Research on the recognition of students' classroom learning state based on facial expressions

Qiuchen Lin, Xiangwei Lai
College of Computer and Information Science, Southwest University,
Chongqing,400715, China
Email: lxw@swu.edu.cn

Abstract. The learning state of students in the classroom can not only affect the learning effect of students, but also reflect the teaching quality of teachers. In this paper, a new image database of classroom learning state is established according to students' learning state in real classroom environment, and an improved ResNeSt network model based on spatial attention mechanism is proposed. In this method, spatial attention mechanism is introduced into the ResNeSt network, which is complementary to the channel attention thought of the original network, so as to improve the recognition accuracy of learning state images. Using this method to identify students' classroom learning state can provide a basis for teaching administrators to evaluate teachers' teaching quality, as well as provide guidance for teachers to better adjust teaching models and teaching methods, and improve teaching efficiency.

Keywords: Learning State; Attention mechanism; ResNeSt network; Teaching quality evaluation

1 Introduction
Education is the basic of a One-Hundred Year Strategy of a Nation. Among all teaching methods, classroom teaching is currently one of the most common in our country. How to find teaching problems from the real classroom and improve the teaching quality has always been the focus of research in the field of education. Relevant researches show that the learning state of students in class is the most direct reflection of the quality of classroom teaching[1], and is an important basis for evaluating the teaching level of teachers.

However, in the traditional teaching, the teaching administrators can only observe the students' class situation and the teachers' teaching situation through the manual in-depth classroom[2], which not only consumes a lot of manpower and time, but also has great one-sidedness and limitations. With the advent of educational information age and the rapid development of computer vision field, it is an irresistible trend to apply computer vision technology to the classroom.

In this study, the deep learning algorithm is used to process and analyze the facial images of students in the real class, so as to obtain the real-time learning state of students in the class. By obtaining the learning state of students in this way, we can quickly understand the degree of interest and acceptance of students in the teaching methods and content of teachers. In this way, the one-sidedness and limitation of traditional classroom teaching administrators in evaluating teachers' teaching quality can be solved, and teachers can better adjust the teaching mode and teaching method to improve the teaching efficiency.
2 Related work

The earliest learning state comes from the study of children with learning disabilities, which is the discussion of the composition of students' learning disabilities. Regarding the concept of learning state, someone gave this view: The learning state refers to the physical and mental functional state of a person while studying, which mainly includes the awake state of the brain and the state of concentration, emotional state, physical function state[3]. Professor Picard from the Massachusetts Institute of Technology (MIT) Media Lab pointed out[4], Researchers can judge and recognize individuals' physiological and mental state by analyzing their physiological signals (heart rate, blood pressure, EEG, skin electricity, etc.), voice, facial expression, body posture and other information. Therefore, the current domestic and foreign research on students' learning state can be divided into the following three aspects.

(1) Study of learning state based on physiological signals

Relevant data show that the EEG activity has a certain regularity, and there is a certain degree of correspondence with the consciousness of the brain. Based on this, Srimaharaj et.al[5]. proposed a way to record students' brain signals using electroencephalography (EEG). Srimaharaj et al. proposed a method to classify students' learning state by cognitive performance recognition using electroencephalogram (EEG) recording students' brain signals. In this article, the decision tree algorithm is used to classify and describe the learning state of students, and the learning state of students is divided into two categories: attention and relaxation. Gunawan[6] et al. recorded learners' brainwave data through EEG sensors, directly measured their attention levels in the form of response time, and used the KNN algorithm to divide attention levels into two state categories, strong attention and weak attention. Although this research method can accurately detect the emotional state of students in theory, it is necessary to make each subject wear a corresponding biosensor in order to obtain the physiological signals of the students, which will cause certain interference to the class performance and behavior of students, affect the class of students, and affect the test results.

(2) Study of the learning state based on eye movement

"The eyes are the Windows of the soul." The behavior of eyes can directly reflect the current cognitive state of students. Yi Jiayue[7] designed a system that can perceive the current learning state of learners through eye movement behavior tracking. This system records the learner's Iris state, line of sight state and eye movement state, etc., and uses Hidden Markov Model (HMM) and Support Vector Machine (SVM) to divide the learner's learning state into three types: "scanning", "searching" and "idle". Qu Lianghao[8] judged the fatigue state of the students by detecting the situation of the students' eyes opening and closing as well as the situation of yawning. The AlexNet network model was used to divide the learning state of the subjects into two states: fatigue and non-fatigue. Cheng Mengmeng et al. divided learners' emotional state into curiosity, boredom, happiness, frustration, distraction, concentration, fatigue and confusion through eye tracking combined with facial expressions. This method based on eye movement is more suitable for online education because it needs to accurately locate the visual state of learners. Only when the learner's eyes are relatively close to the external camera can the student's eye movement data be accurately obtained.

