Inferring the Students' Attention in a Machine Learning Approach: A Feasibility Study

Jiachun Li,Ruiqi Li, Yuting Zhou, Junlin Xian, Xiong Zhang, Xiaojun Hei
School of Information Science and Engineering
Huazhong University of Science and Technology, Wuhan, China
Email:jiachunli@hust.edu.cn,jlxian@hust.edu.cn,heixj@hust.edu.cn

***Abstract*—This paper proposes a multi-modal intelligence teaching system, which uses ordinary student data as a research sample, including video and audio. Through comprehensive analysis as well as the machine learning method, feedback on the student's listening status in real time can effectively help teachers adjust and improve the teaching strategy. This can strive to achieve personalized teaching content recommendation and counseling for students. Through machine learning, the instant feedback of this multi-modal smart classroom can provide a good help to students' academic development.**

Keywords-smart classroom, sensor, face recognition, feedback

I. Introduction

Recently, Y. Kim et al. used data such as eyeball trace, arm gesture etc. to evaluate and promote teaching process with a Restricted Boltzmann Machine[1]. In 2018, Anusha James et al. looking at utilizing video and audio data to classified the activeness of interaction, thus elevating effect of teachers to students[2]. R. Martinez-Maldonado et al. describe a system to monitor the situation of group discussion[3]. The work of N. Gligoric et al. focus on assist teaching through the IoT(Internet of Things) equipment[4]. However, this paper focuses on designing a real-time teaching assist system with a multi-model information analysis method, and will consider designing a specific scene for special people like the autism.

II. Problem Elicited

At present, the recognition of students' listening status is still a difficult problem,it is difficult to achieve full coverage by manual observation. Therefore, we propose video audio multimodal analysis of students' listening status.

Our conceptual classroom is located at Huazhong University of Science and Technology, where many sensors, cameras and wireless microphones are placed to capture audio and video information from students. We apply the edge computing network in the complex scene of the future classroom, aiming to improve the atmosphere, optimizing the student listening experience, and realize the multi-modal smart classroom.

III. System Model

The system uses an edge computing network, and the main module is a CNN(Convolutional Neural Networks) network. The CNN network is responsible for connecting edge nodes and collecting vectors obtained by sensors and other structures. Edge nodes, such as various sensor structures, capture information such as student gestures, pupils, sounds, faces, heart rate, etc., and convert the original audio and video information into processed audio and video vectors to quantify the student's attendance status. The output of the CNN network is the state of attentive and inattentive, specifically the fluctuation curve of the number of children who concentrate on listening to the time axis. Teachers can use this information to adjust the lecture strategy and rhythm. We can also make personalized recommendations for student learning based on the high-frequency vocabulary in the student discussion.

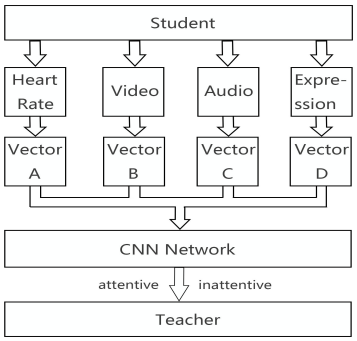

Fig. 1. System model

IV. System Design

A. Video CNN network

The classifier is trained by constructing the CNN network to judge the student's listening state in real time. In the experiment, the face information in the video is mainly collected for network training. And this is based on the implementation of the facial feature point landmark: we pre-processing the input image, accurately identify the positions of different features of the face, and mark 32 pixels. The 400 pictures captured by the video are manually trained, and label two different states: attentive and inattentive. So we can form a trusted input vector to the CNN network, which can count the output of the attentiveand the inattentivestate, and make the change rule of the student's attention concentration with the time axis, including the instant feedback on the seat lecture information. We can automatically implement the following training process during class, which can be applied in real time during teaching.

The data set is constructed based on the orientation of the face to identify the state. The basic unit of the training network

consists of a convolutional layer, an active layer, a convolutional layer, and a pooling layer. The Dropout method with a parameter of 0.5 is used to avoids overfitting. The training uses the SGD method for gradient descent, and the Batch Size is 32. The training image is a 200*200 black and white image, which has 40 iterations. The network contains 32 convolution kernels, and the pooling layer size is 2 units. We use multiple methods to capture and evaluate video information.

