A Teacher and Student Facial Expression Recognition Model Based on Classroom Teaching Videos

2023 International Conference on Intelligent Education and Intelligent Research (IEIR) | 979-8-3503-4289-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEIR59294.2023.10391209

1st Ziyi Liu
CCNU Wollongong Joint Institute
Central China Normal University
Wuhan, China
zeasonliu@mails.ccnu.edu.cn

2nd Li Yang
School of Computer Science
Hubei University of Education
Wuhan, China
hbyangli@hue.edu.cn

3rd Sannyuya Liu
Faculty of Artificial Intelligence in Education
Central China Normal University
Wuhan, China
lsy5918@mail.ccnu.edu.cn

***Abstract*— "Dear teacher make students believe his way". Related literature research shows that positive emotions of teachers' and students have a greater impact on the effectiveness of classroom teaching, and help to improve the learning efficiency of the students. After summarizing the domestic and foreign research status , this paper focuses on the realization of face recognition technology in AI evaluation of teacher teaching. The main research work is as follows. Firstly, this paper preprocesses the obtained classroom teaching video, including framing, removing invalid frames, etc.; followed, based on the Roboflow labeling tool, self-construct a teacher's and student's classroom teaching facial expression data set; then analyzes the characteristics of the general emotion data set; moreover, selects the general data set FER2013 as the training set; and the teacher's facial expressions are divided into three categories: positive, negative, and neutral, which is exported as a verification set subsequently; Finally, a teacher and student expression recognition algorithm is designed based on deep convolutional network. This article mainly uses the classic VGGNet as the main body of the network, and simultaneously completes feature extraction and expression classification. Simulation experiments show that the model has a high accuracy in teacher's and student's FER.**

***Keywords*—FER, Teaching emoticons, AI Evaluation of Classroom Teaching, DCNN**

I. INTRODUCTION

The Ministry of Education introduced the 'Education Informatization 2.0 Action Plan,' which aims to promote the comprehensive application of artificial intelligence in teaching, administration, and other aspects [1]. Furthermore, the '2022 Horizon Report (Teaching and Learning Edition)' also highlights that the application of artificial intelligence in learning analytics and instructional analysis is currently a burgeoning intersection of technology and education [2]. Our university's approval as one of the second batch of pilot institutions by the Ministry of Education for advancing artificial intelligence in teacher workforce development is precisely an advantageous exploration of artificial intelligence in teaching assessment and management. This initiative is aimed at further establishing a digitally-driven system for monitoring the quality of teacher education [3]. Consequently, AI-assisted instructional evaluation has emerged as a focal area in which artificial intelligence technology empowers educational assessment.

Educational assessment encompasses a wide range of elements, including teacher instructional behaviors, student learning behaviors, instructional content and methods, teaching environments, among various other factors [4]. One of its central components is the evaluation of teachers' classroom instruction. This evaluation includes assessing positive teaching attitudes and behaviors exhibited by teachers, such as facial expressions, body language, attire, praise, as well as evaluating student classroom engagement, such as attention and emotional changes.

Benefit from the widespread application of computer vision and natural language processing technologies, today, the evaluation of teachers' positive teaching behaviors, including facial expressions, can be accomplished through AI-assisted assessment methods. This involves AI recognizing teachers' instructional facial expressions, making it a significant aspect of AI-based instructional assessment. With the proliferation and promotion of smart classrooms, it has become favorable to conveniently collect and analyze classroom facial expressions during teaching. This makes it possible to identify and analyze expressions in teacher instructional videos using facial recognition technology.

Research by psychologist Albert Mehrabian found that the overall impact of information is composed of 7% written language, 38% vocal tone, and 55% facial expressions. Facial expressions are a primary form of emotional interaction between teachers and students. The appropriate use of facial expressions by teachers has a direct impact on enhancing student attention, improving learning attitudes and efficiency, and student positive expressions also indicate approval of the classroom experience.

