A Classroom Students Convergent Behavior Analysis System Based on Image Recognition

Qiang Wang
Network Information Center
Xi'an Jiaotong University
Xi'an 710049, China
wangqiang1989@xjtu.edu.cn

Jingru Cui
Network Information Center
Xi'an Jiaotong University
Xi'an 710049, China
cjr1019@xjtu.edu.cn

Zunying Qin
Network Information Center
Xi'an Jiaotong University
Xi'an 710049, China
qzying@xjtu.edu.cn

Xiaofei Ma
School of Software Engineering
Xi'an Jiaotong University
Xi'an 710049, China
mxf013029@stu.xjtu.edu.cn

Guodong Li
Network Information Center
Xi'an Jiaotong University
Xi'an 710049, China
lgdli@xjtu.edu.cn

2021 3rd International Workshop on Artificial Intelligence and Education (WAIE) | 978-1-6654-0068-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/WAIE54146.2021.00019

***Abstract*—Classroom behavior analysis is an effective way to evaluate the teaching effectiveness in the field of learning analytics. However, traditional classroom behavior analysis mainly focuses on the teacher's observation or manual analysis of the classroom videos, which are time-consuming, laborious and subjective. In this paper, we design and implement a classroom students convergent behavior analysis system which based on image recognition. To adapt to the teaching scene, a student face detection method based on MTCNN (Multi-task Cascaded Convolutional Networks) and a student head pose estimation method based on SSR-Net are proposed respectively. The face detection method is improved through NMS (Non-Maximum Suppression), pooling and convolution to alleviate the problems of partial occlusion, variable posture, small scale and large number of students in the classroom environment. For head pose estimation, we embed the ECA (Efficient Channel Attention) mechanism to improve detection accuracy and speed. We use the face detection and head pose estimation methods to identify the behavior of the student's head-up and then analyze the convergence of students. In the experiments, we first demonstrate the head-up detection approach which is the basic of the convergent behavior analysis is feasible and strong timeliness. Then, the equal interval sampling experiments of different classrooms prove that the convergence behavior analysis of the head-up can accurately feedback students' classroom learning and is practicality for teaching evaluation.**

Keywords—teaching effectiveness, convergent behavior, face detection, head pose estimation, head-up

I. INTRODUCTION

In the new era of education construction, teaching reform focuses on improving the quality of talent training. The classroom teaching, as an important part of talent training, reflects the effectiveness of teaching reform and construction to a certain extent. The analysis of classroom behavior is an effective way to evaluate the quality of classroom teaching, which can directly reflect the teaching effectiveness. However, the traditional classroom behavior analysis of learners focuses on the original methods of manual collection, video observation and questionnaires. At the same time, teachers need to manually mark and record students' classroom behaviors, and cannot provide timely feedback. With the advantages of recording and broadcasting systems, the efficiency of Flanders Interactive Analysis System (FIAS) and S-T classroom teaching analysis method have been improved. However, the above methods all require manual observation and analysis, which is time-consuming, labor-intensive, information lagging, and affected greatly by subjective factors of the analyst. It is insignificant in practice to improve the quality of classroom teaching, and cannot meet the large-scale analysis requirements.

With the maturity of artificial intelligence technology, especially with the breakthrough development of CNN in image recognition such as face detection and gesture recognition, classroom behavior analysis based on image recognition technology has become a research hotspot. For the related research of students' classroom behavior, most of the researches are mainly focused on the study of students' behavior itself, such as physical behavior, facial expression recognition, etc and few researches on convergent behavior. The convergent behavior refers to the similar behavior of students with the same needs in the classroom teaching process. Classroom convergent behavior represents the overall trend, and is a an interactive relationship between students and teachers during the class. Compared with the researches of individual behaviors of students, it can truly reflect the state of the classroom and the participation of students. As an important teaching indices, it is conducive to the improvement of classroom teaching quality.

In this paper, we use image recognition technologies constructing a student's classroom convergent behavior analysis system which based on the head-up rate. First, the classroom student face detection model is established based on MTCNN. MTCNN is improved through three aspects: NMS, pooling and convolution optimization to solve the problems of partial occlusion, variable posture, small scale and large number of people in the classroom environment. Then, a lightweight model

978-1-6654-0068-8/21/$31.00 ©2021 IEEE
DOI 10.1109/WAIE54146.2021.00019

of multi-dimensional student head pose estimation is proposed which based on SSR-Net and improved by embedding the ECA attention mechanism. Finally, we use the pitch angle from the student's head pose estimation model to identify the behavior of the student's head-up. In the experiments, we verify that the head-up detected approach is feasible and the convergence behavior analysis of the head-up rate is practicable.

