Journal of Artificial Intelligence Practice (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2022.050210
ISSN 2371-8412 Vol. 5 Num. 2

A Deep Reinforcement Learning Based Emotional State Analysis Method for Online Learning

Jin Lu[1,*]

Guangdong Key Laboratory of Big Data Intelligence for Vocational Education, Shenzhen Polytechnic, Shenzhen, Guangdong, 518055, China
Corresponding author: lujin@szpt.edu.cn

Keywords: Affective Computing, Deep Reinforcement Learning, Expression data, Learning Status

Abstract: With the development of artificial intelligence technology, the basic judgment of students' learning state can be realized through the comprehensive analysis of students' face, expression, behavior posture and other multi-modal data. However, due to the lack of end-to-end recognition model and complete data sets, it is impossible to achieve accurate analysis of learning status. In this paper, based on deep reinforcement learning, an online learning state analysis method based on affective computing is proposed. On the basis of student identity recognition, face recognition is carried out through an unsupervised expression recognition model based on Siam-RCNN, and then 3D CNNs is used to recognize the feature data set for timing extraction. The state of collaborative awareness learning is analyzed by using HMM model. After verification, the accuracy of emotional state recognition can reach 98.88%, which is in the leading level in the industry.

1. Introduction

Online teaching breaks the geographical restrictions, students can learn remotely through the Internet, and teachers can also use abundant information technology to assist teaching. In the process of online teaching, teachers can not grasp the learning status of students in real time, and in the process of teaching, students often do not concentrate, do not understand the teaching content, and are distracted in class, which leads to the overall inefficiency of online teaching. In order to solve this problem, it is necessary to diagnose the emotions of students in online learning intelligently, and feedback them to teachers for real-time adjustment by means of information technology, and using computer vision technology to calculate learners'emotions is a good solution. Deep reinforcement learning is a combination of deep learning and reinforcement learning. It uses the perception ability of deep learning to solve the modeling problem of policy and value function, and then uses the error back propagation algorithm to optimize the objective function. At the same time, it uses the decision-making ability of reinforcement learning to define the problem and optimize the objective. Deep reinforcement learning has general intelligence to solve complex problems to a certain extent, and has achieved success in some fields. At present, deep reinforcement learning can solve the problem of affective computing very well. Deep learning can recognize the expression of the object without making any decision, while reinforcement learning

can further provide decision information on the basis of recognizing the expression.

