IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 5, 2019, accepted October 5, 2019, date of publication October 11, 2019, date of current version October 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2947091

Recognizing Multidimensional Engagement of E-Learners Based on Multi-Channel Data in E-Learning Environment

JIA YUE[1], FENG TIAN[1], (Member, IEEE), KUO-MIN CHAO[2], (Member, IEEE), NAZARAF SHAH[2], LONGZHUANG LI[3], YAN CHEN[1], AND QINGHUA ZHENG[1], (Member, IEEE)

[1]MOEKLINNS Laboratory, Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[2]Department of Computing, Coventry University, Coventry CV12JH, U.K.
[3]Department of Computing Sciences, Texas A&M University-Corpus Christi, Corpus Christi TX 78412, USA

Corresponding author: Feng Tian (fengtian@mail.xjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1004500, in part by the National Science Foundation of China under Grant 61877048, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2019JM-458, in part by the Innovative Research Group of the National Natural Science Foundation of China under Grant 61721002, in part by the Innovation Research Team of Ministry of Education under Grant (IRT_17R86), and in part by the project of China Knowledge Centre for Engineering Science and Technology.

ABSTRACT "Lack of supervision" is a particularly challenging problem in E-learning or distance learning environments. A wide range of research efforts and technologies have been explored to alleviate its impact by monitoring students' engagement, such as emotion or learning behaviors. However, the current research still lacks multi-dimensional computational measures for analyzing learner's engagement from the interactions that occur in digital learning environment. In this paper, we propose an integrated framework to identify learning engagement from three facets: affect, behavior and cognitive state, which are conveyed by learner's facial expressions, eye movement behaviors and the overall performance during short video learning session. To recognize the three states of learners, three channel data is recorded: 1) video/image sequence captured by camera; 2) eye movement information from a non-intrusive and cost-effective eye tracker; and 3) click stream data from mouse. Based on these modalities, a multi-channel data fusion strategy is designed that concatenates time series features of three channels in the same time segment to predict course learning performance. We also presented a new method to make the self-reported annotations more reliable without using external observers' verification. To validate the approach and methods, 46 participants were invited to attend a representative on-line course that consists of short videos in our designed learning environment. The results demonstrated the effectiveness of the proposed framework and methods in monitoring learning engagement. More importantly, a prototype system was developed to detect learner's emotional and eye behavioral engagement in real-time as well as predict the learning performance of learners after they had completed each short video course.

INDEX TERMS E-learning, engagement recognition, multi-channel data fusion, learning performance prediction.

I. INTRODUCTION

As compared to the traditional classroom, the learners' emotions, lack of concentration or motivation can not be monitored dynamically or in real-time in digital learning environments [1]. Although e-learning platforms (e.g., Coursera) indeed provide portable learning ways and abundant quality courses across the globe [2]–[4], "high enrollment and low completion rate" phenomenon still exists in this style of learning [5], [6]. Prior studies have indicated that completion rates on these platforms are as low as 7-11%, and some of the major reasons accounted for this phenomenon are low motivation among the learners and low perceived value for the course [7], [8]. Therefore, it is imperative to detect the

The associate editor coordinating the review of this manuscript and approving it for publication was Kemal Polat.

learning engagement and understand their dynamic learning process in order to deliver timely and individualized feedback.

Engagement is a multidimensional construct that combines diverse psychological constructs such as thoughts, perceptions, feelings, and attitudes [9] from the field of psychology. Research in areas such as education data mining, multimodal learning, cognitive science, psychological and many other fields, has made significant advances in learning analytics, which has shown considerable promise to supervise learner's engagement for improving learning efficiency in e-learning environments [10], [11]. Most of them focused on perceiving or detecting single dimension of Learning Engagement. Some research efforts mainly focus on learner's affective state (e.g., boredom, confusion, frustration and anxiety [12]), or learner's attention (e.g., low, high and normal level [13]), or learning performance assessment through quizzes and assignments [14]. Our team proposed a computational framework of recognizing engagement of e-learners based on multi-channel data granted by Natural Science Foundation of China, Oct, 2014 (grant No. 61472315), where e-learner's facial expression, head pose and mouse behaviors are detected and classified based on multi-channel data in a middle range performance.