The remainder of the paper is organized as follows. Section 2 describes related work for face detection, head pose estimation and classroom behavior analysis. Section 3 presents the proposed system mainly include detailed description of students face detection, head pose estimation and head-up detection. Section 4 evaluates the approach of the convergence analysis based on the head-up rate. Finally, we present our conclusions in Section 5.

II. RELATED WORK

For classroom convergent behavior analysis based on the head-up rate, student face detection and head pose estimation are the key components. After the first convolutional network R-CNN for target detection was proposed in 2014, using deep learning for target detection has been a hot area of research for several years. Girshick [1] proposed Fast R-CNN which builds on R-CNN and SPPNet. The Fast R-CNN improves the detection speed and accuracy through multi-task and shared parameter detection. However, the requirement of region proposal makes it cannot be end-to-end. Ren [2] et al. proposed Faster R-CNN which through the RPN (Region Proposal Network) and anchor box to achieve end to end. This approach improves the detection accuracy of multiple targets, but it has low detection speed. Li [3] et al. proposed TridentNet which is a parallel multi-branch architecture and shares the same transformation parameters with different receptive fields. TridentNet could achieve significant improvements, but it cannot satisfy demand for real-time scene. Josephd [4] et al. presented YOLO which is the first detection algorithm based on regression CNN. YOLO is extremely fast, but show low accuracy in small target detection. Kaipeng Zhang[5] et al. proposed MTCNN which a cascaded architecture with three stages of P-Net, R-Net and O-Net for face detection and alignment.

Recently, deep learning for head pose estimation has become a research hotspot. Vishal [6] et al. proposed HyperFace which based on R-CNN and AlexNet for simultaneous face detection, landmarks localization, pose estimation and gender recognition. Amit [7] et al. proposed H-CNN which captures structured global and local features for face key point detection. However, it has strict requirements on facial posture, and performance reduces when occluded. Nataniel [8] et al. proposed a head pose estimation method based on the combination of Resnet classification and regression, using multiple losses to achieve direct regression of the head pose Euler angle. Yang [9] et al. proposed a single-image head pose estimation algorithm based on age estimation network, combined with fine-grained structure aggregation and soft classification regression strategy to achieve regression prediction of head pose angle.

As an effective way to evaluate the quality of classroom teaching, classroom behavior analysis has been an active research topic in recent years. With the rapid development of machine learning technology, students' classroom behavior has changed from manual observation to automatic recognition. Zhang [10] et al. proposed a method of multi-learner posture recognition based on depth images. They used Kinect sensor obtaining learners' upper limb behavior. Then, the posture feature vector was determined by the angle between the upper limb bone structure vector and the modulus ratio. Finally, SVM was used for the recognition of pose vector features. Wang [11] et al. proposed a gesture recognition method from depth maps based on ConvNet. This method used an improved Depth Motion Map (IDMM) converts the depth sequence to into an image, then feds to a ConvNet for recognition. Shi [12] et al. realized the recognition of students' classroom actions of raising hands, standing, listening to class, writing, reading, sleeping, and looking around based on Fisher and L2EMG. They used multivariate Gaussian distribution to extract student behavior characteristics, and introduced Fisher into the generalized learning system to classify student behaviors.

In summary, the current analysis of students' classroom behavior focuses on the identification of individual student' behaviors, and there are few researches on convergent behavior analysis. Therefore, this article aims to use target detection and head pose estimation technologies to realize the analysis of convergent behavior based on head-up in the classroom.

III. ARCHITECTURE AND DESIGN

The architecture of the student convergent behavior analysis system we proposed is illustrated in Fig. 1. The system mainly composes of class image capture, student face detection, head pose estimation and head-up detection. For class image capture, we regularly capture the image from the live-streaming system which has been implemented in our previous work [13]. The following describes the details of each components.