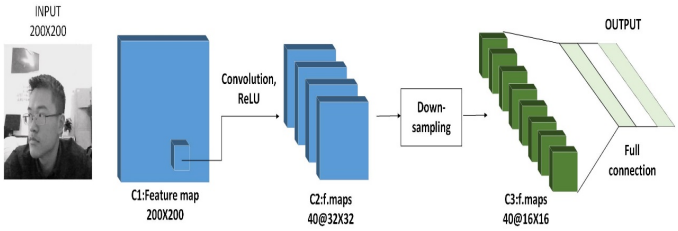

Fig. 2. CNN network

B. TextCNN network

For the processing of sound information, the pkuseg participle is first used to translate the voice information into a statement, The accuracy of the translation is obtained by the change of the sampling rate to determine the appropriate sampling range. Referring to the word2vec model, finally we got a more accurate word vector character.

The TextCNN network is supplemented and the translated statement is input into the trained TextCNN network to obtain the score of the feature vector. The state acquired by the video CNN network is integrated to judge the state of the student's lecture. When the output synthesis of both networks is judged to be attentive, the status of the classmate is confirmed as concentrating in the class.

V. Performance Evaluation

During the experiment, the CPU used i5-7300hq, memory 8g, and the graphics card used Nvida GeForce 1050ti. By training 400 sets of data, we obtained the loss function and the confirmation function curve of the figure.

Loss function maps an value of variables onto a real number intuitively representing some "cost" associated with the event. Manually observe the results of the listening and non-listening states, we compare the output of the CNN network and divide the matching by the total amount of training as the accuracy rate. Defining the delay is the delay from inputting the training data to generating the attentive and inattentive states to the teacher interface. After being optimized, it is shortened to the instant operation. We will also try algorithmic acceleration based on kinect and FPGA to reduce the latency.

Figure 3 shows that as the number of training increases, the training loss function continually decreases, and the compliance with the validation loss function increases the experimental accuracy.In terms of audio, we continue to reduce the error rate of recognition by increasing the sampling rate, focusing on the extraction of high-frequency vocabulary. It can be seen that choosing a suitable sampling rate is a significant aid to the translation of the voice signal.

When analyzed only by video, the accuracy curve is as a in figure 4, and the input is only for audio single-mode data. Combining the two kinds of data, the loss function is more fitted. The performance and accuracy are greatly improved, which is increased from 71.3 percent to 87.4 percent.

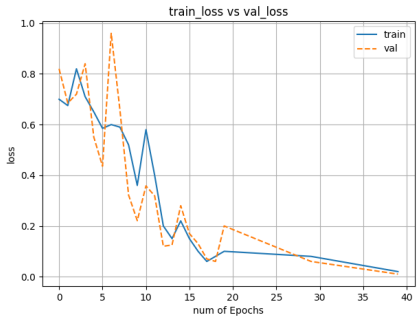

Fig. 3. loss function

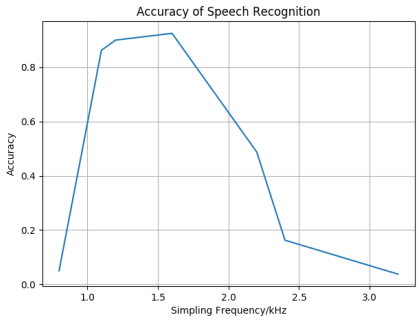

Fig. 4. audio result

VI. Conclusion and Future Work

This paper adopts multi-modal analysis method, which is a smart classroom aid system developed for students. It can provide good feedback for teaching and can also provide personalized learning recommendations for students. The audio and video analysis tools are combined with the techniques related to machine learning face recognition, and the experimental algorithms are continuously optimized.

We're also going to use kinect for algorithm acceleration. Kinect can collect high-quality images every second, which is easy for face detection and tracking, human-computer interaction, 3D modeling, and it can also use LBP (local binary pattern), Gabor, LDA(linear discrimination Analysis) to extract features, improving the accuracy of class status judgment. We will apply more statistical methods to verify the system, and try to apply as early as possible in actual teaching for autism.

References

[1] Y. Kim, T. Soyata, and R. F. Behnagh, "Towards emotionally aware ai smart classroom: Current issues and directions for engineering and education," *IEEE Access*, vol. 6, pp. 5308–5331, 2018.

[2] A. James, M. Kashyap, Y. H. V. Chua, T. Maszczyk, A. M. N?ez, R. Bull, and J. Dauwels, "Inferring the climate in classrooms from audio and video recordings: A machine learning approach," in *IEEE International Conference on Teaching, Assessment and Learning for Engineering*, 4-7 December 2018.

[3] R. Martinez-Maldonado, K. Yacef, and J. Kay, "Data mining in the classroom: Discovering groups' strategies at a multi-tabletop environment," in *Educational Data Mining 2013*, 2013.

[4] N. Gligoric, A. Uzelac, S. Krco, I. Kovacevic, and A. Nikodijevic, "Smart classroom system for detecting level of interest a lecture creates in a classroom," *Journal of Ambient Intelligence and Smart Environments*, vol. 7, no. 2, pp. 271–284, 2015.