In traditional instructional assessment, the evaluation of teachers' instructional practices primarily relies on expert observation and assessment during classroom sessions. However, this often results in inconsistent evaluation criteria and subjective judgments, leading to significant variations and limitations. Moreover, experts find it challenging to systematically record and analyze teachers' facial expressions throughout the entire teaching process. For teachers, understanding their own teaching expressions and paying attention to student classroom expressions can effectively enhance their teaching rapport. Consequently, this research holds the following practical significance:

A. It is a valuable exploration into the objectification of teacher instructional assessment data

Traditional forms of expert classroom observation and assessment face issues such as inconsistent evaluation criteria and subjectivity, leading to significant randomness and lack of support for enhancing educational quality. In classroom instructional assessment, this paper attempts to use facial recognition technology to quantitatively measure teachers' instructional expressions, particularly their rapport with

979-8-3503-4289-5/23/$31.00 ©2023 IEEE

students. This data-driven approach aims to provide empirical support for expert assessment, reduce the burden on experts, and enhance the objectivity of evaluation data.

B. *It provides effective assistance in understanding students' learning situations for teachers*

Student learning status, including their attention and knowledge mastery, will be reflected in their facial expressions. This paper utilizes data analysis of students' learning emotions to help teachers understand students' listening behaviors, knowledge acquisition, and other aspects of their learning situation in the classroom. These information can assist teachers adjust their teaching mode in a timely manner.

II. Related Work

A. *Expression Recognition*

Facial expression recognition, as a hot topic in the field of pattern recognition, has gone through multiple stages of development. In the late 20th century, various facial expression recognition methods were proposed. Deep learning -based facial expression recognition algorithms differ significantly from traditional methods. Traditional algorithms primarily rely on manually defined facial expression features combined with machine learning algorithms such as linear discriminant analysis or support vector machines. In recent years, deep learning facial expression recognition algorithms based on convolutional neural networks have replaced traditional methods, achieving higher accuracy and being applied in various fields, like payment systems.

The process of facial expression recognition generally involves three steps: image preprocessing, feature extraction, and result classification. Image preprocessing is a crucial step, typically involving normalization, median filtering, and image enhancement techniques. Various methods are commonly used for facial expression feature extraction:

- Template-based Matching Algorithms for Facial Expression

  In 1997, Gao Wen et al. established facial expression model templates and combined them with classification decision trees to analyze facial expressions. This algorithm calculates distances and proportions between facial feature points, resulting in fast recognition. However, it demands high facial pose requirements and involves a complex feature extraction process, leading to lower accuracy and poor generalization. Therefore, this algorithm has been phased out.

- Statistics-Based Facial Expression Recognition Algorithms

  In 2007, Zhang Yujin et al. achieved high accuracy in static image facial expression recognition by combining Gabor transform with histogram statistics. Statistical algorithms include Hidden Markov Models (HMM), Singular Value Decomposition (SVD), and Eigenface methods. These algorithms have high complexity and computational demands but exhibit good robustness.

- Deep Learning Recognition Algorithms Based on Convolutional Neural Networks (CNNs)

  In 2016, Lu et al. achieved facial expression recognition without explicit feature extraction using CNNs and improved algorithm robustness through pooling layers [5]. LeCun et al. achieved good facial expression recognition by weight updates through backpropagation.

In recent years, Guo Xingang et al. introduced attention mechanisms and residual network enhancement modules based on CNNs, improving recognition accuracy while reducing parameters [6]. Wang et al. proposed a facial expression recognition model with multi-granularity and self-correction fusion, which can avoid mislabeling of erroneous samples caused by CNN overfitting, thereby improving recognition accuracy [7]. It is evident that experts and scholars continue to propose various solutions to enhance recognition accuracy and robustness in facial expression recognition model design.

Common methods for classifying facial expression recognition results include Support Vector Machines (SVM), Bayesian methods, and k-Nearest Neighbors (k-NN). Generally, a seven-category approach is adopted, classifying recognized facial expressions into happiness, sadness, surprise, anger, disgust, fear, and contempt. Depending on the application scenario, facial expressions can be further categorized into positive, negative, and neutral emotional expressions.