As a popular way of on-line learning, the majority of MOOC courses are designed as short videos of knowledge units, and the duration of each video ranges from about five to ten minutes [15], [16]. However, studies on learning engagement analysis from the perspective of each short video performance have rarely been explored. Hence, they lack an integrated framework to model students' learning from multiple aspects and dimensions, especially the learning performance on knowledge units after they had accomplished each learning task. On the other hand, the rapid development in sensory technology has enabled researchers and practitioners to push the boundaries of learning engagement detection and its analysis by investigating various machine-readable signals or behaviors, such as electroencephalogram (EEG) signal, physiological signal, electrodermal activity, facial expression, gaze, keystroke and mouse movement [17]. In order to obtain high quality data generated from the learning activities, intrusive or wearable devices such as electrode headset [18], wristband [19], or costly sophisticated eye tracker [20] have to be adopted. The adoption of these devices may cause inconvenience or discomfort to learners during data acquisition process. Meanwhile, synchronization of different frequency signals from multi-sources is a nontrivial task, and only a paucity of research exits on online education that focuses on integration of multi sensory data. Thirdly, the limitations of current methods for labeling learner's engagement include annotation biases from observers' views or subjects themselves, and laborious work to check and verify the reliability of these labels. For example, the inconsistent understanding and comprehension on labeling standards varies from person to person, such that some irregular or mistaken labels may occur [21]. At the same time, there is no open-source multimodal learning dataset available in such e-learning environment. The majority of multimodal datasets consist of visual source (e.g., Youtube, Facebook), audio and text, which are not suitable for e-learning analysis [22].

In order to address these problems, we propose an integrated framework to model students' learning engagement in three different facets and dimensions. The proposed framework detects learner's emotional and behavioral state in real-time and predicts their cognitive state from watching short-video episodes. In our study, three components of engagement are expressed by learner's emotions, eye gaze behaviors and knowledge-unit-based course learning performance. Specifically, the emotions and eye behaviors exhibit the objectivity of student's state and they are easy to observe, while the cognitive state emphasizes the learner's mental and psychological state, such as understanding, self-regulation or meta-cognition. To recognize these states, we employ multichannel sensory data: video streams captured by a camera, eye movement information captured by a low cost eye tracker named Tobii Eye Tracker 4C and mouse dynamic log from a standard mouse. Obviously, the devices we choose are non-contacted and cheap, to provide a more spontaneous learning environment for participants. We also collect the relative essence signals from the eye tracker fixed at the bottom of the screen. The rationale behind selection of these three channels is to achieve comprehensive interaction information through visual, eye movement behaviors and mouse dynamic during learning. Particularly, we design a fusion method to combine data obtained from three different channels to predict the student's cognitive state (The performance on each short video). More importantly, a method that integrates learner's prior subject knowledge level, quiz scores and self-assessment data is utilized to raise the reliability of labels, without requiring other observers' intervention to eliminate the label biases. Lastly, we develop an intelligent online learning prototype system, and carry out experiments involving our invited participants, to validate our proposed framework and method. Our major contributions in this paper are as follows.

1) We propose an integrated computational framework to characterize and quantify multi-dimensional engagement in e-learning environment from three facets: affect, behavior and cognition, which provides a new insight into computer-based learning analysis. In this framework, three different channel data (video, eye movement and mouse dynamic) is captured through low cost devices without using intrusive nor wearable equipment.
2) Through fine-tuning parameters by transfer learning, we improve our facial expression recognition model accuracy with insufficient number of Asian images. Moreover, a feature-level fusion method combining multi-channel features is designed to predict learner's cognitive performance.
3) We develop a computer-based learning prototype system to monitor learner's emotional, eye movement and

cognitive state in e-learning environment to evaluate our proposed approach.

The remainder of this paper is organized as follows. Section II reviews the related literature. Section III describes our learning engagement recognition framework in e-learning environment, and introduces the database we have employed. This section also presents the methods for features extraction, selection and fusion from the multi-source channel data. Section IV discusses the experimental setup and experiments involving our invited students to recognize learner's affect, eye behavior and cognition. The last subsection of Section IV provides insight into monitoring learner's facial expressions, eye movement behaviors and course learning performance. Section V concludes this paper and points out future directions for further research.

II. RELATED WORKS

Recent years have witnessed an increased research interest in the area of learning analysis in the context of computer-based learning environment. To date, various research works have been carried out on modeling students learning [23], [24]. In the following subsection, we review related works from two main aspects.

A. LEARNING ENGAGEMENT

There is no consensus on definition and taxonomy of learning engagement in academic communities. The emotions and facial expressions are often considered as the engagement in the majority of current research. These literature works suggest that there is a strong evidence that emotional state of student is easier to be perceived, which may have a strong impact on their learning [25], [26]. For example, Landowska suggested affective learning and affective computing can be combined to assess and improve effectiveness of the education process [27]. Magdin *et al.* [28] drew on Ekman's definition of six emotions to investigate learning effect of students and determine the kind of emotions that students have in a test to help them to deal with stress, anger or disgust. Leony *et al.* [12] infer four kinds of more complex emotions (frustration, confusion, boredom and happiness) in the MOOC platform, based on the four corresponding detection model. The similar taxonomy of learning emotions can be found in [29]–[31]. Some researchers have shown that learner's attention or motivation can be utilized to identify learning engagement in e-learning environment. Narayanan *et al.* [32] studied different attention patterns exhibiting in e-learning classroom, where teacher and students are not geographically separated but connected. Wang [33] detected situation where learner's attention decreases during learning process and suggested desirable/effective feedback. Brandon *et al.* [34] focused on estimating student engagement in distance learning corpus containing unstructured learning sessions. Hussain *et al.* [35] aimed at the lack of student motivation problem and they evaluated student's interaction activities on virtual learning environment. While, some other researchers mainly concentrated on predicting course performance and learning outcomes. Phan *et al.* [36] investigated potential relationships between students' course performance and degree of involvement, their motives of participation as well as their subject matter prior knowledge. Guo and Wu [37] combine students' performance data on homework problems with the results obtained in the first stage (analyzing student learning activities within a chapter, such as video-watching click stream, page-view records and forum interactions, extract interpretive quantities to predict the probability that a student has mastered the knowledge of that specific chapter), and built sequential models to accurately assess student learning outcomes. Zhange *et al.* [38] introduce 19 behavior indicators in the online learning platform, and proposed a student performance prediction model combined with the whole learning process.