In the process of online learning, the multimodal collaborative data of learners is an important element to analyze their emotions, so the focus of online learning intelligent guidance application is to identify and carry out collaborative analysis through multimodal educational data such as micro-expressions and physiological characteristics. The most representative one is Koelstra S [1] of MIT Media Lab, which realizes the confirmation of some data related to emotion, memory and intention in the deep brain through the analysis of learners'facial expressions and action signs. The concept of affective computing was first proposed, and millions of expression databases have been collected all over the world. Referring to the application of K-SVD sparse dictionary in block classification and dictionary optimization, especially K-SVD can perform sparse denoising on images through a complete dictionary, Hao [2] proposed an approximate K-SVD algorithm for facial expression recognition, using a complete dictionary for feature extraction, and using an optimal linear classifier for final classification. Walied Merghani et al. [3] set normalization parameters for different data sets in order to study the influence of temporal and spatial changes of facial expressions on micro-emotion recognition, and conducted comparative experiments on expression recognition on the same platform for LBP-TOP, 3DHOG and HOOF. The global threshold algorithm is used to classify and compare the acceleration modulus and velocity modulus with the tensor modulus, so as to extract the key frames of the expression image sequence, and an expression recognition algorithm based on LBP + optical flow is proposed. Patel et al. [4] tried to use CNN model to replace the traditional LBP-TOP, Gabor filter, optical flow and other algorithms for expression feature extraction. Although the resolution of expression recognition is lower than that of traditional algorithms, it is a pioneer in depth feature extraction. Peng et al. [5] proposed a new apextime network (ATNet) based on the recognition method of spatial information of vertex frames and temporal information of adjacent frames. The improvement obtained by learning temporal information from adjacent frames proves the improvement obtained by adding temporal information learned from adjacent frames around vertex frames. Aiming at the problem that the accuracy of facial expression recognition mainly includes the high difficulty of facial expression recognition in small areas and the insufficient amount of data, Wang et al. [6] proposed to use the combination of residual network and micro-attention for micro-expression recognition, and the basic architecture is based on ResNet network. An attention unit is integrated in each residual unit to focus on the facial expression area, and in order to solve the overfitting problem, a transfer learning mode is used to train the model network; In the CNN + LSTM mode for micro-expression recognition, Khor et al. [7] used the CNN module to extract depth space features, encoded each expression frame into a feature vector, and then realized the feature vector through the LSTM module in the time domain. At the same time, they innovatively proposed two implementation architectures. It respectively tests the recognition effect by enriching the spatial dimension from the superposition of input channels and enriching the time dimension through the superposition of depth features, and uses the advantages of deep learning algorithm in spatio-temporal data processing. Xia et al. [8] proposed a STRCN model, which uses two connection modes to realize the synchronous processing of spatial information and time information respectively. In addition, they also designed a data enhancement strategy to enhance samples. Expression detection technology can also be applied to expression recognition. Michiel Verburg et al. [9] extracted the directional optical flow histogram (HOOF) features to encode the temporal changes of the selected facial region, and then passed them to the RNN composed of long short-term memory (LSTM) units for detection tasks. Huai-Qian Khor et al. [10] proposed a rich long-term recursive convolutional network (ELRCN), which encodes each expression frame into a feature vector through a CNN module, and then classifies the feature vector through a long-short-term memory (LSTM) module.

From the above literature research, it can be seen that the research in the field of learning state

sentiment analysis using expression recognition methods has been very rich, in which the traditional methods use more mature computer vision technology, including LBP, optical flow, sparse dictionary and other feature extraction methods, and have achieved good results. However, the traditional methods are faced with the problems of large amount of calculation and high dependence on data sets, and the research progress has encountered bottlenecks. As the mainstream research direction of computer vision technology, deep learning method is also used in expression recognition. From the research results, deep learning algorithm has gradually surpassed the traditional machine learning algorithm and obtained higher recognition rate. However, due to the insufficient sample data set, most of the existing methods use mature models. The optimal model parameters are not obtained through sample training, resulting in limited improvement of recognition rate. Therefore, it is a new research direction to combine the existing two methods for optimization and improvement, and use the hybrid method for facial expression recognition, but the current hybrid method is simply to "splice" and "combine" the traditional algorithm and the deep learning algorithm, and does not propose an end-to-end system recognition method.

In order to solve the above problems, this paper proposes an online learning state analysis method based on affective computing based on deep reinforcement learning. Firstly, through multi-modal in-depth learning technology, expressionistic information is intelligently recognized from high-definition classroom video and other materials, the basic data set is constructed, and accurate recognition is carried out after time sequence processing, finally, the real learning state is obtained by data modeling through Hidden Markov Model (HMM), and the Viterbi algorithm is used to solve the model. The optimization strategy of learning state based on cooperative awareness is realized.

2. Proposed method

Inspired by the application of twin network in the field of image detection [11], we propose an unsupervised expression recognition model based on Siam-RCNN, that is, face recognition from the Template Frame through the twin network deep convolution model. At the same time, according to the face recognition information, the expression is recognized from the feature frame (Detection Fram). Here, in the face of multi-modal data model, the traditional supervised learning model will have a very large amount of video frame annotation work, and because it will introduce input risks such as manual annotation errors, we propose to use unsupervised mode for model training [12]. The unsupervised target recognition which only needs to carry out first frame labeling is adopted, wherein the'unsupervised 'is embodied in that the training is carried out by adopting video data which only needs to label the first frame of the video, the training process is divided into two parts, namely forward tracking and backward tracking, and in the forward tracking process, a GT template frame is used for predicting a future frame; and in the backward tracking process, a GT template frame is used for predicting a future frame. The GT template frame is predicted using the inferred "pseudo-label" future frame as the template frame.