B. *Expression Recognition in AI-based Classroom*

Teachers using positive facial expressions during instruction can enhance their rapport with students, thus increasing student motivation for active learning. As a primary means of conveying emotions in teaching, facial expression recognition during the teaching process has garnered significant attention from experts in the field of education.

Currently, the application of facial recognition technology in classrooms mainly revolves around attendance taking. Yang et al. implemented a video-based classroom attendance system using a multi-task cascaded neural network and camera control [8]. Fang Shuya et al. employed multi-camera technology to detect target positions in images captured by the primary camera using the Mask R-CNN algorithm. They then captured high-definition photos of each student using secondary cameras and used the FaceNet algorithm for facial comparison, achieving exceptionally high recognition accuracy [9].

Several researchers have experimentally discovered a direct relationship between students' classroom expressions and their level of attention. These scholars use facial expression recognition technology to identify students' head movement or micro-expressions during class to assess their attention and focus. Zhang, utilizing a self-attention mechanism, modified the bottleneck in a deep residual network for multi-scale feature extraction, allowing for the statistical analysis of the attention levels of all individuals in the classroom [10]. Wang et al. tackled the lack of relevant datasets in existing attention assessment by constructing a classroom attention expression dataset and designing a classroom focus evaluation model, addressing the limitations of single-metric models with low accuracy [11].

In addition to classroom facial recognition and analysis, many scholars have made various attempts in learner emotion recognition. Wang Tingting et al. utilized a facial detection

algorithm combining skin color segmentation and template matching. They defined three learning states (focused, fatigued, and neutral) based on the learners' fatigue status and provided interventions for learners experiencing fatigue [12]. Sun Bo et al. addressed the problem of minimal differentiation between facial expressions in the learning process. They designed an expression recognition framework for smart learning environments to eliminate interference from facial differences.

The challenge in AI-based classroom instructional assessment lies in the accuracy of recognition. Wang Liying et al. constructed a multimodal data fusion model for online learning behavior, synchronously fusing temporal data in three dimensions: behavior, emotion, and cognition. This approach allowed for hierarchical diagnostic assessment and statistical clustering analysis, achieving multi-dimensional and hierarchical learner emotion analysis [13]. Tang et al. proposed an intelligent classroom instructional assessment scheme based on facial expression recognition, emphasizing the analysis of students' classroom emotions to assist teachers in post-class teaching process analysis [14]. Yu et al. constructed a multi-path deep attention network and assigned different weights to multiple networks using a self-attention mechanism [15].

III. Proposed Method

This paper chooses to improve based on VGGNet. VGGNet has the advantages of small convolution kernel and simple structure, and also has deeper depth to achieve better model performance.

Through the improvement of the VGGNet model, the model parameters used in this paper are only one million levels, which is far smaller than the number of parameters of the original model which was over 100 million. Moreover, in the process of facial expression recognition, an improved model with fewer parameters is less likely to overfit, which can reduce a lot of calculations for model training.

A. Preprocessing of Teaching Videos

Classroom teaching videos are often lengthy, and due to the diversity of recording equipment and teaching environments, some videos may not be suitable for analysis and need to be excluded. When using classroom teaching videos for model training, a key preprocessing step involves selecting appropriate key frames and removing redundant video frames. Currently, there is no publicly available universal dataset for classroom teaching expression recognition, so it needs to be constructed through manual annotation.

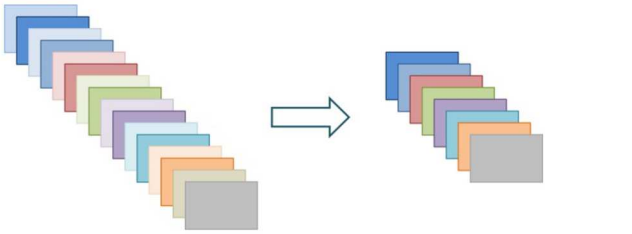

Fig. 1. Schematic diagram of teaching video preprocessing.