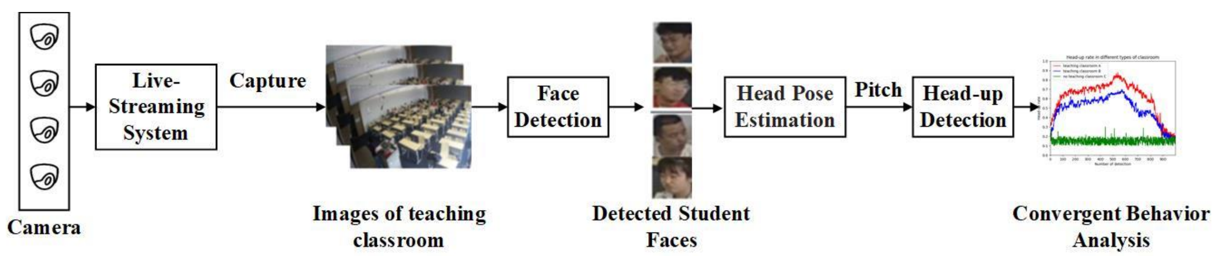

Fig. 1. Architecture of the convergent behavior analysis system

A. Student Face Detection

We construct a student face detection model based on an improved MTCNN (Multi-Task Cascade Convolutional Neural Networks) which is a popular face detection model. For student face detection in the classroom environment, MTCNN has the following problems. First, when a dynamic occlusion occurred or existing small-scale at the back of the class, it will cause misdetection and error detection. Then, P-Net is time-consuming for high-resolution images. Also, when large number of people in the image the time consumed by O-Net and R-Net also increases. In view of above problems, we propose several improvements.

1) NMS optimization

MTCNN employs NMS (Non-Maximum Suppression) to merge highly overlapped predictions and use the predefined IOU (intersection-over-union) threshold to suppress all the other predictions having overlap with the selected predictions in the end of each stage for improving detection performance. Equation (1) is the original NMS score function.

$$s_i = \begin{cases} s_i & , \ iou(M, b_i) < N_t \\ 0 & , \ iou(M, b_i) \geq N_t \end{cases} \tag{1}$$

For the prediction b_i, calculates the IoU with the highest confidence score prediction M. If the IoU is greater than the threshold N_t, the score of b_i will be set to 0. In the case of targets occlusion, the original NMS will bring misdetection. To this end, we optimize the original MS by reducing the confidence score to avoid the suppression of the lower-scoring predictions under occlusion. The improved NMS score function is illustrated in Equation (2).

$$s_i = \begin{cases} s_i & , \ iou(M, b_i) < N_t \\ s_i * \exp(-iou(M, b_i)) & , \ iou(M, b_i) \geq N_t \end{cases} \tag{2}$$

For the IoU is greater than the threshold N_t, we introduce the gaussian function as an overlapping weighting function to cut down the score of the prediction. Under the improved NMS, the prediction more overlap M gets more punishment. Instead, there is no punishment to the prediction which is without overlapping the M.

2) Pooling optimization

In CNN, pooling is a layer structure sandwiched between the continuous convolutional layers. Through maximum pooling or average pooling, the feature maps output by the upper layer can be used as the input image to filter the feature information again. Pooling can reduce the feature dimension and increase the receptive field. At the same time, the model size can be reduced to improve the calculation speed. However, it will lead to the reduction of parameters in the subsequent layers and cause the loss of a lot of important information of the feature image. For example, after a maximum pooling of 3×3, nearly 8/9 of the original image information will be lost. Especially, for small-scale student target detection in the classroom, pooling operation will cause the loss of many students' location information, leading to less accuracy in target detection. To alleviate this problem, we replace the maximum pooling with the same stride convolutional layer in each stage of MTCNN to construct the full convolutional network. In this way, fine-grained features and semantic information will be captured. It can effectively avoid the loss of a large amount of important information, especially the coordinate information in the target detection task, and improve the performance of target detection.

3) Convolutional optimization

Depthwise separable convolution can improve training speed without loss of accuracy. However, depthwise separable convolution directly performs DW (depthwise) operations on the input feature maps causing poor information flow due to the feature channel and space are completely decoupled. Meanwhile, since DW cannot change the channel, the input low-dimensional features are directly extracted from the low-dimensional features, which cannot obtain a large amount of information leading to reduction of accuracy. To solve this problem, we propose lightweight conv-block (LCB) based on the depthwise separable convolution. For each stage of MTCNN, we use the proposed LCB to replace all the convolution except the first convolution in each stage. Fig. 2 shows a simple view of the LCB.

