Automatic Student Engagement in Online Learning Environment Based on Neural Turing Machine

Xiaoyang Ma, Min Xu, Yao Dong, and Zhong Sun

***Abstract*—With the continuous and rapid growth of online courses, online learners' engagement recognition has become a novel research topic in the field of computer vision and pattern recognition. While a few attempts to automatic engagement recognition has been studied in the literature, learning a robust engagement measure is still a challenging task. To address it, we propose a new automatic engagement recognition method based on Neural Turing Machine in this paper. In particular, we firstly extract student's eye gaze features, facial action unit features, head pose features, and body pose features respectively, then combine these multi modal features into the final feature of our recognition task. Moreover, we propose the engagement recognition framework based on the idea of Neural Turing Machine to learn the weight of each short video feature. In consequence, the feature fused by different weights will be applied to identify the students' engagement in learning online courses. Empirically, we show improved performance over state of the art methods to automatic engagement recognition on DAiSEE dataset.**

***Index Terms*—C3D network, engagement recognition, features fusion, neural turing machine, OpenFace.**

I. Introduction

With the enrichment and development of Internet technology, online education courses are becoming more and more popular with students. Students can gain knowledge through online education anytime, anywhere, without being constrained by time and space, and can repeatedly watch video that they do not understand. Greatly enhancing the convenience of students' learning, while promoting the sharing of educational resources.

Engagement is one of the affective state [1] which is a link between the subject and resource. It has various aspects such as emotional, cognitive and behavioral aspect. In the online education environment, to help the students to retain their attention level or track those parts of the video where they loss the attention it is mandatory to track student engagement based on a series of external expressions such as the subject's emotions [2], behaviors, and postures. Now, a large number of research experts at home and abroad have recorded the process of students' online learning by video, and then judged the student's learning engagement by observing the behavior of the students in the video, however, there are also some scholars who conduct research on student engagement recognition through questionnaires, quizzes [3], etc. But engagement intensity is a complex emotional state, which is not only related to the subject's emotions [4], but also to the subject's posture and some actions. Judging students' engagement in learning through limited video data is still a serious challenge.

In this paper, our task is to automatically identify students engagement in learning from multiple feature perspectives, making full use of student's eye gaze information, head pose information, facial action unit information, and body posture information. We think that the student's engagement in learning can be reflected by some behaviors [5] of the students during the course of the lesson. Such as when the students look away from the screen or outside the screen range, when watching the video, the eyes are half closed or fully closed, yawn and restlessness in a chair, etc. These behaviors reflect that the student's engagement in learning is not high, otherwise it indicates that the student has a higher degree of engagement in learning.

We first fuse the extracted features from multiple angles into a new feature, and then use the idea of Neural Turing Machine [6] read heads to perform video feature fusion. Through the feedforward of neural network and the back propagation process of the neural network, self-learning a group weights, and then weight the sum of multiple video features to fuse into a complete video feature.

In the traditional method of video feature fusion is to sum multiple short video features and average them. This method defaults that the quality of each short video feature is the same, but due to factors such as lighting and information volume, the quality of each short video feature is different, so we propose a new fusion method.

We consider the engagement recognition of learning as a classification task, the purpose of which is to use Neural Turing Machine [7] read heads to build a model to minimize the cross-entropy loss function between the predicted value and the actual engagement value. The database used in this paper is DAiSEE, which includes video data of 112 students participating in online courses in real environments.

According to the research standards of domestic and foreign experts and scholars, the level of learning engagement is divided into four levels [1], [8], from 0 to 3 representing completely disengagement , Barely engaged, engaged and highly engaged, respectfully. As the value increases, the level of student engagement increases. Examples of four different levels of engagement in the DAiSEE are shown in Fig. 1.

II. Related Work

In the work of Whitehil *et al*. [8] in 2014, research was conducted on the question of whether student engagement

Manuscript received February 24, 2020; revised November 16, 2020. This research was partially supported by the National Key Research and Development Program of China (No. 2018YFB1402903), the National Natural Science Foundation of China under Grants 61601310, 62177048.

