kinds of expressions or behaviours. The following formula is used to calculate the total number of students recognized from the class video frequency:

$$\begin{aligned}\text{Sum} = {} & \text{count}_{pos\_\text{behaviour}} + \text{count}_{neg\_\text{behaviour}} \\ & + \text{count}_{neu\_\text{behaviour}}.\end{aligned} \tag{1}$$

Then, calculate the probability value of the product table and the product line as follows:

$$U_{pos\_\text{emotion/behaviour}} = \frac{\text{count}_{pos\_\text{emotion/behaviour}}}{\text{Sum}}. \tag{2}$$

And then weight is assigned to expression recognition results and behaviour recognition results, and the final comprehensive evaluation value is calculated as follows:

$$U_{com} = U_{pos\_\text{emotion}} * \alpha + U_{pos\_\text{behaviour}} * \beta. \tag{3}$$

Finally, according to the comprehensive evaluation value to analyze the students' overall class status evaluation grade. The class status assessment grade under different conditions is divided into 7 grades. They are A+ grade, A grade, B+ grade, B grade, B– grade, C grade, and so on. The meanings of each level are shown in Table 2.

4. Experiment and Analysis

4.1. Data Set. In order to further verify the accuracy of the proposed facial expression recognition method and the intelligent teaching evaluation method, this paper uses the public facial expression data set FERPlus and the self-built classroom teaching video data set for verification. FERPlus is an extension of FER2013 introduced in the ICML 2013 epitaph learning challenge. It is a massive data set of real facial expressions collected by Google's search engine. It includes 28,709 training images, 3,589 validation images, and 3,589 test images. All face images in the data set will be aligned and adjusted to 48×48. FERPlus is primarily labelled with 8 emoticon tags (neutral, happy, surprised, sad, angry, disgusted, fearful, and contemptuous). The authors evaluate several training schemes, such as single hot tag (majority voting) and tag distribution with cross-entropy loss. This paper mainly tests the overall accuracy by the majority voting method.

The self-built classroom teaching video data set is the video containing student expression and student behaviour trends. There are 90 video clips of classroom teaching. It is collected from real classroom scenes on the Internet. The data set is mainly used for the verification of intelligent teaching evaluation methods.

4.2. Experimental Settings. Adjust the size of all video frequency frames to 224×224. The actual environment is Pytorch 1.6. On all data sets, the learning rate is initialized to 0.01, divided by 10 after 30 epochs.

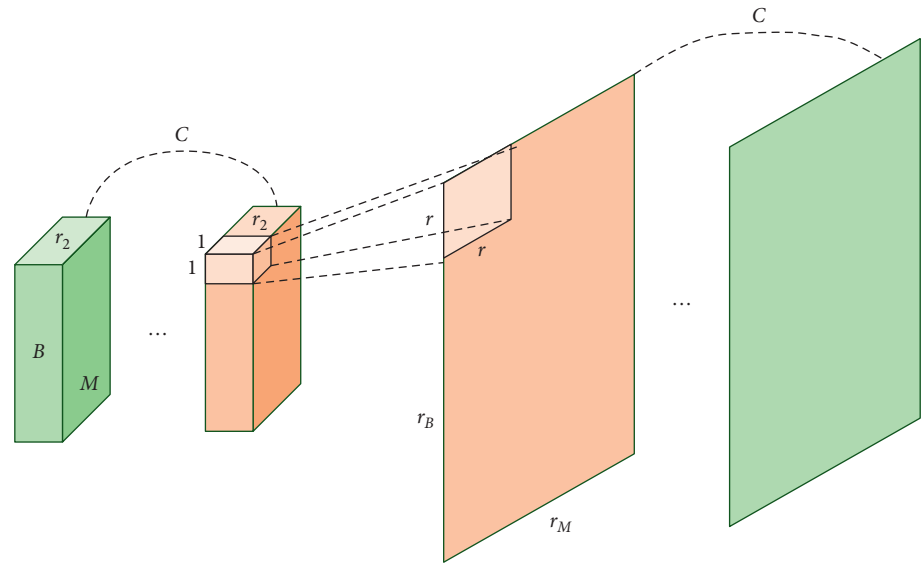

Figure 3: Process of scale transformation of the feature map.

Table 2: Grade evaluation form.

Grade	Grade description
A+	The proportion of students who listen carefully and actively participate in the classroom accounts for more than 75% of the total number of students in the classroom.
A	The overall classroom status is close to the A+ level, accounting for more than 70% of the total number, indicating that the whole classroom is good.
B+	The overall classroom state is close to grade A. Most (>67%) students are in the state of recognizing B+ listening to the class. Only a few students are not serious and can communicate with students after class.
B	A small number of students in the classroom do not listen carefully and cannot actively participate in class B. However, on the whole, the number of serious students (>65%) is still greater than the number of careless students.
B−	The number of students who do not listen carefully and actively participate in the classroom exceeds half of the total number of students in the whole classroom (>45%), indicating that the effect of the whole classroom is not ideal.
C	There are too many students (>C50%) who do not listen carefully and do not actively participate in the classroom.
C−	There are many problems in the classroom. We should find out the root problems in time and make corresponding adjustments.