Siam-RCNN networks need to add timing information. Reference [13] proposes a method for 3D CNNs to identify feature data sets for timing extraction. A 3D CNNs convolutional neural network is used to encode video data temporally, and a trainable filter and local neighborhood pooling operation are applied to the original input to obtain a hierarchical and gradually complex feature representation. The temporal and spatial characteristics of the human body in the video data output by the Simac module are extracted through the SSD core. These SSD kernels operate in the spatio-temporal dimension, so they can capture the motion information of the human body in the video stream. A 3DCNN [14] is constructed based on the original frame and crop image respectively, and then the two 3DCNNs are used as feature extractors to extract the features of the original frame

and crop frame respectively. This architecture can generate multiple channels of information from successive video frames, and then perform convolution and downsampling operations separately in each channel. Finally, the information of all channels is combined to get the final feature description. The model [15] is enhanced by computing the auxiliary output from the high-level motion features, and the features are data fused and timing information is added. The expression features extracted by the double 3D feature extractor are processed by the SVM classifier, and the feature data are retained while the noise data are removed. The temporal coding method for 3D CNNs to identify feature data sets will be based on the KTH data set for training [15] [16]. Four consecutive 16-frame videos are randomly taken from each training set as training samples, and one consecutive 16-frame video is randomly taken from each test set as test samples.

According to the characteristics of the traditional MDPs model, the expression can be taken as an action, and the learning state can be taken as a state, and the learning state strategy of a student agent can be perfectly formed into a reward function for feedback, but for time series data such as learning state data, MDPs is difficult to realize data modeling. As a derivative process of MDP [17], HMM's ability to process time series data is very suitable for the real learning state modeling of this project. The HMM model is $\lambda=(A, B, \pi)$ It belongs to the two-state sequence model, in which the observable sequence represents the emotional state and the implicit sequence represents the real learning state of students. Through the above analysis, the hidden Markov model (HMM) can be used to model the relationship between students' learning state and expression emotion. The specific implementation steps are shown in Figure 1:

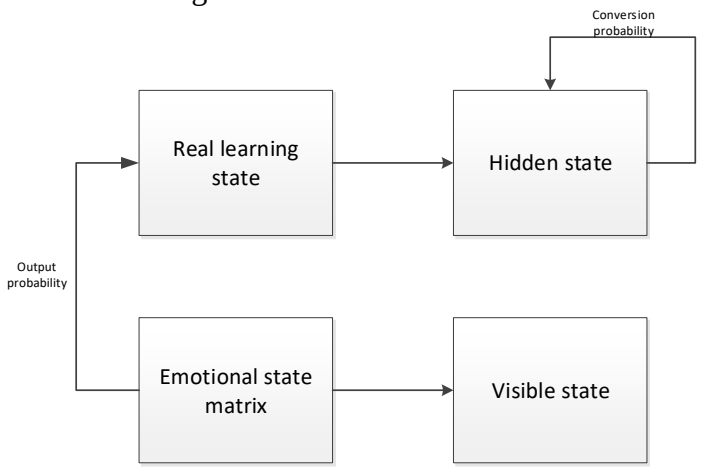

Figure 1: HMM calculation model

The original data state, namely the visible initial state, is constructed through the facial expression. π 。 The initial learning state transition probability matrix A is obtained by performing motion estimation on the mapping matrix through the logarithmic likelihood value and the sliding window method. Observed transition probability B between real learning state and facial expression is obtained by online planning. Via the HMM model. $\lambda=(A, B, \pi)$ Realize the modeling of collaborative perception learning state, and finally get the real learning state of the learning guidance system.