(3) Study of learning state based on video image

The learning state of students in real classrooms is often more obvious through facial expressions or body postures. For example, when students are listening carefully to a class, they often show that their eyebrows are stretched, their eyes are staring at the teacher or blackboard, their hands are naturally placed on the table, and their upper body is slightly leaning forward. When students do not understand the teacher's class content, they tend to frown and close their lips. Sun Bo[9] et al. used the algorithm based on tensor decomposition to recognize the facial expressions of students' images, and divided the emotional states of students into 7 types: concentration, fatigue, happiness, surprise, confidence, confusion, and boredom. Liu Dongxing[10] et al. used the head posture estimation method (R-CR-C+EPnP) to calculate the position and rotation angle of the head to study whether students are paying attention to the blackboard area, and divide the students' classroom states into three categories: participation, attention, and neglect. Gao Yudou[11] uses the VGG19 model to recognize the expressions of students' images, and divides the students' learning states into seven kinds of learning emotions: happy,
surprised, confused, disgusted, fatigue, distracted and concentration. Yu[12] et al. used Microsoft Kinect equipment to record the video images of students in the classroom, detected typical classroom behaviors (such as sitting, raising hands, standing, sleeping and whispering) of students through face recognition and gesture recognition, and proposed a queue-based analysis engine divides these behaviors into positive and negative state. Fu[13] et al. used Open Pose to extract the key points of human bones, face and fingers, and then identified five classroom learning behaviors, including listening, fatigue, hand raising, side turning and reading and writing, through the classifier based on convolutional neural network (CNN). Han Li[14] et al. combined with existing intelligent monitoring devices to propose a method for studying the state of students' classroom learning based on facial expression analysis. Through real-time intelligent detection of students' facial features and efficient facial expression analysis, this method can deduce students' listening states (including understanding, resistance, listening, confusion and disdains).

The study method based on video image is currently the most popular method for studying the learning state of students in a real classroom environment. However, for the study method of learning state based on body posture, results obtained by analyzing 2D images taken by ordinary cameras are not ideal. Instead, deep-sensing cameras such as Kinect are needed to record students' 3D body postures. However, in the existing studies of learning state analysis based on facial images, most of them use the standard database captured in the laboratory environment to train and verify the model, so there are two shortcomings in the study of students' learning state in the classroom environment:

First, the facial expressions in the existing student learning image database are all based on the basic expressions (happy, sad, surprised, disgusted, fear, anger, etc[15]). However, in the classroom environment, students rarely show basic expressions such as fear, anger, and sadness due to teachers' teaching content and methods. Therefore, the classification of these basic expressions is not suitable for studying the learning state of students in the classroom.

Second, the images collected in the laboratory environment are all positive faces, and they have the characteristics of single background, uniform illumination and no occlusion. However, in the classroom environment, face images will have problems such as head deflection, face occlusion (hair occlusion, mask occlusion, eye occlusion, hand occlusion, etc.), poor sharpness and so on. Therefore, the database in the laboratory environment is not suitable for the study of real classroom scenes. As shown in the figure below, it is a comparison between the images collected in the laboratory environment and the images collected in the classroom environment.

3 Establishment of database

3.1 The learning state concerned in classroom learning

If we want to discuss the relationship between students' performance in class and their academic performance, we should pay the most attention to whether students pay attention or not in class, because only when students pay attention to the course in class, can they learn the most impressive and profound knowledge from the teacher's teaching. W. Srimaharaj et.al[5]. divided the learning state of students into
two types: attention and relaxation. F. E. Gunawan[6] et al. divided the learning state of students into two types: concentration and decreased attention. Yang[16] et al. studied the degree of students' concentration in class.

In this study, the main concern is the teaching quality of teachers and the degree of interest and acceptance of students in the course content. Through the study of pedagogy, psychology and other related disciplines, and the observation of a large number of students' class videos, and by referring to the relevant research of Sun Bo, Han Li et al., this study summarizes four learning states: "understanding, confused, boredom, and wandering".

3.2 Relationship between classroom learning state and facial features

After determining the types of learning state in this study, we summarized the corresponding relationship between the four learning states (understanding, confused, boredom, wandering) and facial features and the corresponding combination of AUs through interviews and discussions with professional teachers, combined with FACS(Facial Action Coding System). The facial features corresponding to each learning state are shown in Table 1, and the AUs corresponding to each learning state is shown in Table 2.