However, most of these research efforts only investigate single facet or dimension of the learning engagement, and do not fully study the whole learning process. In addition, few investigators attempted to analyze student's state or performance from a short video, which lasts for 5-10 minutes and usually contains one knowledge unit. In this paper, we investigate three elements of engagement encompassing learner's affect, eye behaviors and cognition, to reflect different facet and dimension of learning engagement, including physiological, physical and mental state.

B. STATE-OF-THE-ART METHODS AND TECHNIQUES ON DETECTING ENGAGEMENT

Current methods for detecting and modeling engagement can be categorized into three types: self-report questionnaires [39], learning-logs-based data mining methods [40], [41] and sensors-based techniques [42]. Traditionally, the questionnaire with several questions is a most straightforward way to assess students' engagement. However, the bias in results is an inevitable limitation as standards may vary from one learner to other. There is a growing interest in employing machine learning methods to explore learner's log emerged from the interactions with his or her learning environment. For example, Tian *et al.* [43] recognized and regulated the e-learners' emotion from interactive Chinese text. They compared Support Vector Machines (SVM), Naive Bayes, LogitBoost, Bagging, MultiClass Classifier, RBF network and J48 machine learning algorithms to classifying the emotions. Hershkovitz and Nachmias [44] focused on two types of analysis: 1) investigating a learner's activities, to learn about her or his learning process, and 2) examining the activities of a large group of learners, in order to develop a log-based motivation measure. However, these machine learning methods are not able to extract intuitive cues for monitoring student's visible learning behaviors, such as emotions or body movements. Fortunately, great advances in sensory techniques have made it possible to study complex human-machine interactions. S. Saha proposed a system to classify engagement based on body gesture using

Kinect sensor [45], Booth *et al.* [34] recorded EEG data of students watching online lecture videos and used it to predict engagement rated by human annotators. Other sensory channels devices like Tobii T60 [46], keyboard [47] and mouse [48] have leveraged the detection of learner's affect. Recently, researchers believe that multi-sensor data fusion methodology has ability to increase the accuracy and reliability of the estimates [49], which shows the significance and feasibility of multi-channel data fusion methodology in diverse research fields. Gogia *et al.* [50] used facial features and brain signal of user captured from a camera and a Brain Computer Interfacer (BCI) module. Di Mitri *et al.* [19] proposed an approach based on multimodal data such as heart rate, step count, weather condition and learning activity that can be used to predict learning performance in self-regulated learning settings. They employed a biosensor called "Fitbit HR wristband", to enhance learning effectiveness. However, some sensors devices adopted in the above literature are either intrusive or wearable, such as BCI module, wristband, and costly sophisticated eye tracker. Though Li *et al.* [13] proposed a low cost multimodal fusion framework that only used webcam and mouse, they focused only on subject's attention while reading an article.

In addition to the field of education and e-learning, the engagement is also investigated in other research. Yu *et al.* [51] proposed a multilevel structure based on coupled hidden Markov models (HMM) to estimate engagement levels in continuous natural speech. The first level is comprised of SVM-based classifiers that recognize emotional states, which could be (e.g.) discrete emotion types or arousal/valence levels. A high-level HMM then uses these emotional states as input, estimating users' engagement in conversation by decoding the internal states of the HMM. Rich *et al.* [52] developed and implemented an initial computational model for recognizing engagement between a human and a humanoid robot, based on a study of the engagement process between humans.

III. OUR FRAMEWORK AND MODEL

Our proposed framework is shown in Fig. 1. Three channels data are input, including video stream sequence, eye movement learning logs and mouse dynamics, which are represented in different colors, and the output of this framework is three elements of learner's Engagement: emotional state, eye behavioral state and cognitive state. In this paper, the learner's emotional state set is comprised of seven basic facial expressions: joy, disgust, sadness, surprise, fear, anger and neural [53]. The probability values of facial expression will be predicted at each learning moment real-time, and the maximum one will be output as learner's emotion state. The eye movement behavioral state set consists of watching video, reading teaching materials and typing notes, which are also tracked in real-time. The output of eye movement behavioral state will be predicted through the classifiers, such as Random Forest. The cognitive state is represented as a value ranging from 0 to 1, and predicted by our multi-features fusion model,