For a given model $\lambda=(A, B, \pi)$ And observable sequence O, solving for the maximum probability yields O The sequence of States for the S, that is, to solve $\arg\max_{S_1,\ldots,S_t}(O_1,\ldots,O_t,S_1,\ldots,S_t)$。 Approximation algorithm and Viterbi algorithm are generally used to solve this kind of decoding problem. The approximation algorithm is simple, but it does not consider the timing relationship, so

it can not ensure that the state sequence is the maximum probability state sequence, and the state sequence may have some parts that do not actually occur. The Viterbi algorithm uses dynamic programming to solve for the path with the largest probability, starting from$t = 1$Begin at time t, that state at time t is calculated recursively asS_tThe maximum probability of each partial path of, until it is calculated to time T, the state isS_TThe maximum probability of each path at time T is the probability of the optimal path, and the node of the optimal path is also obtained. The Viterbi algorithm effectively reduces the time complexity through optimization. If the length of the observable sequence is m, the number of hidden States is n, and the hidden state transition diagram is Fig. 2.

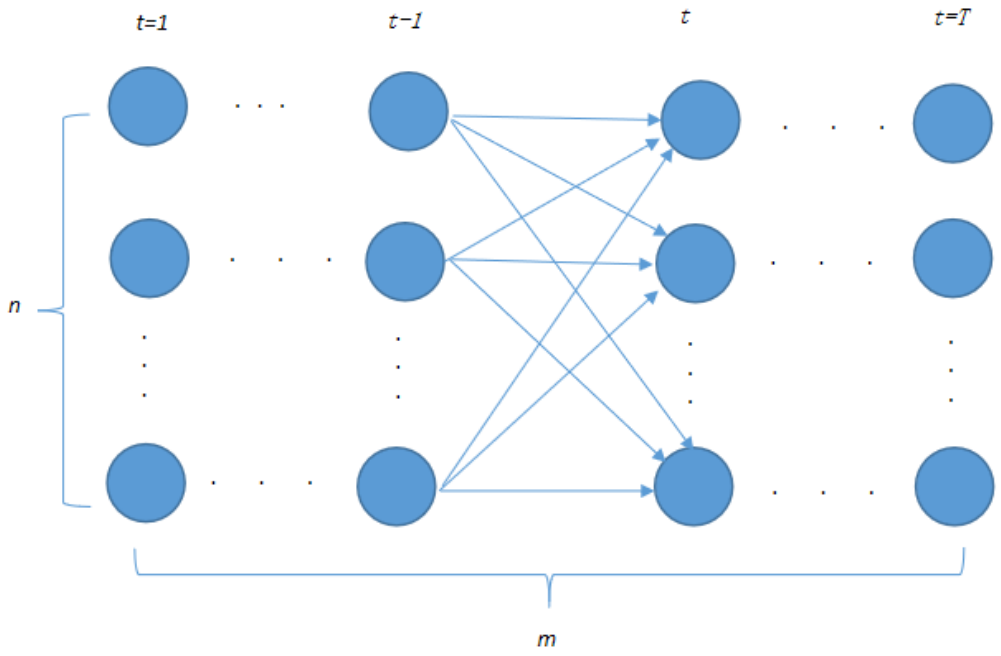

Figure 2: Viterbi algorithm

It can be seen that the hidden state probability at time t is only calculated at time T-1, so the state reuse theory in online planning can be applied to the Viterbi algorithm to reduce the time and space complexity.

3. Experimental process

All the experiments are carried out on the workstation, the equipment parameters are intel I9-10900k, 64GB, 2 * RTX3090, 1 T pcie3.0 SSD, and the experimental platform is TensorFlow. During the experiment, the CIFAR-10 data set was used to train the constructed convolutional neural network.

In this paper, a convolutional neural network is constructed and trained. The parameters of the trainNetwork function are: training data set, training set label, network structure, training strategy, and the classification effect of the convolutional neural network is verified.

First, as shown in Fig. 3, a face detection algorithm is used to detect the position in the face, and the face position is obtained by cropping from the original picture.