Table 1. Corresponding relationship between learning state and facial features

| Learning states | Eyes | Facial features | Mouth | Nose |
|-----------------|------|----------------|-------|------|
| understanding    | Upper eyelid lift, Fix eyes on, Blink at a uniform rate | Flat, Raised | Open or close naturally | natural |
| confused         | Wide open, Blink frequency becomes lower | Tightly | Lips closed, Corners of the mouth stretching | Nostrils expansion |
| boredom          | Blink frequency becomes higher, Upper eyelid contraction, Eyes wandering | Tightly, Eyebrows down | pout, yawning, Upper lip ascension, Nasolabial groove, One side of the mouth is upturned | Nostrils expansion |
| wandering        | Blink frequency becomes lower, Keep eyes fixed on one spot, Empty eyes or closed eyes | Flat, stretch | Yawning or Lips open naturally | natural |

Table 2. Corresponding relationship between learning state and AUs

| Learning states | AUs                             |
|-----------------|---------------------------------|
| understanding    | AU5+AU1+AU2+AU20/AU25           |
| confused         | AU7+AU4+AU14+AU20+AU38          |
| boredom          | AU7+AU4+AU17+AU10/AU12+AU9+AU38 |
| wandering        | AU43+AU27/AU25                  |

3.3 Establishment of the image database of classroom learning states

By collecting classroom video images of primary and middle school students and referring to the design
process of the learning emotion database of Beijing Normal University (Beijing Normal University Learning Affect Database, BNU LAD), we have established our student image database in the real classroom environment. We collected a total of students' classroom video images in one semester including the third grade, fifth grade, first grade and third grade of elementary school, and accumulated more than 500 hours of effective video materials. At the frequency of intercepting once every 4 minutes, the picture of the whole class is first intercepted, and then the upper body image of the target student is intercepted by drawing software. The principle of intercepting is the clearest and least interfered student image in the video. In the end, a total of 872 valid facial images were cut out.

Due to the small number of images available in the database at present, prior to the experiment, the database was expanded by data enhancement operations such as rotation, flip, cropping and scaling of the images, and the number of images was increased to 1600. With the original data source, we analyzed the facial features of the image and worked with professionals to manually label the learning states categories of all images to form a complete learning states image database.

The categories of learning states in the database include understanding, confused, boredom and wandering. Among them, there are 250 images in the understanding state, 235 in the confused state, 240 in the bored state and 212 in the disengaged state. In the experiment of this paper, two classification methods are discussed, the dichotomy (focused, unfocused) and the quaternary classification (understanding, confused, boredom, wandering). Among them, the images of focused in dichotomy are composed of the images of understanding and confused in quaternary classification, while the unfocused images are made up of bored and wandering images in the four categories. As shown in Figure 2, they are examples of understanding, confused, boredom, and wandering.

![Example diagram of learning states categories](image)

Figure 2. Example diagram of learning states categories

Since the image data set is collected based on the real classroom environment, we first need to detect and locate the face in the image, separate the object we care about from the image, and then go through gray normalization, size normalization, face occlusion filling and other pre-processing operations.

4. ResNeSt network based on spatial attention mechanism

The network structure adopted in this paper is ResNeSt network based on the spatial attention mechanism. That is, the spatial attention mechanism is introduced on the basis of the ResNeSt network structure, so that the network can pay more attention to the pixels with higher weights on the attention feature map of the spatial dimension, so as to achieve the purpose of adaptive feature optimization.

4.1 ResNeSt network

ResNeSt[17] is a new variant of ResNet[18]. This variant is inspired by other previous variant methods of ResNet, such as the multi-path multi-channel representation mechanism in GoogleNet[19], channel attention mechanism introduced in SE-Net[20] and feature graph attention introduced through two network branches in SK-Net[21]. ResNeSt network generalizes the attention of channel direction to feature mapping group representation. The feature map is divided into K groups along the channel direction, K is the number of feature map groups, and each feature map group is divided into R branches, each branch has its own weights, that is, the attention of feature map is divided into a single network block, so as to form a Split-Attention Block. Finally, each module is connected through the output after Split Attention. As shown in Figure 3 below, it shows the Split-Attention Block on the feature map.
groups in the ResNeSt network. Figure 4 shows the Split-Attention within a single feature map group.

**Figure 3.** Split-Attention Block on the feature map groups

**Figure 4.** Split-Attention within a single feature map group

### 4.2 Attention mechanism

Attention is a mechanism that reflects that people usually pay more attention to a certain part of information while ignoring other information. Wang et al. defined the attention mechanism in the field of computer vision, which usually includes two aspects: 1) Determine which area of the input needs to be focused on; 2) These important areas will be processed separately with limited computing resources. In this way, by introducing the attention mechanism, the network can be guided to assign different weights to different regions in the input image through learning, so that the key information in the image can be paid attention to and other non-key information can be suppressed.