4.3. Experimental Results and Analysis

4.3.1. Experiment 1: Experimental Results and Analysis of Different Algorithms on FERPlus Open Data Set. In this paper, the performance of the proposed model is compared with that of the comparative model on the open data set FERPlus. The comparison models are as follows: literature [17] is a model for extracting expression features from noise based on a deep convolution neural network. Literature [18] is a model that focuses on extracting facial regions that have an important influence on emotion recognition based on attention mechanism. Literature [19] is an expression recognition model based on shared representation integration. Their accuracy on the test set is shown in Table 3.

By observing the experimental results in Table 3, it can be found in this paper that compared with the other three models, the model proposed in this paper achieves the highest accuracy on the open data set FERPlus. This is mainly because the method in this paper integrates local and overall facial expression features and gives higher weight to the effective area of facial expression, which solves the problem of failing to learn effective facial expression features caused by the loss of facial expression information under occlusion.

Table 3: Grade evaluation form.

Models	Accuracy (%)
Literature [18]	66.91
Literature [19]	84.83
Literature [17]	86.33
Proposed model	**88.57**

Bold value represents the best result.

In order to further verify the validity of the proposed expression recognition model based on a deep attention network, the proposed model is compared with literature [18] model based on the attention mechanism. The experimental results are shown in Figure 4, showing the accuracy of different expression recognition models (the orange curve represents literature [18], and the green curve represents the proposed model).

According to Figure 4, compared with literature [18] model, the accuracy of the model proposed in this paper has reached 80.9% in the first round, which is nearly 30% higher than literature [18] model. This shows that the proposed model has achieved a good facial expression recognition effect in the first round. In addition, the accuracy curve of the

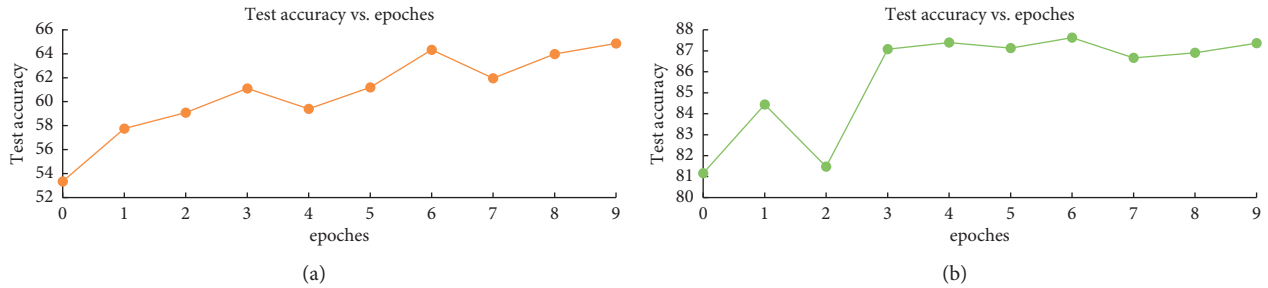

Figure 4: The accuracy of the test set: (a) literature [18] accuracy and (b) proposed model accuracy.

proposed model has gradually levelled off in the third round. While the literature [18], which is also based on the attention mechanism, only levelled off in the eighth round. Therefore, compared with the conventional attention-based facial expression recognition model, the proposed model can quickly learn effective facial expression features. In order to verify the weight value of the constrained loss function α for the influence on the accuracy of the model, this paper further sets different weight values α to verify the sensitivity of the proposed model to parameters. The accuracy results are shown in Figure 5.

In this paper, the weight values α are set as 0.9, 0.5, and 0.7, respectively. It can be observed from Figure 4 that the best effect can be achieved when the weight value α is 0.5. And it can achieve high accuracy in the first round of training. And, in the subsequent rounds, the performance is always stable, always at a high accuracy level. When the weight value α is 0.9, the accuracy is always low. Therefore, too low weight value α will make the threshold value setting unable to produce its due effect and unable to effectively optimize and increase the weight distribution of each branch. However, the setting of too high weight value α will make the weight difference of each branch too much and lose the significance of setting multichannel network. When the weight value α is too small, the threshold value setting cannot produce the desired effect. Therefore, the weight value α is set as 0.5 in this paper.