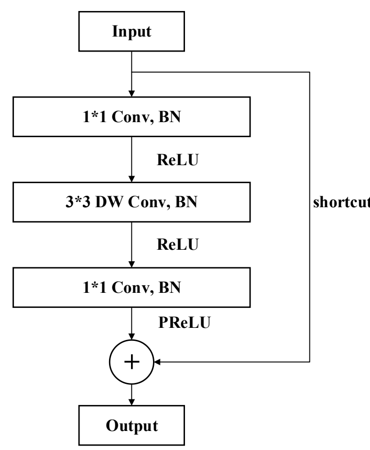

Fig. 2. Overview of the proposed LCB

For LCB, firstly we add a 1×1 convolution ascending layer before DW which can improve the feature extraction ability of convolution benefiting from the high-dimensional tensor. Secondly, to solve the problem of damage or loss of characteristic information caused by ReLU, before output, we replace the ReLU with the PReLU [14] which uses a linear transformation with a slope of α as shown in Equation (3).

$$PReLU(x_i) = \begin{cases} x_i & , \ x_i > 0 \\ \alpha_i x_i & , \ x_i \leq 0 \end{cases} \tag{3}$$

As the number of layers increases, the error increases due to the divergence of the gradient, which makes it difficult to train the network model. Finally, we employ shortcut in LCB to solve the problem of deep network gradient divergence.

B. Head Pose Estimation

In order to improve the ability of students' head pose estimation, in this paper we modify the SSR-Net and embed the ECA attention mechanism to construct a lightweight model of multi-dimensional student head pose estimation. The model consists of Building Module, Fusion Module, and Feature Regression Module.

1) Building Module

Building module builds the basic network structure through two basic blocks BR and BT which is similar to the FSA-Net [9]. The expressions of BR and BT are shown in Equation (4) and (5).

$$B_R(c) = \{SepConv2D(3\times 3, c) - BN - ReLU\} \quad (4)$$

$$B_T(c) = \{SepConv2D(3\times 3, c) - BN - \mathrm{Tan}\,h\} \quad (5)$$

SepConv2D represents a 2D convolution, BN represents batch normalization. ReLU and Tanh are activation functions. c is the number of channels. In order to explore different characteristics, we use a two-stream heterogeneous approach to build the model. Each stream is divided into 3 stages to perform multi-classification tasks to realize the estimation of head attitude angle from coarse to fine. At the same time, based on the SSR-Net, the number of basic blocks is appropriately adjusted to explore more posture angle features for a three-dimensional estimation of the head posture. The network structures for stream1 and stream2 are shown in Equation (6) and (7) respectively.

$$B_R(16) - AvgPool(2\times 2) - B_R(32) - B_R(32) - AvgPool(2\times 2) \quad (6\text{-}1)$$

$$B_R(64) - B_R(64) - AvgPool(2\times 2) \quad (6\text{-}2)$$

$$B_R(128) - B_R(128) \quad (6\text{-}3)$$

$$B_T(16) - MaxPool(2\times 2) - B_T(32) - B_T(32) - MaxPool(2\times 2) \quad (7\text{-}1)$$

$$B_T(64) - B_T(64) - MaxPool(2\times 2) \quad (7\text{-}2)$$

$$B_T(128) - B_T(128) \quad (7\text{-}3)$$

2) Fusion module

The structure of fusion module is shown in Fig. 3. For each stage, the feature maps from two streams are input into a 1×1 convolution and a pooling respectively. Then, we fuse a size of W×H×C feature map U by means of two different activation functions which can gain different levels of features at different stages. For obtaining more expressive features, we add attention mechanism in fusion module. The attention mechanism can enhance the effective head posture, meanwhile restrain interference head posture. It can reduce error and improve detection performance. Although SE [15] and CBAM [16] can obtain higher accuracy, we adopt ECA (Efficient Channel Attention) [17] due to multi-classification strategy and balance accuracy and speed. ECA adopts a local cross-channel interaction strategy to avoid dimensionality reduction and effectively capture the information of cross-channel interaction. As shown in the figure: feature map U generates the attention feature map A through the ECA module. Finally, we fuse U and A to obtain the final optimized feature map U′ and pass it to the next module.