There are two main types of visual attention mechanisms, soft attention and strong attention. Among them, soft attention is mainly focused on regional information and channel information, namely spatial attention and channel attention. The channel attention mechanism is to assign different weights to the signals on each channel through learning. The greater the weight, the greater the correlation between the channel and the key information, so it is necessary to pay more attention to these channels with significant weight. For example, the Squeeze-and-Excitation (SE) in SE-Net, in this module, the feature graph is obtained through the convolution operation. Then, the Squeeze operation is performed on the feature graph to obtain the global features on the channel dimensions, then performs the Excitation operation on these global features. Finally, it can obtain the connection between each channel through learning, thus get the respective weights of different channels. Spatial attention mechanism is to find out the regions with key information in the image through spatial transformation of the input image, and then give higher weight to these regions, so as to extract the local important information. Such as the spatial attention module in CBAM, it can obtain two feature graphs of global average pooling and global maximum pooling by calculation, and then connect them together. Finally, the feature graphs of spatial attention channels can be obtained by using sigmoid activation function.

### 4.3 An improved ResNeSt network model based on spatial attention mechanism

Inspired by the serial connection of the channel attention mechanism and spatial attention mechanism in CBAM, and since there was already the idea of channel-based attention mechanism in the original ResNeSt network, so this study introduced the spatial attention mechanism into the ResNeSt network. By calculating the attention map of the feature map of the spatial dimension, and integrating it with the original attention map of the channel dimension, so that the network has the ability of adaptive feature learning. Since spatial attention mainly focuses on "which position" is the important feature information, which is complementary to channel attention, so it can extract the key features of the student images in
this paper more effectively.

The network model adopted in this paper is shown in Figure 5. The original image is divided into K feature map groups, this study introduces the mechanism of spatial attention into the network of ResNeSt. We endorse network the capability of self-adaptation by calculating the attention map that is the characteristic graph of spatial dimension and then merge it with the original attention map of the channel dimension.

![Figure 5. Improved ResNeSt network based on attention mechanism](image)

5 Model training and result analysis

5.1 Experimental parameters

This experiment is based on Windows system and implemented in Python language. The deep learning frameworks used are Tensorflow and Keras. In the experiment, the Centerface algorithm was used to detect human faces, and preprocesses the detected faces through operations such as size normalization and gray-scale normalization, and then 10,000 iterations were carried out. Finally, the classification results of understanding, confused, boredom and wandering were obtained.

5.2 Experimental results and analysis

This experiment is based on the dataset constructed in this paper, and the results of the proposed algorithm and the existing algorithm are compared and analyzed. It includes the original ResNeSt, ResNet, ResNeXt, SeNet and the ResNeSt algorithm that introduces a spatial attention mechanism proposed in this paper, which is called Sd_ResNeSt. As shown in Table 3, it shows a comparison of results based on dichotomy. Similarly, Table 4 shows a comparison of the results based on the four classifications.

| Table 3. Comparison of classroom learning state results based on dichotomy |
|-------------------------------------------------|---------|---------|---------|---------|
|                                | Accuracy | Precision | Recall  | F1-score |
| ResNet                          | 0.797    | 0.796    | 0.793    | 0.795   |
| ResNeXt                         | 0.813    | 0.809    | 0.810    | 0.808   |
| ResNeSt                         | 0.820    | 0.815    | 0.816    | 0.815   |
| Sd_ResNeSt                      | 0.835    | 0.830    | 0.828    | 0.825   |

| Table 4. Comparison of classroom learning state results based on four categories |
|-------------------------------------------------|---------|---------|---------|---------|
|                                | Accuracy | Precision | Recall  | F1-score |
| ResNet                          | 0.698    | 0.697    | 0.695    | 0.696   |
| ResNeXt                         | 0.705    | 0.701    | 0.699    | 0.700   |
| ResNeSt                         | 0.712    | 0.710    | 0.711    | 0.710   |
| Sd_ResNeSt                      | 0.735    | 0.731    | 0.730    | 0.728   |
Through comparison, we find that ResNeSt has a maximum accuracy of 82% in the dichotomy and of 71.2% in four categories based on the dataset used in our article among the existing algorithms. By introducing the spatial attention mechanism into ResNeSt in this paper, the learning state recognition accuracy of dichotomy could reach 83.5%, which is 1.5% higher than the original ResNeSt, and the accuracy in four categories could reach 73.5%, which is 2.3% higher. The result demonstrates that the introduction of the attention mechanism can effectively improve the accuracy of learning state recognition in a real classroom environment. Nevertheless, the algorithm in this paper is not the fastest in terms of running time due to the addition of the spatial attention mechanism. Therefore, we are going to improve the detection speed of our model, to make it has a better balance between detection accuracy and detection speed.