For the obtained classroom teaching videos, the first step is to segment the images into multiple frames, as shown in the figure, and then perform facial expression recognition. In this paper, a batch of classroom teaching videos collected from our university's AI teacher classroom teaching ability evaluation laboratory was used. Through manual judgment, some videos with students wearing masks or having excessively lowered heads were deemed unsuitable for training and needed to be excluded from the training set.

Fig. 2. Classroom video of too many students wearing masks.

To address issues such as high frame redundancy and unclear facial expressions in some frames, a frame comparison revealed that, compared to frame separation at a 1-second interval, which resulted in excessive redundant expression data and increased computational burden for model training, frame separation at a 5-second interval might lead to the loss of some student expression variation information. Therefore, this paper adopted a frame separation and sampling interval of 3 seconds and annotated the sampled data. Since the focus of this paper is on evaluating the facial expressions of teachers and students during classroom teaching, videos with more teacher appearances were primarily annotated. Furthermore, the captured images were compressed to reduce the complexity of model training.

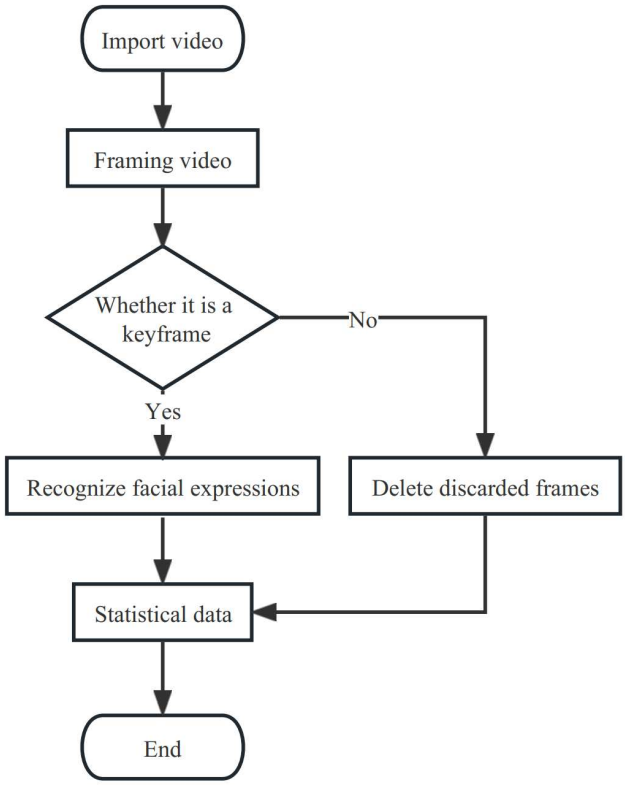

Fig. 3. Frame statistics flow chart.

B. Expression Recognition Datasets

Due to the limited availability of datasets that can be constructed solely by annotating classroom teaching videos, the process is slow. Moreover, the subjectivity introduced by a small number of annotators is a concern. Additionally, the small size of facial images in classroom videos poses challenges for training facial expression recognition models. Therefore, this paper introduces commonly used facial expression recognition datasets from the current research field into the model training process: FER2013, JAFFE, and CK+. Furthermore, the seven facial expressions are categorized into three classes that are more relevant to classroom scenarios: 'happy' as positive expressions, 'sad,' 'angry,' 'disgusted,' 'fearful,' and 'contempt' as negative expressions, and 'surprised' and other neutral expressions as neutral expressions. A comparison of these datasets is presented in the table below.

TABLE I. COMPARISON TABLE OF PUBLIC EXPRESSION DATASETS

Dataset name	Specific Information	
	Number of pictures	*Expression type*
FER2013	35886	7
JAFFE	213	7
CK+	981	7
KDEF	4900	7
RaFD	8040	8
SCface	4160	5

C. Deep CNN model and network

Currently, due to DCNN's advantages of effective feature extraction and high classification accuracy, utilizing Convolutional Neural Networks (CNNs) for facial expression recognition is the preferred choice among researchers.