The authors are with Capital Normal University, Beijing, China (e-mail: 2181002054@cnu.edu.cn, xumin@cnu.edu.cn).

doi: 10.18178/ijiet.2021.11.3.1497

can be identified through student facial expressions. This paper is one of the most influential studies by far which combine computer vision with machine learning method for student engagement recognition.

Fig. 1. Label 0 to 3 show four different engagement intensity levels: [0(low)-3(high)].

They mainly make two contributions: 1. for artificial tagging, two video lengths are used, that is, a long clip of 60 seconds and a short clip of 10 seconds. Experiments have proven that people have a better understanding of 10-second short clip videos. 2. They identify student engagement by extracting manual features and use the classifier to classify engagement. The experiment results show that, in the short video binary classification task of 10 seconds, the recognition result of the machine learning model is equivalent to the level of manual annotation. It is clear that the method of machine learning can be used for automatic recognition of learning participation.

Since then, domestic and foreign scholars have carried out a series of more in-depth studies on the prediction of student engagement. The main research methods include frame-based methods, space-time-based features methods, and multi-modal features-based methods. Besides, Spatio-temporal-based methods are sometimes used in combination with the other two methods.

- The frame-based method is to use a traditional manual method or convolutional neural network (CNN) [9] to perform single feature extraction on 2D images. After extracting corresponding low-level features or high-dimensional features, use a certain feature fusion method to obtain video features, where the way of weighted average sum is the most common and effective feature fusion method. The work of Whitehil et al. in 2014 was based on a frame-based method. Three sets of comparative experiments were performed. Different filters were used to extract the facial features and the classifier was used to determine the level of student engagement. Jacob et al. [10] trained their CNN model on a large number of existing datasets, and then saved the trained model for feature extraction on their own datasets.
- The method based on spatio-temporal features is to use the Long Short-Term Memory (LSTM) and Convolutional 3D to extract the temporal and spatial features of the video, mainly to save the time information of the input signal. The multi-modal feature-based method refers to a combination of feature analysis of multiple angles of a student to determine the state of student engagement. In the EmotiW-2018 challenge, Yang *et al*. [11] proposed a multi-instance learning framework that integrates traditional machine learning features from multiple perspectives such as OpenFace, OpenPose, LBP-TOP, and deep learning features such as 3D convolution (C3D) [12]. The fused features are input to the LSTM network to determine the degree of participation, which improves the accuracy of the regression results of engagement recognition and won the championship. Monkaresi *et al*. [13] collected facial videos and heart rate data of 22 students at the same time, using face tracking features (from Microsoft Kinect), LBP-TOP features, and heart rate features to construct machine learning models to predict students'engagement. The elimination experiment results show that the facial tracking features from Kinect have the best experimental results among the three types of feature input.

III. Our Approach

We use OpenFace 2.0 [14] and 3D convolution network model these open-source tools to extract students' multi-modal features in online learning videos, and analyze students'facial expressions, head pose, eye movements, body movement postures, etc. In order to achieve a more reasonable fusion of multiple short video features into a complete video feature, we have adopted a new method: through the increase of neural network training rounds, each group of short video features is compared with the label value and learns one. Learn appropriate weights to better fit the engagement status.

A. Experimental Technology

OpenFace 2.0: this is an open source face framework that includes landmark, head pose, action unions, eye gaze, etc., and contains all source code for training and detection. OpenFace uses Conditional Local Neural Field (CLNF) to detect and track key points on the face. OpenFace is suitable for use in computer vision, machine learning and affective computing communities.