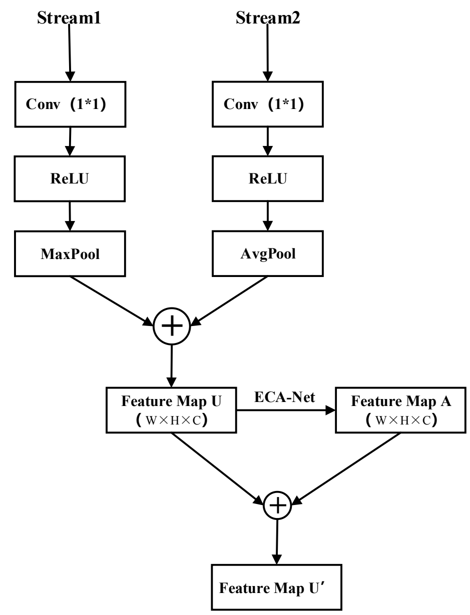

Fig. 3. Overview of fusion module

3) Feature regression module

The feature map U′ output by the fusion module is used to generate three feature vectors of probability distribution, offset vector and scale factor through a fully connected layer and activation function. The K set of feature vectors $\{p^{(k)}, \eta^{(k)}, \Delta k\}$ are used for the estimation of Pitch, Roll and Yaw by means of SSR [18]. The output of SSR is the attitude angle in the form of Euler angle.

C. Head-up Detection

We employ student face detection model and head pose estimation model to calculate the head-up rate. Then, analyze the convergence based on head-up rate in the real teaching scence. The head-up detection procedure is shown as Fig. 4. First, the classroom image captured from live-streaming system is input into the improved face detection model based on LCB. The face detection model will output the instance bounding boxes of student face and the corresponding scores. The bounding boxes which the *improved_NMS (score)* is greater than the thresh are added to the *detected_face_list*. Then, through the head pose estimation model with ECA, we will get the pitch angles of each detected faces. Finally, the face which pitch angle in $[\beta_1, \beta_2]$ will be marked as head-up and added to the head_up_list. The β_1 and β_2 are specified according to the specific classrooms.

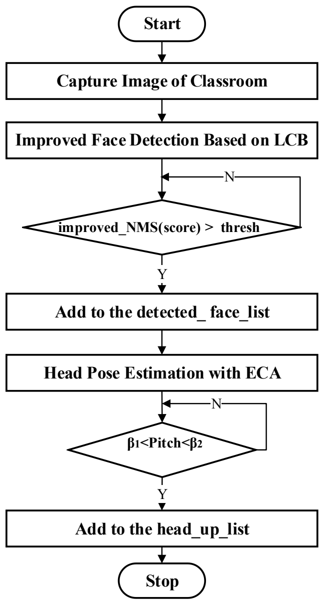

Fig. 4. Procedure of head-up detection

1) Head-up Recognition

we use the pitch angle from the student's head pose estimation model to identify the behavior of the student's head-up behaviors. Usually, peoples are accustomed to look directly at the target object of attention rather than obliquely. We establish a coordinate system based on the real classroom environment and calculate the angle between the student's line of sight and the blackboard boundary. The classroom coordinate system is shown in Fig. 5.

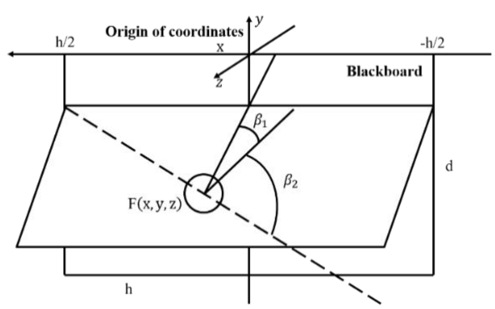

Fig. 5. Classroom coordinate system for head-up recognition

With the middle point of the upper boundary of the classroom blackboard as the origin of coordinates, select the left direction of the center point and the vertical direction of the center point as the x and y axes in turn, and make the direction pointing to the student's vertical x, y plane as the positive z axis.

If the length of the blackboard is h, the width is d, and the coordinates of the center point of the student's head are (x, y, z). The angle between the center point and the upper and lower borders of the blackboard β_1 and β_2 are the maximum rotation range of the student's head posture pitch. β_1 and β_2 indicate the pitch range of students in the classroom. When the ptch rotation range exceeds the threshold, that is, the student's head's line of sight is not concentrated within the boundary of the blackboard, it is regarded as a non-head-up state. The calculation of the pitch rotation range is shown in Equation (8).

$$\begin{aligned}\beta_1 &= -arctab(y/z)\\ \beta_2 &= -arctab((d-y)/z)\end{aligned} \tag{8}$$