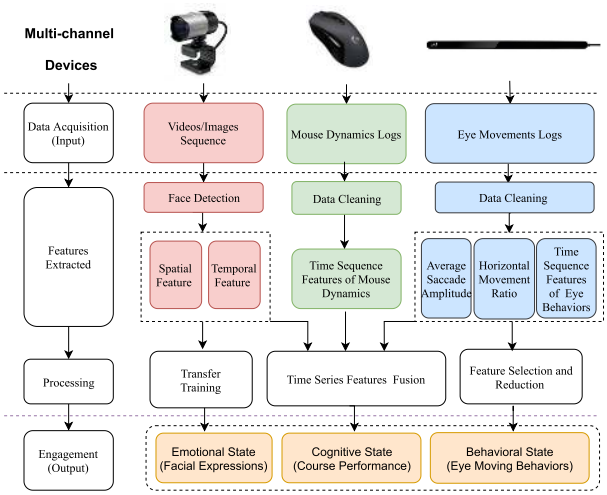


FIGURE 1. Three components of learning engagement recognition framework.

which indicates the course performance. The higher the value is, the better performance the learner achieves. The performance value conveys a subjective level facet of engagement. Note that the course performance value is predicted after finishing each short course video.

Next, we describe the dataset we employed, the emotions and eye gaze behaviors recognition models, three modality feature fusion strategy and a method to label cognitive state reliably.

A. EMPLOYED DATABASE

We use two available image datasets to train facial expression model: ImageNet database [54] and USTC-NVIE (Natural Visible and Infrared facial Expression) database [55]. The first database is open and free, and the second one is obtained for research purpose after receiving the author's consent. In our study, ImageNet database is used to pre-train models, and USTC-NVIE dataset is used for learning Asian face features. Meanwhile, we use an online eye learning behavioral database built by our research group earlier to train our eye gaze model.

USTC-NVIE is an Asian face database which was constructed by The Key Laboratory of Computing and Communication Software of Anhui Province(CCSL). This dataset consists of a natural visible and infrared facial expression database of 70GB size including static image and facial expression image sequence. It also contains both spontaneous and posed expressions of more than 100 subjects with or without glasses, recorded simultaneously by a visible and an infrared thermal camera, with illumination provided from three different directions. Finally, we obtained 22906 facial images which are most relevant to our study.

The eye learning behavioral database was recorded using Tobii Eye Tracker 4C commercial equipment. Twenty-two subjects were asked to watch a short course video, read teaching material and type some notes or comments, and approximately 120 minutes worth of data was collected.

For each task carried out by a subject, we removed 10 seconds at the beginning and end of the recording fragment as the learner's eye gaze or fixations may be out of the screen scope during that time. Additionally, some outliers will be detected and cleaned in the data process period. The emergence of outliers mainly caused by incorrect use of eye tracker. For example, when the distance from the participant's eyes to the eye tracker is too far or too close, none eyes will be detected by eye tracker and null value will be marked. Another kind of outlier is that the horizontal and vertical coordinates of gaze point are beyond the current resolution of monitor, which are caused by participant's eye gaze or fixations are not on the screen, or the hardware instrumental errors.

B. FACIAL FEATURES EXTRACTION AND EMOTIONAL STATE RECOGNITION

A change in facial expression goes through three stages: Onset, Apex and Offset [56]. During this process, the intensity of facial expression gradually increases from a neutral to peak, and then gradually decrease to neutral again. Inspired by this, we take face image sequences instead of static image as input in order to take an integrated consideration of spatial and temporal features of the image. Therefore, we adopt a two-step features extraction strategy: 1) spatial image characteristics of the representative expression-state frames are learned using Convolutional Neural Network (CNN), 2) temporal characteristics of the spatial feature of the facial expression are learned using Long Short Term Memory(LSTM). The feature extraction procedure and facial expression model can be seen in Fig. 2. In spatial feature extraction period, we follow the idea of transfer Learning that pre-train a CNN model (VGG16 with 16 layers [57] or Inception-ReNet-V2 with 572 layers [58]) on ImageNet database firstly, and fine-tune the pre-trained convolutional layer's parameters on USTC-NVIE database to learn more about Asian face features, which improves the generalization ability and recognition accuracy. Then, we extract the spatial features of each frame from input image sequences by the pre-trained models and learn temporal features of these spatial feature sequences by LSTM. Lastly, seven basic face expressions are predicted through Softmax Layer using these features.

In this paper, we take facial expressions to represent learner's emotional state. As the recognition targets are only 7 classes, we need to modify network structure of both CNN model (VGG16 and Inception-ReNet-V2). The specific steps taken are as follows:

1) Remove full connection layer of the two CNN models, and set only one layer with 512 cells in VGG16 or 1024 cells in Inception-ReNet-V2.
2) Adjust the output number of classes at Softmax Layer
3) Add a Global Average Pooling Layer [59] behind last convolutional layer, in order to transform a four-dimensional tensor to a two-dimensional tensor, and cut down the size of the training parameters as well.