C3D (Convolutional 3D): a general, compact, simple, and efficient model proposed by Tran *et al*. This model is a method based on 3D ConvNets for extracting spatio-temporal features of human behavior. Compared with other methods, the features learned through the C3D model can be used to input a simple linear classifier to obtain better results. Features obtained through C3D extraction encapsulate information related to objects, scenes, and actions in the video, making them useful for a variety of tasks without having to assign a model to each task.

2D ConvNets would lose the time information of the input signal after each convolution operation and can only be modeled spatially. However, in 3D ConvNets, convolution layer and pooling layer operations are performed on space and time, which can better model time information. Compared with 2D ConvNets, 3D ConvNets are more suitable for spatio-temporal feature learning.

B. System Pipeline

Fig. 2 shows our proposed experimental framework. First, we divide the video into a set of continuous frames as input to OpenFace and C3D networks, and extract features separately. Then we serially stitch the multi-modal features to combine them into a complete feature. In the experiments of Niu *et al*.,

they proposed the idea of Gaze-AU-Pose (GAP) [15] features, which make the eye gaze characteristics, head pose characteristics, facial action unit characteristics serially stitched. We added body pose characteristics based on their experiments, and then input a series of short video features into a fully connected layer and a softmax classification layer to ensure that the learned values meet the probability distribution. Each of weight parameters is self-learned through a deep neural network. The weight of the short video features makes the high-quality short video features account for a larger proportion of the complete video features, which is more conducive to the accurate judgment of student engagement and improves the accuracy of the experiment.

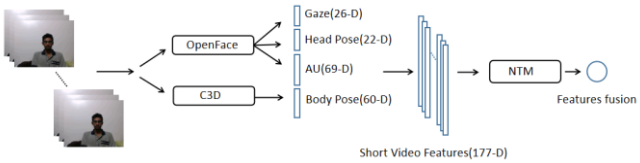


Fig. 2. The system pipeline of our approach.

C. Pre-processing

If the multi-modal features of the entire long video are directly extracted, the details of local video features will be missing, or a large number of video features will be redundant. In order to avoid these shortcomings, we take the method of segmenting a long video and divide a long video V into m segments , such as $V = [V_1, V_2, \ldots V_m]$, where V_i represents the i-th video segment, extracting 4 features for each video segment, and finally merge the features of the m video segments into a video feature. In our work, $m = 3$, and each video segment is used as input for our experimental model. The video is approximately 10s in length, the training and validation video data are captured at 30fps and we preserve the video frame to a resolution of 640 × 480 at 10fps.

D. Multi-modal Features

Evaluating students' engagement in learning can be based on multiple perspectives [16], [17]. Generally, we believe that students' facial expressions [18], head posture, eye gaze status, body posture [19] and other factors can reflect the intensity of students' engagement in learning. If the student's facial expression is pleasant and serious, his eyes are focused on the screen, his body posture is upright, and there is no unnecessary action except writing and thinking, then we think that the student's learning engagement is high [20]. For the above reasons, we extract our features from four perspectives including OpenFace 2.0 and C3D features. The gaze, head-pose, facial action unit features are captured by OpenFace while body-pose features are extracted by C3D network.

Eye gaze: feature extraction was performed on the eye position of the research subject through OpenFace 2.0, and we can obtain the gaze direction vector and radian gaze direction. In addition, we calculated the difference between the direction vector of each frame in each dimension of the video clip and the average position of all frames value to further describe gaze information. In order to comprehensively analyze the eye condition, we calculated the average position of all detection points of each eye and added it to the feature.

Head pose: we first describe the head pose information using the extracted head position of each frame and the head direction vector expressed in radians, and then calculate the difference between the direction vector of each frame and the average position of all frames, which is added to the feature. In addition, we also detect the 68 facial key points and use the average position of all key points in each frame to represent the head posture state. Find the average position of all key points in all frames in the video clip, then calculate the arithmetic mean, subtract the difference between the result and the average position of all key points in each frame, and use it in combination with the above pose features.