2) Head-up rate calculation

Classroom student head-up rate is the ratio of the effective head-up students to the total number of students. Through time series analysis of the whole class head-up rate, the regularity of student behavior could be further explored in the time dimension. The calculation formula is shown in Equation (9). r^{ent} represents the average head-up rate of the time interval（$T_n - T_1$）, m_i and M_i represent the head-up and total students in the i-th detection respectively. n is the number of detections in（$T_n - T_1$）.

$$r^{ent} = \frac{1}{T_N - T_1}\sum_{i=1}^{n}\frac{m_i}{M_i}\times 100\% \tag{9}$$

IV. Evaluation

In the experiments, we first evaluate head-up detection approach compared with manual in classroom environment. Then, analyze the convergence behavior on head-up in different classes to prove that it is practicality and can reflect students' classroom learning effect.

A. Evaluation of Head-up Detection

Since the head-up rate is the basic of convergence analysis, we evaluate the effect of the head-up detection approach we proposed. We compare the detection algorithm against MTCNN, FSA-Net and the manual statistical results in a real teaching environment image set. The image set contains 200 images which is captured from the live-streaming system every 3 seconds and lasts 10 minutes. The experiment results are shown in Fig.6.

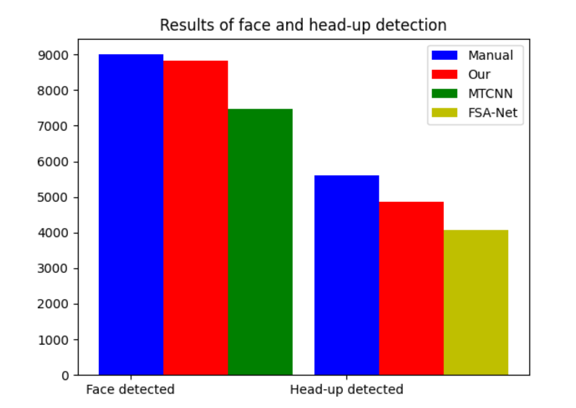

Fig. 6. Results of face and head-up detection

For the image set, each image contains 45 students, so total 9,000 students should be detected during the experiment. In effect, we detect a total of 8,514 students which can acquire 94.6% accuracy and improve 18% compared with MTCNN. For student face detection in the classroom environment, the error comes from students occluding each other. We also manually count the number of heads up for each image, a total of 5,616 people looks up during the 10 minutes. In our head pose estimation approach, we detect a total of 4862 head-up, which can acquire 86.8% accuracy and improve 19.1% compared with FSA-Net. The error also comes from students occluding. At the same time, the total time of the 200 image experiments is 378 seconds, and the average time for each image is 1.89 seconds. The results show that the head-up detection method has the characteristics of accuracy and real time, and can meet the detection demand of real teaching environment.

B. *Convergence Behavior Analysis*

To verify that the head-up behavior of students has convergence over a period of time. We select multiple teaching and no-teaching (such as self-study and examination) classrooms for comparative analysis. We calculate head-up rate of the different classrooms every 3 seconds and lasts 50 minutes which is a whole lesson. The compared result is shown in Fig. 7. The head-up rate is fluctuant during the whole lesson. In the first 550 detections student head-up rate is increasing overall, proves that the convergence of head-up and concentration is gradually increasing. After 550 detections students head-up rate and convergence is gradually decreasing. On the contrary, the head-up rate of no-teaching classroom maintains a relatively steady state. Experiments prove that students' head-up behavior do have the convergence during the teaching.