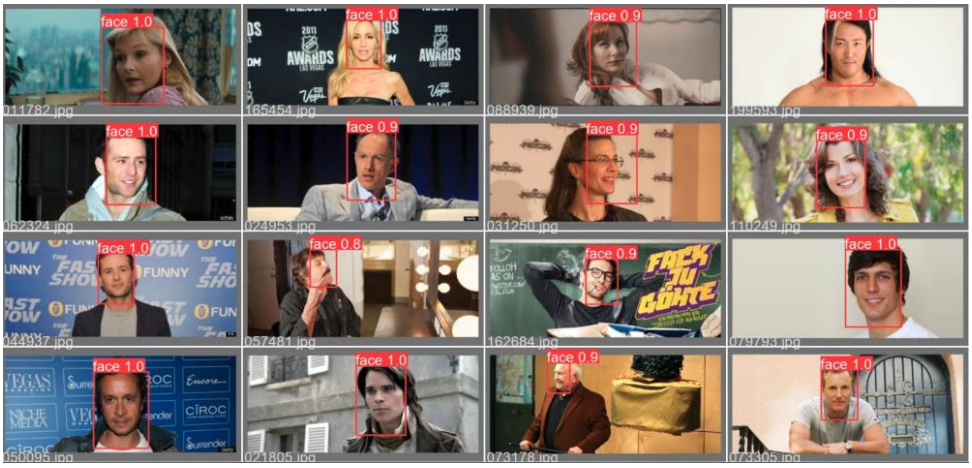

Figure 3: Face detection effect

As can be seen from Fig. 3, face detection is visualize by randomly extracting data from that data set. By using the SGD optimization algorithm to train the two basic models, the learning rates of 0.01, 0.001, 0.0005 and 0.0001 are tried to adjust the hyperparameter training (Figure 4). It can be observed that when the learning rate is 0.01, the training has a large shock and the convergence is slow; when the learning rate is 0.01, the training has a large shock and the convergence is slow. Convergence is relatively faster at 0.001; the learning rate I The convergence is the fastest at 0.0005 and 0.0001, and the convergence is stable at the sixth iteration.

Figure 4: Training convergence with different learning rate loss

It can be seen from Fig. 5 that the visual analysis of the tag correlation of the synthetic data set and the observation of the correlation between different tag data show that most of the images in the data set are distributed in the middle along the horizontal direction, especially the samples with large face size. In the vertical direction, smaller faces are distributed above the picture, while larger faces tend to be distributed in the middle of the picture. The box of face annotation is close to a square, or a rectangle whose height is greater than its length.

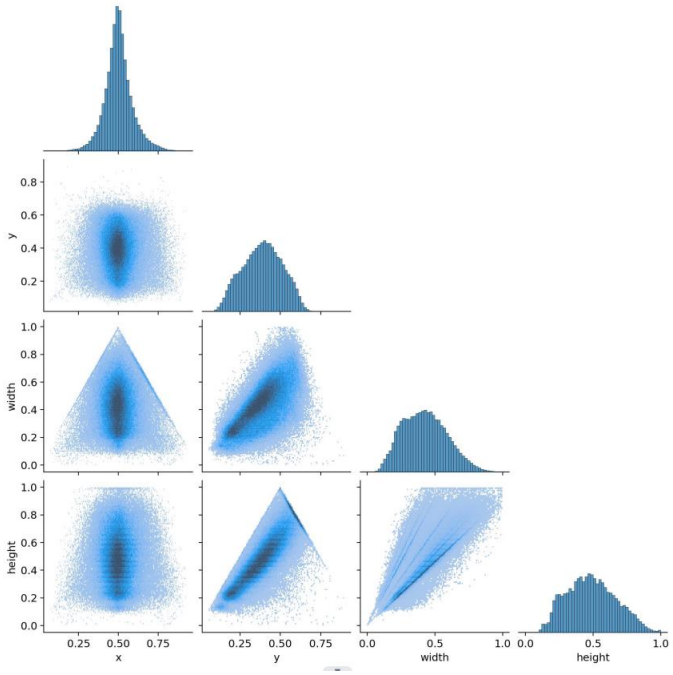

Figure 5: Correlation between labels of a data set

Each attribute is independent of each other, labeled with 0 and 1, and activated with Sigmoid function at training time [18]. In all experiments, resnet is used as the basic network, and all images are normalized to fall on [-1,1]. SGD optimization method is used, and the learning rate is 0.001. The convergence process is recorded and compared. See Figure 6 for details.