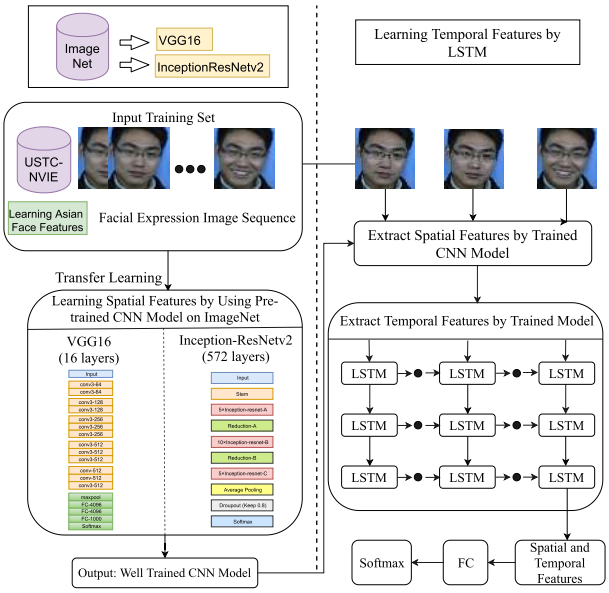


FIGURE 2. Spatial and temporal features extraction.

TABLE 1. Modified network of models.

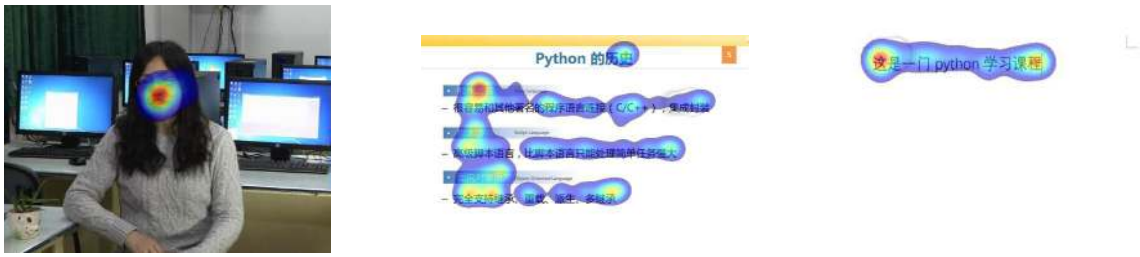
	Layer	Input	Output
Input	Input	N× (S × S × 3)	N × (S × S × 3)
Spatial features	CNN layer	N × (S × S × 3)	N × (M × 1)
temporal features	LSTM layer	N × (M × 1)	(512 × 1)
Dropout	Dropout	Dropout(0.2)	Dropout(0.2)
Output	Softmax	(512)	(7)

4) Add a dropout layer with a parameter value of 0.2, to reduce the risk of over-fitting while training the models.

The modified parameters of networks are presented in Table 1. Where N is the number of facial image sequence frames, the value of S is 244 and M is 512 when choosing VGG16 model, while the value of S and M is 299 and 2048 when choosing Inception-ReNet-V2.

C. TIME SERIES FEATURES EXTRACTION AND EYE BEHAVIORAL STATE RECOGNITION

In a real-world E-learning scenario, student's eye movements are focused on targeted screen during most of the learning time, and different learning behaviors present different eye movement tracks, gaze and fixation [60]. Take a learning instance in our developed system for example, as shown in Fig. 3a, when watching a video course, learner's gaze may focus on the teacher at the beginning of course, then moves to other points with the change in teaching content. Fig. 3a demonstrates the watching eye movement behavior. While Fig. 3b and 3c illustrate the eye movement behaviors of reading teaching materials and typing notes, respectively. As can be seen Fig. 3b and 3c, the student's eye gaze moves horizontally with the text while reading and typing, while the density of the former shows more sparse than that of the later. Because the eye gaze moves much more slowly while typing



(a) (b) (c)

FIGURE 3. Eye tracking of three eyes movement behavior during learning. (a)Eye gaze and fixation of watching behaviors. (b)Eye gaze and fixation of reading behaviors. (c)Eye gaze and fixation of typing behaviors.

notes. In this paper, we mainly concentrate on these three eye movement behaviors of learning, i.e. watching, reading and typing, to convey learner's eye gaze behavioral state.