6 Conclusion
This paper studies the students' classroom learning state in a real classroom environment. By adopting the self-designed image database of classroom learning states, we propose a new ResNeSt network model, which is a method of identification that is based on the spatial attention mechanism. The spatial attention mechanism has the feature information of images in the spatial dimension, which is complementary with the channel attention-based features in the original ResNeSt network. To do so, network can simultaneously focus on the information of both spatial dimension and channel dimension, which improves the recognition accuracy of image data in the real classroom environment.

References
[1] Zhijun Liu. Discussion on the Evaluation Standard of Classroom Teaching Quality[J]. Journal of Chinese Education, 2000, 2(2) pp 55-58.
[2] Yang Zhang. Research on the Problems and Countermeasures of Classroom Teaching Quality Evaluation[D]. Hunan Normal University, 2007.
[3] Jingjing Zhang. A Summary of the Study of Learning States[J]. Journal of Hubei Radio and Television University, 2010, 30(11) pp 15-16.
[4] Picard RW. Affective Computing[M]. MIT Press, 1997.
[5] Wanuss Srimaharaj, Supansa Chaising and Punnarumol Temdee. Brain Cognitive Performance Identification for student learning in classroom[J]. 2018 Global Wireless Summit, 2018 pp 102-106.
[6] Fergyanto E. Gunawana, Krisnantus Wanandib and Benfano Soewito. Detecting the early drop of attention using EEG signal[J]. 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2017 pp 1-6.
[7] Jiayue Yi. Research and application of online learning states real-time evaluation system based on eye movement detection[D]. Shanghai Jiao Tong University, 2016.
[8] Lianghao Qu. Analysis and Research of Students' Classroom Fatigue State Based on Deep Learning[D]. Chongqing Normal University, 2019.
[9] Bo Sun, Yongna Liu, Jiubing Chen and Di Zhang. Facial expression-based sentiment analysis in intelligent learning environments[J]. Modern Distance Education Research, 2015(2) pp 96-103.
[10] Dongxing Li, Heng Liu, Wenquan Chang, Pengfei Xu and Zuying Luo. Visualization Analysis of Learning Attention Based on Single-image PnP Head Pose Estimation[J]. International Conference on Education, 2017.
[11] Yudou Gao, Facial Expression Recognition and Emotion Analysis of Learners in Natural Scenes[D]. North China Electric Power University, 2019.
[12] Mengling Yu, Jing Xu, Jinrong Zhong, Wei Liu and Wenqin Cheng. Behavior Detection and Analysis for Learning process in classroom environment[J]. 2017 IEEE Frontiers in Education Conference (FIE), 2017 pp 1-4.
[13] Rong Fu, Tongtong Wu, Zuying Luo and Fuqing Duan. Learning Behavior Analysis in Classroom Based on Deep Learning[J]. 2019 Tenth International Conference on Intelligent
Control and Information Processing (ICICIP), 2019 pp 206-12.

[14] Li Han and Yang Li. Analysis of Teaching Effect Based on Facial Expression in Classroom Environment[J], Modern Distance Education Research, 2017, 000(4) pp 97-103.

[15] Dan Wang, Rong Fu, et.al. The expression of the emotions in man and animals[M]. London: John Murray, 1872.

[16] Bohong Yang and Zeping Yao. In-classroom learning Analytics Based on Student Behavior, topic and teaching characteristic mining[J]. Patter Recognition Letters, 2020, 129 pp 224-31.

[17] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu and Haibin Lin. ResNeSt Split-Attention Networks[J]. arXiv.org cs arXiv:2004.08955, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition[J]. IEEE Conference on Computer Vision & Pattern Recognition, 2016 pp 770-78.

[19] Christian Szegedy, Yangqing Jia and Pierre Sermanet. Going deeper with convolutions[J]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp 1-9.

[20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017 pp 7132-41.

Acknowledgments
This work was supported by program of “The Construction of Big Data Platform for Early Childhood Education and its Application in Promoting the Balanced Development of Regional Education", sponsored by the "Program of Technology Innovation and Application Development [The Special Program for Alleviating Poverty by Technology]" by Chongqing Committee of Science and Technology (Program No. CSTC2019JSCX-KJFP0004).