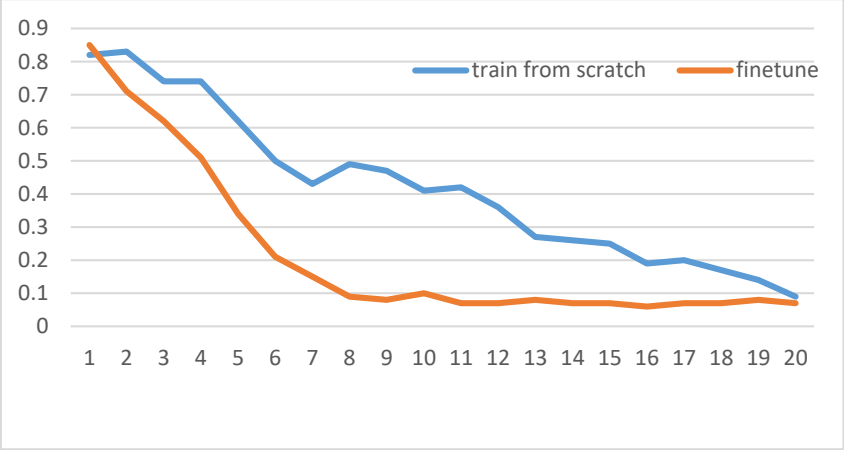

Figure 6: Comparison of loss reduction using transfer learning method and traditional method

As can be seen from Figure 6, orange is the transfer learning method, blue is the new training method, and orange converges faster and tends to stabilize earlier.

4. Experimental results

After 100 iterations of training using tensorflow, the two tasks of face detection and face attribute recognition are visualized, where train/box _ loss is the variance of model prediction bbox and ground truth during training. The train/obj _ loss is the iOU loss for detecting bbox and ground truth, and the train/CLS _ loss is the face attribute classification loss. As can be seen from Fig. 7, it can be observed that the face detection task converges rapidly in the first 20 iterations, while the

face attribute classification converges stably in the early iterations due to transfer learning, and the evaluated mAP fluctuates in the early iterations but tends to be stable after 20 iterations.

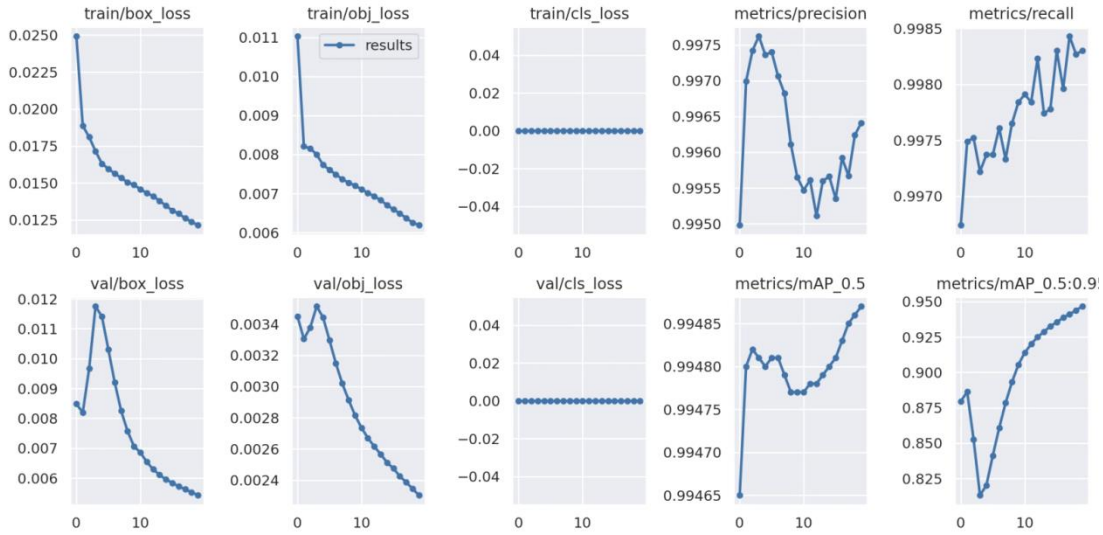

Figure 7: The change of tensorflow training process loss and the validation results on the validation set