Through analyzing the characteristic of the three eye movement behaviors of learning, it is obvious that the eye movement trajectories of reading and typing learning behaviors are analogous, which shows large horizontal moving distance and short vertical moving distance, while the eye movement speed of reading is faster than that of the typing. On the other hand, watching eye movement is quite distinct with other eye movement behaviors, which passively follows teacher's instructions. Liu *et al.* [60] proposed two metrics (Average Saccade Amplitude and Horizontal Movement Ratio) to indicate the three gaze movement of online learners. We followed the former feature and adapt the Horizontal Movement Ratio (denoted as R_h) feature, to better classify learner's eye behaviors. The first one is represented by $S_{average} = \frac{D_{total}}{N-1}$. Where, $S_{average}$ represents the Average Distance of Eye Gaze Movement, D_{total} is the total moving distance during a period of time, and N is the number of eye gaze points appeared in that time. And The second one is defined by

$$R_h = \frac{\sum_{n-1}^{n-N-1} 1\{4V_n < H_n\}}{N-1} \tag{1}$$

Due to the influence of eye blinks, it is impossible to keep our eye movement absolutely parallel while learning. Therefore, we deem the eye move horizontally as long as $4V_n < H_n$. Where, V_n, H_n indicates different vertical and horizontal values between two adjacent gaze points, respectively. If not, the movement should be considered as a vertical moving in that time.

What's more, we extract 40 general features from eye log data time series, involving statistics, wavelet and Fourier Transform. These time series features are come from tsfresh, which is a open and free Python package to process time series data. Table 2 give a concise description, more details can be found in [61]. Particularly, these time series features are applied in mouse dynamic logs analysis, as with the same time series characteristic of eye movement logs.

However, the log data time series captured by eye tracker or mouse often contains a large amount of noise and redundant data. So, the extracted features from logs may be sensitive or highly irrelevant. In order to better recognize learner's eye behavioral state and to improve robustness of the models, we apply Kolmogorov-Smirnov (KS) Test [62] and Benjamini-Yekutieli (BY) procedure [63] to filter relevant and robust features. Since the KS Test is only suitable for binary classification or regression problems, we firstly transform our multiple eye behavioral states classification into three binary classification problems: P_W, P_R, P_T, as described in following representation.

$$P_W = \{(Watching), (Reading, Typing)\}$$
$$P_R = \{(Reading), (Watching, Typing)\}$$
$$P_T = \{(Typing), (Watching, Reading)\}$$

TABLE 2. Description of general time series features.

	Feature
1	Absolute energy of time series which is the sum over squared values.
2	Sum over the absolute value of consecutive changes.
3	Calculates the value of an aggregation function.
4	Calculates a linear least-squares regression for values.
5	Implements a vectorized Approximate entropy algorithm.
6	Calculator fits the unconditional maximum likelihood.
7	Autocorrelation of the specified lag.
8	First bins the values of x into max_bins equidistant bins.
9	Measure of non linearity in the time series.
10	Calculates the number of peaks of at least support n in time series.
11	First fixes a corridor given by the quantiles.
12	Estimate time series complexity through distance between features.
13	Number of values that are higher than the mean of time series.
14	Number of values that are lower than the mean of time series.
15	Calculates a Continuous wavelet transform for the Ricker wavelet.
16	Mean over the differences between subsequent time series values.
17	Spectral centroid, variance and skew of the absolute FFT spectrum.
18	Fourier coefficients of the one-dimensional DFT.
19	The first location of the maximum value of time series.
20	The first location of the minimal value of time series.
21	Coefficients of polynomial function.
22	Checks if any value in time series occurs more than once.
23	Checks if the maximum value is observed more than once.
24	Checks if the minimal value is observed more than once.
25	Relative last location of the maximum value of time series.
26	Last location of the minimal value of time series.
27	Length of time series vector.
28	Length of the longest subsequence that is bigger than the mean.
29	Length of the longest subsequence that is smaller than the mean.
30	Highest value of the time series.
31	Mean of time series.
32	Mean over the absolute differences between subsequent time series.
33	Lowest value of the time series.
34	Median value of the time series.
35	Difference between max and min of time series.
36	Standard deviation of the time series.
37	Sum of all data points present in the time series more than once.
38	Sum over the time series values.
39	Variance of the time series.
40	Denoting if the variance is greater than its standard deviation.