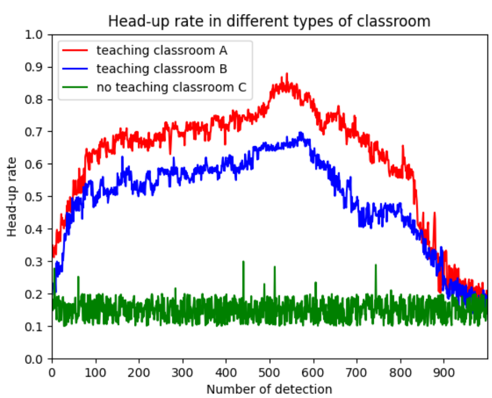

Fig. 7. Head-up rate in different types of classroom

To investigate the variation for head-up convergence over time, the whole lesson is divided into several time periods and analysis using the equal interval sampling method. We calculate the average head-up rate every 5 minutes and the result is shown in Fig. 8. In the first 5 minutes, the head-up rate is lower than 0. 5 and weak convergence. From 5 to 30 minutes, the head-up rate is gradually increasing and reaches the peak of class concentration between 20 and 30 minutes. After about 30 minutes, the head-up rate begins to decline but remains above 0.45. During this time, the head-up behavior is still very similar and focused. However, in the last 10 minutes, the head-up rate is lower than 0.5, indicating that the students' concentration gradually decreases during this time. The experiment results are in good agreement with actual teaching situation. According to the head-up convergence investigated from the experiments, teachers could suitably schedule the teaching contents. The major content should to be taught in the middle of the class, and it is best not to lay out important knowledge in the end of the class. Meanwhile, to achieve good teaching effectiveness, teachers also can remind students to adjust when weak head-up convergence happens. In addition, the convergence of head-up represents the overall trend of students' behavior which can be one of the important metrics for the evaluation of teaching quality. Further, the analysis of classroom behavior is an effective way to evaluate the quality of classroom teaching, which can directly reflect the teaching effectiveness.

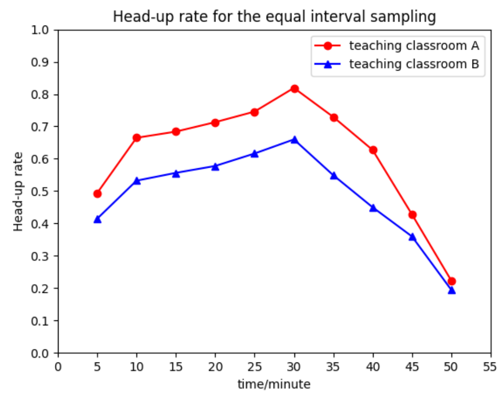

Fig. 8. Head-up rate for the equal interval sampling

V. Conclusion

In this paper, we use image recognition technologies constructing a students classroom convergent behavior analysis system which based on the head-up rate. A student face detection method based on MTCNN and a student head pose estimation method based on SSR-Net are proposed to adapt to the teaching scene. we use the pitch angle from the student's head pose estimation model to identify the behavior of the student's head-up and do convergent behavior analysis. In the experiments, we verified that the head-up detected approach is feasible and the convergence behavior analysis of the head-up rate is practicable. In addition to the head-up behavior，facial expression is also an important aspect for convergent analysis. In future work, we will try to do convergent analysis combining with facial expression.

References

[1] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[2] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[3] Y. Li, Y. Chen, N. Wang and Z. Zhang, "Scale-Aware Trident Networks for Object Detection," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6053-6062.

[4] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.

[5] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, Oct. 2016.

[6] R. Ranjan, V. M. Patel and R. Chellappa, "HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 1, pp. 121-135, 1 Jan. 2019.

[7] A. Kumar, A. Alavi and R. Chellappa, "KEPLER: Keypoint and Pose Estimation of Unconstrained Faces by Learning Efficient H-CNN Regressors," 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017, pp. 258-265.

[8] N. Ruiz, E. Chong and J. M. Rehg, "Fine-Grained Head Pose Estimation Without Keypoints," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 2155-215509.

[9] T. Yang, Y. Chen, Y. Lin and Y. Chuang, "FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation From a Single Image," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1087-1096.

[10] ZHANG Hong-yu, LIU Wei, XU Wei, et al. Depth Image Based Gesture Recognition for Multiple Learners[J]. Journal of Computer Science,2015Vol.42 No.9

[11] P. Wang, W. Li, S. Liu, Y. Zhang, Z. Gao and P. Ogunbona, "Large-scale Continuous Gesture Recognition Using Convolutional Neural Networks," 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 13-18.

[12] Shi Y , Wei Y , Pan D , et al. Student Body Gesture Recognition Based on Fisher Broad Learning System[J]. International Journal of Wavelets Multiresolution and Information Processing, 2018, 17.

[13] Q. Wang, B. She, J. Cui, Z. Qin and G. Li, "Design and Implementation of Distributed Video Live-Streaming System Based on SRS," 2021 10th International Conference on Educational and Information Technology (ICEIT), 2021, pp. 233-238.

[14] K. He, X. Zhang, S. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026-1034.

[15] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011-2023, 1 Aug. 2020.

[16] Sanghyun Woo, Jongchan Park, Joon-Young Lee and In So Kweon, "CBAM: Convolutional block attention module", ECCV, 2018.

[17] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11531-11539.

[18] Yang T Y , Huang Y H , YY Lin, et al. SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation[C]// Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18. 2018.