Facial action unit: we need to focus on 18 facial key points (AU 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, 45) and use OpenFace 2.0 to extract the facial action unit feature of the research object to evaluate the existence frequency [20] and intensity of each AU. In order to further analyze facial expressions, we calculated the difference between the intensity of each AU frame and the average intensity of all its frames, and the proportion of each AU intensity in all AUs, and added it to the final feature.

Body pose: recently, deep learning technology has made great contributions to the development of computer vision. We chose 3D ConvNet because it can better model the time signal of the input information through 3D convolution and 3D pooling operations, thereby ensuring the timeliness of the input data. The C3D network model includes 8 convolutional layers, 5 maximum pooling layers to extract temporal and spatial features, 2 fully connected layers, and a softmax layer to obtain video classification results. Our C3D network model is only used to extract body pose features, removes the softmax layer, and adds a fully connected layer.

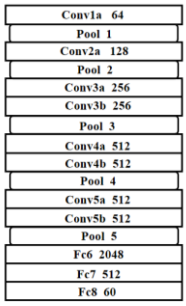


Fig. 3. The structure of our C3D model which include 8 convolutional layers, 5 maximum pooling layers and 3 fully connected layers.

Among them, all 3D convolution kernels are 3, and the stride is 1. The 3D pooling layer is represented by pool1 to pool5. In order to save the time information of the previous stage, the first pooling layer kernels are 1×2×2, and stride is 1×2×2.However, the remaining 3D pooling layers are 2, and the stride is 2. The first fully connected layer output dimension is 2048, the second fully connection layer is 512, the third fully connection layer is 60 .Then C3D features are extracted by body images in a segment which is composed by 16 consecutive frames. The C3D structure is shown in Fig. 3.

E. Feature Fusion

Different from the traditional short video feature fusion method, we adopt the idea of Neural Turing Machine read

heads to fuse multiple short video features. The Neural Turing Machine (NTM) [6] architecture contains two basic components: a neural network controller and a memory bank. Like most neural networks, the controller network receives inputs from an external environment and emits outputs in response. The workflow of the NTM mainly includes two parts: read heads and write heads.

Read heads: we consider memory as an $N \times M$ matrix M_t, where t represents the current moment and that memory will change over time. Our reading process is to generate a positioning weight vector W_t with a length of N, which represents the memory weight corresponding to N addresses. The last read heads vector R_t is as in:

$$R_t = \sum_{i}^{N} W_t(i) M_t(i) \tag{1}$$

Besides, the sum of the weight vectors is 1.We consider a short video feature as an $F \times 1$ vector. The short video feature is represented by V_i, and the complete video feature is represented by V:

$$V_i = \left[V_{1i}, V_{2i} ... V_{Fi}\right]^T \tag{2}$$

$$V = \left[V_1, V_2 ... V_N\right] \tag{3}$$

Initially, a set of weight parameter vectors W is randomly initialized, and its length is F, which represents the proportion of the features corresponding to the F positions. We get the calculation results of W and V through the feedforward neural network that can be called Q.

As we all know, the vector Q is composed of Q_1, Q_2 ... Q_N, such as $Q = [Q_1, Q_2, ... Q_N]$,which represents the corresponding probability of our short video feature. In order to conform to the probability distribution, a softmax layer is added. After such processing, the P (Q is represented by P after processing) vector sums to 1.And finally, the result of our video feature fusion is (4), and our loss function is shown in formula (5) where p (x_i) represents the true probability distribution and q (x_i) represents the predicted probability distribution.

$$F_v = \sum_{i=1}^{N} P_i V_i \tag{4}$$

$$H(p,q) = -\sum_{i=1}^{N} p\left(x_i\right) \log\left(q\left(x_i\right)\right) \tag{5}$$

The cross-entropy loss value is obtained through the forward propagation of the neural network, and then the back-propagation of the neural network is used to adjust the weights to obtain the probability distribution of a single video feature. In this way, we have realized multi-modal video feature fusion based on Turing Neural Machine read heads process.