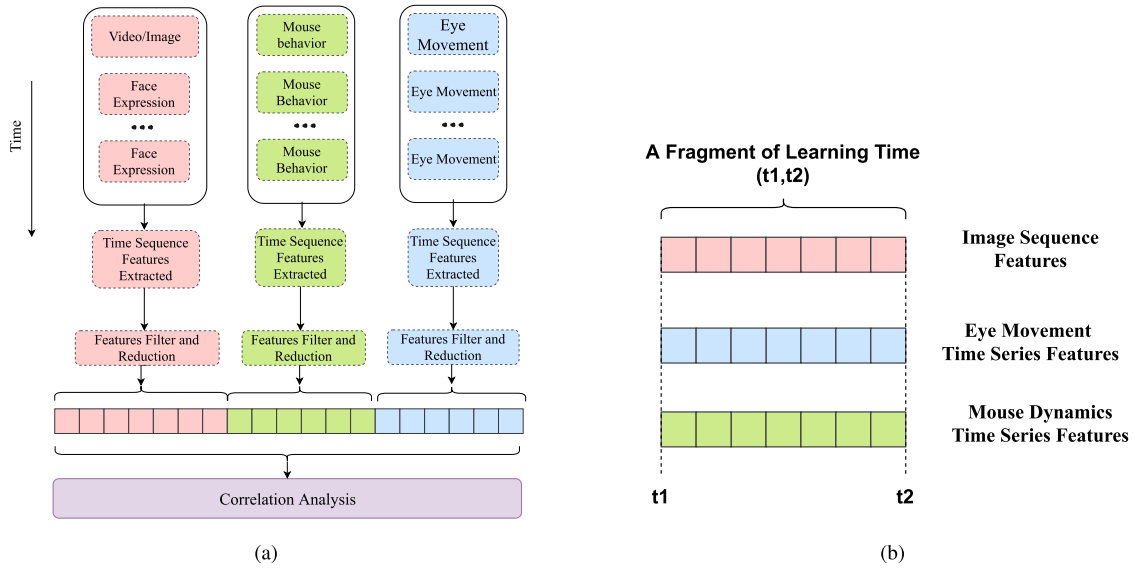


FIGURE 4. Multi-source channel features fusion framework. (a)Features fusion strategy. (b)Alignment of three channels features.

Then, KS test and BY procedure are carried out on each feature set of binary classification problem (P) in turn [64], to obtain filtered features:

$$F_W = filter(feature(P_W))$$
$$F_R = filter(feature(P_R))$$
$$F_T = filter(feature(P_T))$$

The functions *feature*() and *filter*() of KS test and BY procedure are used to filter and select features, and the integrated features of eye behaviors can be represented as $F = F_W \cup F_R \cup F_T$. Finally, the dimensions of these features will be reduced using Principal Components Analysis (PCA) [65], to enhance the generalization ability.

D. COGNITIVE STATE PREDICTION

The cognitive state refers to learner's investment in learning task, such as how they allocate effort toward learning, and their understanding and mastery of the material [21]. Unlike emotional and eye behavioral state, cognitive state reflects distinct facets of student's learning engagement, which emphasizes learner's mental and psychological state, which is hard to observe and label. In this study, we investigate learner's cognitive state through their performance on short video course, and from the perspective of knowledge unit mastery degree. We also use a specific value ranging from 0 to 1, to represent learner's performance in each short video. In what follows, we will describe our designed cognitive state labeling method and multi-sensory fusion strategy.

1) MULTI-SOURCE FEATURES FUSION METHOD

Fig. 4a illustrates the features fusion procedure of the three channel data captured at three different channels. Firstly, learner's facial expression, eye gaze movement and mouse dynamics time series features within the same time interval are extracted separately. To make it clear, the alignment of time series features of the multi-channel data is shown in Fig. 4b These time series features are described in Table 2. Then, we apply KS Test and BY procedure to sift most relevant features, and use PCA method to reduce the features' dimensions, because the combination of sifted features often reach thousands of dimensions. Thirdly, we concatenate the processed features of three channels, to form an integrated fused vector.

2) LEARNER'S COGNITIVE STATE LABELING

Generally, the annotations of supervised learning method require the input from learners themselves or external observers. These two approaches have access to different types of information and may be influenced by different biases. In this work, we follow the self-annotated approach, since it is difficult for external observers to label a learner's performance from learning a short session of videos. However, this kind of self-report suffers from the problem when respondents aim to appear admirable to others and when they inflate responses to preserve their own self-esteem [21]. Additionally, different interpretations of rating standards may occur among learners. In order to overcome this limitation, we propose to combine the learner's prior subject knowledge level, self-assessment and quiz scores, to increase the reliability of learners' self-report data. Fig. 5. elucidates the label amendment method, and the steps taken are as follows.

1) Normalize the learners' self-assessment data to the interval [0,1], to unify the self-report standards and rescale the scores on the ratings. We define normalization formula as follow:

$$S_c = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (2)$$

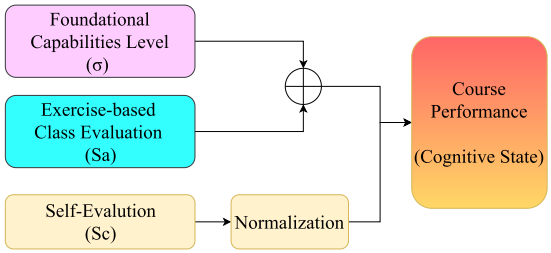


FIGURE 5. Method to amend Self-assessment labels.

where, S denotes each learner's self-assessment value, S_{min}, S_{max} are the minimum and maximum of S.