However, using CNNs for facial expression recognition comes with its set of challenges. The recognition performance of the network is directly related to the training data, which means that training based on existing models can result in weak model generalization. On the other hand, increasing network depth or using larger convolutional kernels significantly increases computational complexity and can lead to overfitting.

Therefore, this paper chooses to improve upon the VGGNet for model training. VGGNet has the advantage of small convolutional kernels and a simple structure while achieving good model performance with sufficient depth.

In this paper, three convolutional modules are named as conv1, conv2, and conv3. Within the model, conv1-1 contains a 1x1x32 convolutional layer, conv2-1 contains a 3x3x64 convolutional layer, conv2-2 contains a 5x5x64 convolutional layer, conv3-1 contains a 3x3x64 convolutional layer, and conv3-2 contains a 5x5x64 convolutional layer. Additionally, there is a max-pooling layer with a feature map size of 24x24x64 between convolutional layers 2 and 3, and another max-pooling layer with a feature map size of 12x12x64 between convolutional layer 3 and the fully connected layer. Finally, a softmax layer is used to output the recognized facial expression type.

With the enrichment of datasets and improved computational capabilities, deep convolutional algorithms have become one of the commonly used algorithms in the field of computer vision. Compared to manually designed image recognition features, convolutional operations are better at discovering features within images, achieving precise recognition accuracy. The convolutional operation formula as in (1).

$$z(u,\ v) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} x_{i,j} \cdot k_{u-i,v-j} \tag{1}$$

In this equation, 'k' represents the convolutional kernel, and 'x' represents the input. If the input original information is three-dimensional data and the corresponding convolutional kernel is $K^{(l)}(n \times n)$, and the feature map is $X^{(l-1)}(m \times m)$, then the convolution operation with input bias is as in (2).

$$z(u,\ v) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} x_{i,j}^{l-1} \cdot k_{i,j}^{l} + b^{l} \tag{2}$$

Extensive experimentation by researchers has shown that the accuracy of deep convolutional algorithms is related to network depth; generally, deeper networks yield better recognition results. However, when constructing models, there is a need to strike a balance between model performance and training complexity. The calculation of the parameter count 'P' for convolutional operations is shown in (3). The calculation of computational complexity 'F' is shown in (4), where k_h represents the convolutional kernel's height, k_w represents the convolutional kernel's width, c_{in} is the number of input channels, c_{out} is the number of output channels, and 'H' and 'W' represent the length and width of the input image, respectively.

$$\mathrm{P} = k_h \times k_w \times c_{in} \times c_{out} \tag{3}$$

$$\mathrm{F} = k_h \times k_w \times c_{in} \times c_{out} \times H \times W \tag{4}$$

To reduce the computational complexity of deep convolutional neural networks, depth separable convolution methods have been introduced. Depth separable convolution helps reduce parameters, thereby lowering computational complexity. The parameter count for this convolution is shown in (5), and computational complexity is presented in (6).

$$\mathrm{P} = k_h \times k_w \times c_{in} \tag{5}$$

$$F = k_h \times k_w \times c_{in} \times H \times W \tag{6}$$

Compared to traditional convolution, the input and output channels of this convolution network are the same, with each channel having a corresponding convolutional kernel, thus reducing parameters and computational complexity. Subsequently, researchers further improved this convolution by designing the bottleneck module, as shown in the figure 4.

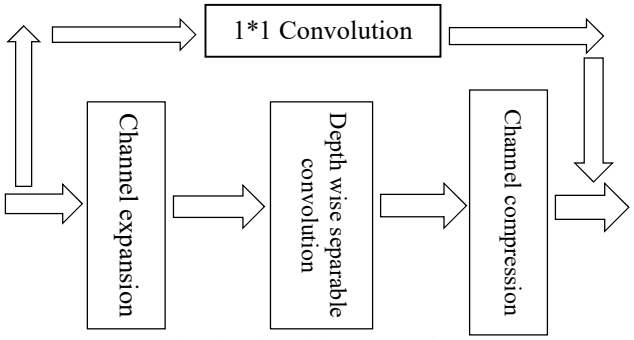

Fig. 4. bottleneck module structure diagram.