4.3.2. Experimental Results and Analysis of Different Algorithms on Classroom Teaching Video Data Sets. The following experiments will further test the performance and effectiveness of the proposed classroom teaching evaluation algorithm in real classroom scenarios. Aiming at the proposed intelligent teaching evaluation algorithm (WCA) integrating classroom expression and behaviour, the performance is evaluated on the collected classroom video data of 90 students, and the state of each student is predicted accurately. Judge the status of students in class by recognizing their expressions and behaviours. The video accuracy results of the test part are shown in Table 4.

The proposed algorithm has achieved good accuracy in each video. The highest is 74.7%. However, the addition of behaviour recognition algorithm will also produce certain errors, so the recognition accuracy of classroom teaching videos is very low. In addition, in order to better verify the

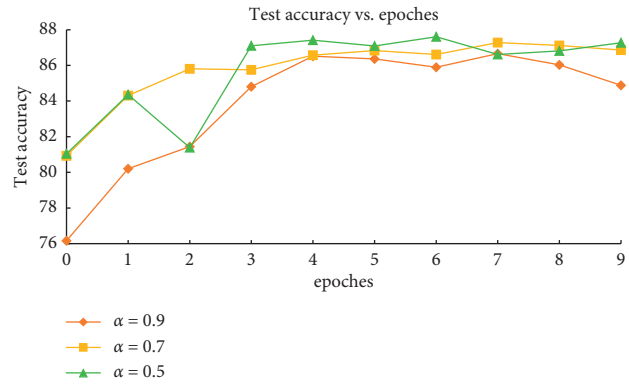

Figure 5: Parameter sensitivity experiment.

proposed intelligent teaching evaluation algorithm integrating classroom expression and behaviour, the proposed algorithm is compared with the classroom teaching evaluation algorithm based on expression, the classroom teaching evaluation algorithm based on behaviour, and the classroom teaching evaluation algorithm based on probability fusion. Some prediction and annotation results are shown in Tables 5 and 6.

According to the experimental results in Tables 5 and 6, it can be found that the average accuracy of the proposed intelligent teaching evaluation algorithm integrating classroom expression and behaviour is 57.9%. It is between the accuracy of expression recognition and behaviour recognition and higher than the accuracy of classroom assessment based on probability fusion. Because the fusion algorithm proposed in this paper distributes the weight of the results of expression and behavior, it can be used to classify the comprehensive values. The recognition results of facial expression and behaviour are more fully integrated. And, through many experiments, the selection of appropriate weight values will make the result more accurate. The classroom teaching evaluation method proposed in this paper combines both expression and behaviour, and the errors of expression and behaviour recognition will have a certain impact on the evaluation of the algorithm. Although the average accuracy of the algorithm proposed in this paper is lower than that of intelligent teaching evaluation based on behaviour recognition. However, the method based on behaviour recognition does not add the error of expression recognition, so the factors considered in this algorithm are

Table 4: Accuracy test table.

Video number	WCA accuracy (%)	Video number	WCA accuracy (%)	Video number	WCA accuracy (%)
1	57.3	9	47.7	17	61
2	55	10	54.9	18	50.6
3	51.1	11	59.4	19	62.3
4	48.3	12	74.7	20	60.9
5	52.8	13	42.9	21	70.4
6	65.5	14	42.3	22	49.3
7	57.7	15	44.8	23	52.8
8	70.3	16	76.2	24	64.6

Table 5: Comparison of classroom teaching evaluation results.

How to get the results of class assessment	Video number									
	1	3	8	9	15	19	22	23	24	37
The result of state scaling based on the expression	A+	B–	C–	A+	A+	A+	A+	A	A+	B
Classroom status assessment based on facial expression recognition	B	B	C–	A	A+	A+	A+	A+	A+	C
Behaviour-based significant status labelling results	B	A+	C–	C–	B–	A+	C–	C–	B	C–
Easy to conduct class status assessment	A+	A+	C–	C–	B	B	C	C–	C–	C
Annotation results based on the fact-fusion method of expression and behaviour	A	B+	C–	B	A	A+	B	C–	A	B
Probabilistic integration of classroom status assessment based on expression and behaviour	A	A	C–	B	A	A	B	B	B	B–
Annotation results based on expression and behaviour under WCA method	A+	A	C–	A	A+	A+	A	A	A	C
WCA classroom status assessment based on expression and behaviour	A	A+	C–	B	A+	A+	A	B	A+	C–

Table 6: Average accuracy of classroom teaching evaluation results.

Category	Accuracy (%)
Classroom teaching evaluation algorithm based on expression	54.5
Behaviour-based classroom teaching evaluation algorithm	61.2
Classroom teaching evaluation algorithm based on probability fusion	40.1
An intelligent teaching evaluation algorithm integrating classroom expression and behaviour is proposed	57.9

more comprehensive, and the accuracy result is more objective.