IV. Experiment

A. Experimental Dataset

The dataset used for engagement intensity prediction of DAiSEE which is collected from 112 Asian subjects (32 female and 80 male) who are participating in online lessons in a real environment in total. Their ages vary from 18 to 30 years. Each of 9068 video clips is 10 seconds. Video engagement tags are divided into four levels based on student participation from low to high, which are represented by {0,1,2,3}. These video clips were recorded under three lighting conditions, including bright, dark, and mild. Video recording locations include: dormitory, laboratory and library. The database simulates the real environment when collecting data, and to ensure that the data is closer to the real situation, it records the attendance status of the subjects before training. The data set is divided into training set, validation set and test set, the ratio is 60:20:20.

B. Experimental Setup

For the model configuration, our experiment mainly includes two parts: feature extraction and feature fusion. Our experimental framework is implemented using tensorflow. Feature extraction part: using OpenFace 2.0 to extract features, the obtained GAP [15] features are 117 dimensions, of which the head pose information is 22 dimensions, the eye gaze information is 26 dimensions, and the facial action unit information is 69 dimensions. The body posture is extracted using C3D network of which the information is 60 dimensions, and the batch of each experiment is 8. Rectified linear unit (ReLU) is used as the activation function, and a dropout layer with the dropout rate of 0.6 is applied to avoid the overfitting situation. We save the model after 100000 trainings to facilitate subsequent feature fusion. Feature fusion part: we use two fully connected layers to achieve the fusion of multiple short video features. The two fully connected layers are connected with size of 1024 and 128 respectively .The initial learning rate is set to 0.01 and multiplied by 0.1 every 20 epochs.

C. Experimental Results and Analysis

TABLE I: Accuracy of Engagement Prediction by Different Features and Method

Method	Accuracy (Validation)	Accuracy (Test)
Traditional method	57.5%	58.1%
Our method	60.2%	61.3%
GAP	58.3%	58.9%
C3D	56.2%	57.8%

The method we proposed first train on the training set of the DAiSEE database, and then evaluate the classification accuracy on the test and validation dataset. At the same time, as a comparison, we also tested the experimental accuracy of video feature fusion that is commonly used for summation and averaging. The results can be found in Table I. Through comparison we can find that our proposed model is superior to the most commonly used model of feature fusion in the past. In order to make the experiment richer, we also performed two sets of comparative experiments. The extracted GAP [15] features and C3D body posture features were input into our model for video feature fusion to evaluate the accuracy of student engagement respectively. The experimental results are also listed in Table I. Through comparative experiments, it is found that the use of multiple

angle features to judge the accuracy of students' engagement in learning is better than that of a single gesture feature. Through multiple groups of comparative experiments, we can conclude that our proposed multi-modal video feature fusion method can obtain a more accurate judgment for predicting students' engagement in online learning.

V. Conclusion and Future Work

Predicting student engagement in an online learning environment is a challenging task. In order to make a more accurate judgment of student learning participation, in this paper, we propose a method of video feature fusion based on the idea of Neural Turing Machine read heads. From the perspective of multiple features, we get the fusion of multiple short video features by using two fully connected layers. It can be found from Table I that our method can more accurately predict the student's learning participation than the traditional method of weighted summation and averaging.

In future work, we will continue to study which of the features could play a greater role in judging students' engagement in multi-modal features, and are more related to their accuracy. At the same time, we will consider adding an LSTM network to our model in order to better grasp the time signal of the input information.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Min Xu and Xiaoyang Ma conceived of the study. Xiaoyang Ma and Yao Dong analyzed the data, conducted the research. Xiaoyang Ma wrote the paper. Min Xu and Zhong Sun revised the paper. All authors had approved the final version.

References

[1] R. G. A. Dhall, A. Kaur, and T. Gedeon, *EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction*, 2018.