The bottleneck module differs from depth separable convolution in that it performs channel expansion before the

operation, resulting in more feature maps, which is beneficial for subsequent feature extraction. Channel expansion using a 1x1 convolutional kernel is an example of this. The calculation of computational complexity for the expansion operation is shown in (7).

$$P = 1 \times 1 \times c_{in} \times \alpha \times c_{in} \quad (7)$$

The parameter count for the expansion operation is presented in (8).

$$F = \quad c_{in} \times \alpha \times c_{in} \times H \times W \quad (8)$$

The calculation of computational complexity for the compression operation is shown in (9). The parameter count for the compression operation is presented in (10).

$$P = 1 \times 1 \times c_{in} \times c_{out} \quad (9)$$

$$F = \quad c_{in} \times c_{out} \times H \times W \quad (10)$$

The model flowchart for the deep convolutional algorithm from video input to statistical analysis in the recognition process is shown in Figure 5.

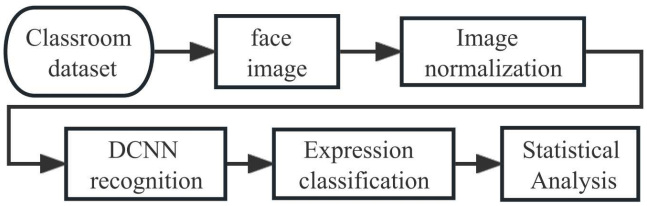

Fig. 5. Deep convolutional neural network algorithm flow chart.

During the recognition process, the calculation of facial position is as shown in (11).

$$\sigma = \varphi(MAX_x + \varepsilon_1,\ MIN_x + \varepsilon_2,\ MAX_y + \varepsilon_3,\ MIN_y + \varepsilon_4) \quad (11)$$

IV. Experimental Results and Analysis

A. Experimental setup environment

Due to necessity of training deep learning models in this experiment， PyTorch was selected as the primary deep learning framework. The experimental hardware setup involved utilizing a remote instance of the Zenith Cloud AI container server, with Ubuntu 18.04 as the operating system. The specific hardware configuration comprised a 16-core CPU, an NVIDIA RTX 3090 GPU with 32GB of system RAM and 24GB of GPU memory, and CUDA version 11. The main software and their respective version information used during the experiment are listed in the table below.

TABLE II. Experimental software and version information

Library	***Version***
Python	3.7
opencv	3.4.4.19
tensorflow-gpu	2.3.1
Numpy	1.24.2
Scipy	1.4.1
Matplotlib	3.7.1
Plotly	4.8.2
……	……

Due to the significant impact of batch size and epoch settings on model performance, this paper employs the stochastic gradient descent algorithm for model optimization. Batch size is closely associated with the model's generalization performance; increasing it can expedite the training process but may diminish generalization performance. Additionally, the configuration of these parameters must also take into account the hardware capabilities of the training device. Therefore, considering the aforementioned factors, the experimental parameters chosen for this study are presented in the table below.

TABLE III. Experimental hyperparameters

Experimental parameters	***Parameter value***
Learning_rate	1e-6
Batch_size	16
dropout	0.2
epoch	200
optimizer	Adam

B. Dataset selection and expansion

To objectively evaluate the performance of the trained models and facilitate direct comparisons with results from other researchers, this experiment chose publicly available datasets for facial expression recognition, include FER2013, JAFFE, and CK+. The use of publicly available datasets also expedited the training and testing process.

In the experiment, the dataset was divided into training, testing, and validation sets in a ratio of 7:2:1. To ensure consistency in the facial expression images across the three datasets, only those emotion categories present in all three datasets were selected for training. Additionally, a Java program was employed to standardize the annotation formats.

To enhance the training effectiveness of the models, image augmentation techniques such as rotation, flipping, and masking were applied to the dataset using the Roboflow platform to expand the dataset.