It can be seen from Figure 8 that in this paper, the task of face recognition and attribute recognition will be split, face detection model will be used to detect faces, and then face recognition model and face attribute classification model will be used to recognize face attributes, which leads to slower speed. At the same time, face recognition needs to be called many times in the scene with many faces. The mean value of a single frame obtained by reasoning different images on the Jetson Xavier NX [19], in which the blue is the joint recognition method consisting of mtcnn face detection small model and two resnet50 for face recognition and face recognition, and the orange is the unimodular multitask method in this paper. The two groups of models are used to reason on images with face density of 5 and 25 respectively. It can be seen that when the face density is 5, the method in this paper is ahead of the multi-model method. With the increase of face density, the time consumption of the method in this paper is almost unchanged, while the time consumption of the multi-model method is greatly increased.

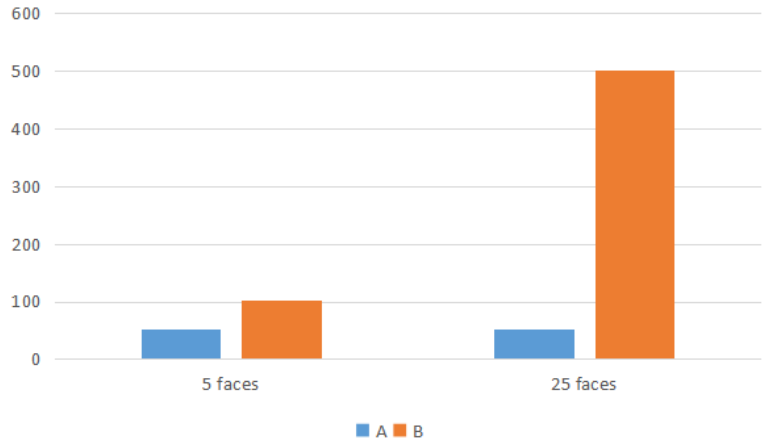

Figure 8: Speed comparison of other methods

The lfw data set is used as a verification set for face recognition, the lfw face is resized to 128 * 128, the resized picture is filled to 608 * 608, and the face feature vector detected and recognized is

used for recognition and comparison. It can be found from Table 1 that the method proposed in this paper can achieve a test efficiency of 98.88 by calculating the accuracy of LFW and comparing with other algorithms.

Table 1: Accuracy Comparison

Model	Accuracy
CosFace[20]	99.73%
Facenet[21]	99.63%
Sphereface[22]	99.42%
Dlib	99.38%
VGG-Face[23]	98.78%
Ours	98.88%

5. Conclusion

At present, the new generation of information technology, represented by 5G, artificial intelligence, big data, cloud computing, Internet of Things and block chains, is accelerating its penetration and integration with various fields of economy and society, and promoting the overall transformation of economy and society. Intelligent education based on the construction of intelligent campus will speed up the reform of talent training mode and teaching methods, and reconstruct the education system. The core of intelligent education is education and teaching, and the fundamental goal is to improve the quality of personnel training. How to break through the key technology of artificial intelligence in the application of intelligent education and lead the demonstration of innovative application of intelligent education has become the focus and difficulty of scholars, universities and companies at home and abroad. This paper constructs a multi-modal collaborative education data detection model for online learning, designs and builds an emotional state monitoring and auxiliary system for online learning, and explores the changing rules and influencing factors of emotional state in online learning, which provides a reference for the design of emotional state intervention strategies and intelligent guidance systems and activities for online learning.

Acknowledgement

This work is partially supported by Shenzhen Education Science “14TH FIVE-YEAR PLAN”2021 Subject: Research on online learning emotion analysis and intelligent tutoring based on collaborative perception of multi-modal education data(ybzz21015), Key technology research and innovative application demonstration of intelligent education(2019KZDZX1048), Guangdong Key Laboratory of Big Data Intelligence for Vocational Education(2019GKSYS001), Shenzhen Vocational Education Research Center Jointly Established by the Ministry and the Province(6022240004Q).