[2] T. Thales, W. Michel, and P. Rik, "Emotion-induced engagement in internet video advertisements," *Journal of Marketing Research*, vol. 49, no. 2, pp. 144-159.

[3] Z. Zainuddin *et al*., "The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system," *Computers & Education*, 2020, p. 145.

[4] C. Cheng *et al*., "An ensemble model using face and body tracking for engagement detection," in *Proc. the 2018 on International Conference on Multimodal Interaction*, ACM, 2018.

[5] A. Jalal and M. Mahmood, "Students' behavior mining in e-learning environment using cognitive processes with information technologies," *Education and Information Technologies*, 2019.

[6] A. Graves, G. Wayne, and I. Danihelka, *Neural Turing Machines*, October 2014.

[7] W. Z. Qiao and X. J. Bi, "Ternary-task convolutional bidirectional neural turing machine for assessment of EEG-based cognitive workload," *Biomedical Signal Processing and Control*, 2020, p. 57.

[8] Y. C. Lin *et al*., "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, 2014, vol. 5, no. 1, pp. 86-98.

[9] T. Chinchu, N. Nair, and D. B. Jayagopi, "Predicting engagement intensity in the wild using temporal convolutional network," in *Proc.* the 2018 on I*nternational Conference on Multimodal Interaction*, ACM, 2018.

[10] W. Jacob *et al*., "The faces of engagement: automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol 5, no. 1, pp. 86-98.

[11] J. Yang *et al*., "Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction," in *Proc. the 2018 on International Conference on Multimodal Interaction*, pp. 594-598, 2018.

[12] D. Tran, L. Bourdev, R. Fergus *et al*., *Learning Spatiotemporal Features with 3D Convolutional Networks*, 2014.

[13] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, 2017, pp. 15-28.

[14] T. Baltrusaitis *et al*., *OpenFace 2.0: Facial Behavior Analysis Toolkit*, 2018, pp. 59-66.

[15] X. S. Niu *et al*., "Automatic engagement prediction with GAP feature," in *Proc. the 2018 on International Conference on Multimodal Interaction*, ACM, 2018.

[16] W. B. Park *et al*., *Affective Tutors: Automatic Detection of and Response to Student Emotion. Advances in Intelligent Tutoring Systems*, Springer Berlin Heidelberg, 2010.

[17] A. C. Cruz, B. Bhanu, and N. S. Thakoor, "Vision and attention theory based sampling for continuous facial emotion recognition," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 418-431, 2014.

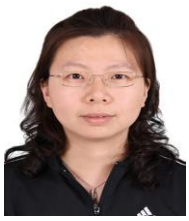
[18] J. Han, *et al*., "Towards emotion-sensitive learning cognitive state analysis of big data in education: Deep learning-based facial expression analysis using ordinal information," *Computing*, 2019, no. 3.

[19] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective posture and body motion to detect engagement with a game companion," *HRI*, 2011.

[20] Y. Liu *et al*., "Student engagement study based on multi-cue detection and recognition in an intelligent learning environment," *Multimedia Tools and Applications*, 2018.

Copyright © 2021 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).

Xiaoyang Ma is currently a master student in the College of Information Engineering at Capital Normal University, Beijing, China. She was born in Hebei province in 1994. Her research interests include computer vision and pattern recognition.



Min Xu is an associate professor with the College of Information Engineering, Capital Normal University, Beijing, China, who was born in Jiangxi province in 1978. She received her Ph.D. from Renmin University in 2012. Her research interests include computer vision, pattern recognition, and machine learning.

Yao Dong is currently a master student in the College of Information Engineering at Capital Normal University, Beijing, China. She was born in Zhejiang Province in 1996. Her research interests include computer vision and pattern recognition.

Zhong Sun is professor in the College of Informational Engineering at Capital Normal University, Beijing, China, who was born in Liaoning province in 1973. She received her Ph.D. from Beijing Normal University in 2008. Her research interests include mobile learning and technology enhanced teacher professional development.