C. Analysis of experimental results

Figure 6 illustrates the accuracy variation curves of the deep convolutional neural network model used in this study across the FER2013, JAFFE, and CK+ datasets. Here, accuracy and loss were recorded every 5 epochs. The experiment tested the model's performance over epochs ranging from 0 to 200.

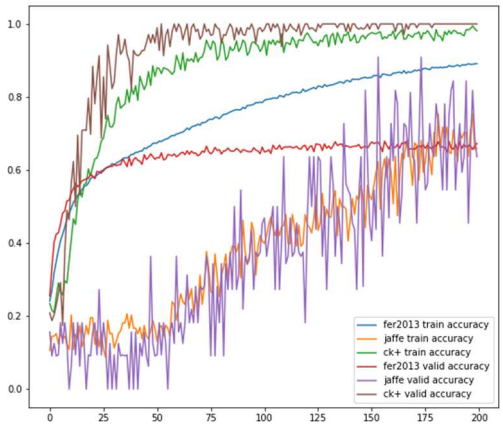

Fig. 6. Model Accuracy Variation Curve.

Figure 7 presents the loss variation curves of the deep convolutional neural network model used in this study across the FER2013, JAFFE, and CK+ datasets.

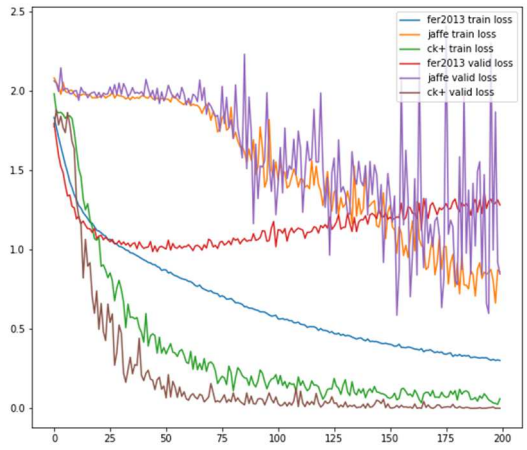

Fig. 7. Model Loss Variation Curve.

By observing the accuracy and loss curves, one can assess the model's training progress and estimate the fitting process. From the figures, it can be observed that the model tends to stabilize after approximately 75 epochs, indicating good fitting speed.

Due to the FER2013 dataset being collected through web scraping, the model achieved an accuracy rate of around 67%. In contrast, when applied to the laboratory-captured JAFFE and CK+ datasets, the model consistently achieved recognition accuracy rates of approximately 99% in cross-validation experiments.

V. Conclusion

In traditional teaching evaluations, the assessment of a teacher's instructional performance often relies on expert classroom observations. However, subjective judgment based on experience can lead to inconsistent evaluation standards, and experts may find it challenging to systematically record and analyze a teacher's expressions throughout the entire teaching process. For teachers, paying attention to their own teaching expressions and students' classroom emotions can effectively enhance their teaching rapport.

By introducing facial expression recognition technology into classroom teaching evaluations, it becomes possible to help experts comprehensively record emotional changes in each individual during a class, thus assisting teachers in improving their teaching.

This paper focuses on the application of facial expression recognition technology in AI-based classroom teaching assessments and accomplishes the following tasks:

- Using classroom videos, a dataset of facial expressions during teacher-student interactions is created. The obtained classroom teaching videos undergo preprocessing, including frame extraction and the removal of invalid frames. Subsequently, features of general facial expression datasets, such as FER2013, are analyzed and selected as the training dataset. Using the Roboflow annotation tool, a dataset of teacher and student facial expressions in the classroom is constructed, with expressions categorized as positive, negative, or neutral. This dataset is exported for validation purposes.
- A classroom expression recognition algorithm based on deep convolutional networks is designed. This paper primarily adopts the classic VGGNet as the core network and improves the model by enhancing the convolutional layers, following advancements from recent research papers. This approach simultaneously extracts features and classifies expressions. Simulation experiments demonstrate that this model achieves high accuracy in recognizing classroom expressions and exhibits fast fitting speed.