References

[1] Koelstra S, Muhl C, Soleymani M, et al. DEAP: A Database for Emotion Analysis; Using Physiological Signals [J]. IEEE Transactions on Affective Computing, 2012, 3(1): 18-31.

[2] Hao Zheng, Xin Geng, Zhongxue Yang. A Relaxed K-SVD Algorithm for Spontaneous Micro-Expression Recognition [J]. 2016.

[3] Merghani W, Davison A K, Yap M H, et al. The implication of spatial temporal changes on facial micro-expression analysis [J]. Multimedia Tools and Applications, 2019, 78(15): 21613-21628.

[4] Patel D, Hong X, Zhao G, et al. Selective deep features for micro-expression recognition [C]. International

conference on pattern recognition, 2016: 2258-2263.
[5] Peng M, Wang C, Bi T, et al. A Novel Apex-Time Network for Cross-Dataset Micro-Expression Recognition[C].// 8th International Conference on Affective Computing and Intelligent Interaction (ACII). 2019.
[6] Wang C, Peng M, Bi T, et al. Micro-Attention for Micro-Expression recognition.[J]. arXiv: Computer Vision and Pattern Recognition, 2018.
[7] Khor H Q, See J, Phan R C W, et al. Enriched Long-term Recurrent Convolutional Network for Facial Micro-Expression Recognition [J]. 2018.
[8] Xia Z, Hong X, Gao X, et al. Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-Expressions[J]. IEEE Transactions on Multimedia, 2020, 22(3): 626-640.
[9] Verburg M, Menkovski V. Micro-expression detection in long videos using optical flow and recurrent neural networks[C]. ieee international conference on automatic face gesture recognition, 2019.
[10] Khor H Q, See J, Phan R, et al. Enriched Long-term Recurrent Convolutional Network for Facial Micro-Expression Recognition[C].// IEEE International Conference on Automatic Face & Gesture Recognition. arXiv, 2018.
[11] Voigtlaender P, Luiten J, Torr P, et al. Siam R-CNN: Visual Tracking by Re-Detection [J]. 2019.
[12] Vara P, D'Souzakevin B, Bhargavavijay K. A Downscaled Faster-RCNN Framework for Signal Detection and Time-Frequency Localization in Wideband RF Systems [J]. IEEE Transactions on Wireless Communications, 2020.
[13] Ren X, Lei X, Dong N, et al. Interleaved 3D-CNNs for Joint Segmentation of Small-Volume Structures in Head and Neck CT Images [J]. Medical Physics, 2018, 45(5).
[14] Schroff F, Kalenichenko D, Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering[C].// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.
[15] Liu W, Wen Y, Yu Z, et al. SphereFace: Deep Hypersphere Embedding for Face Recognition[C].// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
[16] Deng J, Guo J, Zafeiriou S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition [J]. 2018.
[17] Johannes S. Protein homology detection by HMM–HMM comparison [J]. Bioinformatics, 2005(7):951-960.
[18] Wang H, Song Z. Improved Mosaic: Algorithms for more Complex Images [J]. Journal of Physics Conference Series, 2020, 1684:012094.
[19] Chen Y, Weng Q, Tang L, et al. Thick Clouds Removing From Multitemporal Landsat Images Using Spatiotemporal Neural Networks [J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, PP (99):1-14.
[20] Wang H, Wang Y, Zhou Z, et al. CosFace: Large Margin Cosine Loss for Deep Face Recognition[C].// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.
[21] Schroff F, Kalenichenko D, Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering[C].// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.
[22] Liu W, Wen Y, Yu Z, et al. SphereFace: Deep Hypersphere Embedding for Face Recognition[C].// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
[23] Gyawali D, Pokharel P, Chauhan A, et al. Age Range Estimation using MTCNN and VGG-Face Model[J]. 2021.