5. Conclusion

This paper takes student expression and behaviour recognition in classroom teaching videos as the research object and proposes an online English teaching quality assessment based on K-means and an improved SSD algorithm. VGG16 was replaced by DenseNet as the basic network, and feature extraction capability and computing speed were improved by optimizing the DenseNet structure. Based on the regression idea of selecting the default box from coarse to fine in the region-based candidate detection algorithm, the target and background are simply distinguished, and then classified and position regression are performed to obtain the accurate information about the default box. Feature information is extracted by designing a feature graph fusion module. Feature image scale transformation is used to fuse feature images. The number of default boxes and the optimal aspect ratio were obtained by k-means clustering analysis. Experimental results on classroom teaching video data set and public data set show that the proposed student expression recognition model and intelligent teaching evaluation method are superior to existing models in accuracy. The following research direction will identify the interaction between students and teachers and analyze its impact on teaching quality.

Data Availability

The labelled data sets used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported by the Huanghe Science and Technology College.

References

[1] T. Hoel and J. Mason, "Standards for smart education - towards a development framework," *Smart Learning Environments*, vol. 5, no. 1, pp. 1–25, 2018.

[2] A. Rowe and J. Fitness, "Understanding the role of negative emotions in adult learning and achievement: a social functional perspective," *Behavioral Sciences*, vol. 8, no. 2, p. 27, 2018.

[3] W. P. Bambang, "Integrating body language into classroom interaction: the key to achieving effective English language teaching," *Humanities & Social Sciences Reviews*, vol. 7, no. 3, pp. 121–129, 2019.

[4] J. Marín-Morales, C. Llinares, J. Guixeres, and M. Alcañiz, "Emotion Recognition in Immersive Virtual Reality: From Statistics to Affective Computing," *Sensors*, vol. 20, no. 18, p. 5163, 2020.

[5] Z. Zhang, X. Ji, X. Cui, and J. Ma, "A Survey on Occluded Face recognition," in *Proceedings of the 2020 the 9th International Conference on Networks, Communication and Computing*, pp. 40–49, Sydney, Australia, December 2020.

[6] W.-L. Zheng, W. Liu, Y. Lu, B. L. Lu, and A. Cichocki, "EmotionMeter: a multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2019.

[7] G. Lu, X. Cheng, X. Li, J. Yan, and H. Li, "Multi-modal emotion feature fusion method based on genetic algorithm," *Journal of Nanjing University of Posts and Telecommunications*, vol. 39, no. 5, pp. 41–47, 2019.

[8] Y. Lu, H. Zhang, L. Shi, F. Yang, and J. Li, "Expression-EEG bimodal fusion emotion recognition method based on deep learning," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 9940148, 10 pages, 2021.

[9] M. Aminu and N. A. Ahmad, "New variants of global-local partial least squares discriminant analysis for appearance-based face recognition," *IEEE Access*, vol. 8, pp. 166703–166720, 2020.

[10] P. Li, H. Liu, Y. Si et al., "EEG Based Emotion Recognition by Combining Functional Connectivity Network and Local Activations," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2869–2881, 2019.

[11] G. Li and Y. Wang, "Research on Leamer's Emotion Recognition for Intelligent Education system," in *Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 754–758, IEEE, Chongqing, China, October 2018.

[12] S. Zhai, D. Shang, S. Wang, and S. Dong, "DF-SSD: an improved SSD object detection algorithm based on DenseNet and feature fusion," *IEEE Access*, vol. 8, pp. 24344–24357, 2020.

[13] Q. Yin, W. Yang, M. Ran, and S. Wang, "FD-SSD: an improved SSD object detection algorithm based on feature fusion and dilated convolution," *Signal Processing: Image Communication*, vol. 98, p. 116402, 2021.

[14] S. Zhu, G. N. Mujiang, H. Jumahong, and P. L. N. Maiti, "A Remote Sensing Image Segmentation Method Based on Fusion Mechanism," *Journal of Physics: Conference Series*, vol. 2138, no. 1, p. 012016, 2021.

[15] S. Thakoor, C. Tallec, M. G. Azar, R. Munos, P. Veličković, and M. Valko, "Bootstrapped Representation Learning on graphs," 2021, https://arxiv.org/abs/2102.06514.

[16] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin Matters: Towards More Discriminative Deep Neural Network Embeddings for Speaker recognition," in *Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1652–1656, IEEE, Lanzhou, China, November 2019.

[17] S. M. S. Abdullah and A. M. Abdulazeez, "Facial expression recognition based on deep learning convolution neural network: a review," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 53–65, 2021.

[18] Y. Li, J. Zeng, and S. Shan, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.

[19] C. Luna-Jiménez, J. Cristóbal-Martín, R. Kleinlein et al., "Guided spatial transformers for facial expression recognition," *Applied Sciences*, vol. 11, no. 16, p. 7217, 2021.