References

[1] Ministry of Education, "Notice on Issuing the 'Education Informatization 2.0 Action Plan'," http://www.moe.gov.cn/srcsite/A16/s3342/201804/t20180425_334188.html, Apr. 18, 2018.

[2] EDUCAUSE, "2022 EDUCAUSE Horizon Report: Teaching and Learning Edition," https://library.educause.edu/-/media/files/library/2022/4/2022hrteachinglearning.pdf?la=en&hash=6F6B51DFF485A06DF6BDA8F88A0894EF9938D50B, Apr. 18, 2022. [Accessed: Aug. 1, 2022].

[3] Ministry of Education, "Notice on Implementing the Second Batch of Pilot Projects for the Construction of Teacher Workforce Boosted by Artificial Intelligence," http://www.moe.gov.cn/srcsite/A10/s7034/202109/t20210915_563278.html, Sep. 8, 2021.

[4] Q. Li, "A Study on Teacher Classroom Teaching Behavior in the Informationized Teaching Environment," Ph.D. dissertation, Huazhong Normal University, 2012.

[5] G. Lu, J. He, J. Yan, et al., "A Convolutional Neural Network for Facial Expression Recognition," Journal of Nanjing University of Posts and Telecommunications (Natural Science), vol. 36, no. 01, pp. 16-22, 2016. DOI: 10.14132/j.cnki.1673-5439.2016.01.003.

[6] X. Guo, C. Cheng, and Z. Shen, "Facial Expression Recognition Based on Convolutional Network with Attention Mechanism," Jilin University Journal (Engineering and Technology Edition), [Online]. Available: https://doi.org/10.13229/j.cnki.Jdxbgxb20221345..

[7] J. Wang, M. Ma, A. A. Alimjan, et al., "Facial Expression Recognition Based on Fusion of Multi-Granularity and Self-Repair," Computer Engineering and Design, vol. 44, no. 02, pp. 473-479, 2023. DOI: 10.16208/j.issn1000-7024.2023.02.021.

[8] K. Fang, "Design and Implementation of Classroom Attendance System Based on Video Stream Face Recognition," Ph.D. dissertation, Huazhong Normal University, 2018.

[9] S. Fang and S. Liu, "A Non-Intrusive Classroom Attendance Method Based on Student Body Detection," Computer Applications, vol. 40, no. 09, pp. 2519-2524, 2020.

[10] W. Zhang, "Method and System for Recognizing Student Concentration Based on Self-Attention Mechanism," Ph.D. dissertation, Huazhong Normal University, 2022. DOI: 10.27159/d.cnki.ghzsu.2022.001703.

[11] N. Wang, "Research on Multi-Object Tracking in Online Video under Complex Scenes," Ph.D. dissertation, Beijing Jiaotong University, 2022. DOI: 10.26944/d.cnki.gbfju.2022.000059.

[12] T. Wang, Y. Wu, and X. Ai, "Recognition and Intervention of Learning Fatigue Based on Facial Expression Recognition," Computer Engineering and Design, vol. 31, no. 08, pp. 1764-1767+1778, 2010. DOI: 10.16208/j.issn1000-7024.2010.08.013.

[13] L. Wang, Y. He, and J. Tian, "Construction and Empirical Research of Multimodal Data Fusion Model for Online Learning Behavior," China Distance Education, 2020, no. 545(06), pp. 22-30+51+76. DOI: 10.13541/j.cnki.chinade.2020.06.002.

[14] Q. Tang, L. Zhang, Z. Xia, et al., "Research on the Application of Facial Expression Recognition in Classroom Teaching Assessment," Modern Information Technology, 2022, vol. 6, no. 20, pp. 191-195. DOI: 10.19850/j.cnki.2096-4706.2022.20.045.

[15] W. Yu, M. Liang, X. Wang, et al., "Recognition of Student Facial Expressions and Intelligent Teaching Assessment in Classroom Teaching Videos Based on Deep Attention Network," Computer Applications, 2022, vol. 42, no. 03, pp. 743-749.