2) To amend the irregular or mistaken self-assessment value (S), we refer to each learner's prior knowledge level and quiz score, to recalculate a new label value. We consider that the higher subject knowledge foundation level of a learner has, the less his or her quiz scores contribute to recognize his learning performance. As a result, we take the prior knowledge level as an impact factor for quiz scores, and denote it as σ. The prior subject knowledge level is categorized into l grades, so $\sigma \in \{1, 2, \ldots, l-1, l\}$. The formula to calculate final label value (denoted as S_n) is

$$S_n = \alpha S_c + \beta \frac{S_a}{\sigma}, \quad \alpha + \beta = 1 \tag{3}$$

where, α and β are the weights of each learner's self-assessment (S_c) and quiz score (S_a), and the sum of them are equal to 1. Note that S_a should be normalized to [0,1]. From Fig. 5, we can see the final cognitive state of the participants is determined by their self-assessment and quiz score. When rating, each learner has its own mind on allocating the two values of α and β. For example, one learner may think the weights of two parts are the same, i.e. both α and β are 0.5; another learner may think the self-assessment part is more important than quiz score part, i.e. α is 0.7 and β is 0.3. Note that the value 0.5 is just a default setting. The two values but will be changed by learners themselves. We assume the default value of both α and β are 0.5, and the values can be changed by learners themselves when rating.

IV. EXPERIMENTS

In this section, we report a set of experiments carried out to demonstrate the effectiveness of the proposed framework and methods. We will introduce the experimental setup followed by the design. The prototype system we developed is presented in the following subsections.

A. EXPERIMENT SETUPS AND DESIGN

As shown in Fig. 6, this experiment was carried out in an indoor Computer Lab environment. We utilize three external devices, a Microsoft LifeCam webcam located on the top of the computer screen, a Tobii Eye Tracker fixed at the bottom of the screen and a common wired mouse.

FIGURE 6. Experimental setups in lab environment.

TABLE 3. Parameters and specification of channel sample data.

Device source	Dimensions	Frequency	The sample frequency is fixed or not?(Y/N)
Image/Video	(640,480,3)	15Hz	Y
Eye movement	(3,1)	10Hz-90Hz	N
Mouse Dynamic	(5,1)	0.2Hz-60Hz	N

Meanwhile, we have developed two software tools to collect eye tracking logs and mouse dynamic logs. The parameters and specifications of sample data from three channels are listed in Table 3, in which, four dimensions of eye tracking logs record are represented as: <*event_type*, *x*, *y*, *timestamp*>. Where, *event_type* denotes the data type of the fixation, such as *BEGIN* and *DATA* or *END*, *x* and *y* are the coordinate of eye fixation, and the *timestamp* records the event time of each log. Similarly, the mouse dynamic logs are recorded as: <*message*, *wheel*, *x*, *y*, *window*, *timestamp*>, in which, *message* denotes the type of click event, *wheel* is the direction of wheeling, *window* refers to the current active window of cursor, *x* and *y* mean the coordinates of mouse cursor, and *timestamp* records the time of each mouse dynamic event.

We chose a representative MOOC course titled "Data Processing Using Python" with course videos, teaching materials and quizzes. The duration of these videos ranges from five to ten minutes, and each video mainly contains one knowledge unit. To better execute the experiments and meet our requirements, we developed a prototype computer-based learning environment, which is shown in Fig. 7. On clicking a video, the left hand side panel presents course videos, and the corresponding teaching materials emerges on the right hand side panel. Before starting course learning, the subjects were required to fill some information, encompassing name, major, sex, age and their prior knowledge level of Python language. In this system design, we removed the embedded or in-video quiz mechanism during learning. After accomplishing each short video learning, the learners were asked to self-assess their learning performance with a value ranging from 10 to 100. Lastly, subjects took a quiz to test their degree of mastery of learning knowledge units. The entire process took approximately 50 minutes for a participant to complete.

Overall, we invited 46 subjects to participate in Python course learning on our designed system. There were 32 males

TABLE 4. Comparison of four models on training time, execution time and memory size occupied.

Model	Training Time	Time on CPU (95% confidence interval)	Time on GPU (95% confidence interval)	Size
VGG16 without LSTM	5h	(3046.7, 3119.3)ms	(45.25, 46.19)ms	80MB
Inception-ResNetV2 without LSTM	12h	(13420.5, 13521.4)ms	(38.57, 39.11)ms	219MB
VGG16 with LSTM	2h	-	(56.69, 58.57)	(78+5)MB
Inception-ResNetV2 with LSTM	4h	-	(46.81, 47.85)ms	(217+9)MB

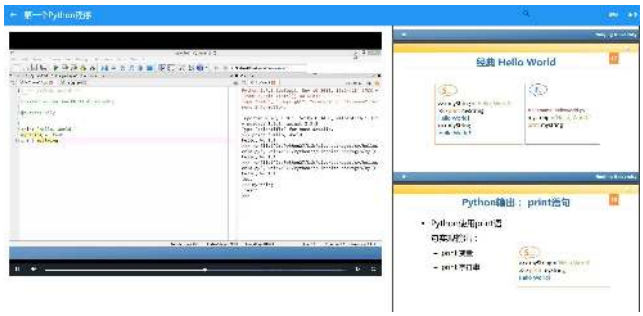


FIGURE 7. Prototype computer-based learning system course video and teaching material page.

and 14 females, with the age between 19 and 29. Moreover, the educational background distribution is: 33 subjects in undergraduates level, 11 subjects in graduates and 2 subjects in doctoral study. About two thirds of the subjects major in Computer Science and Technology. Finally, we collected a multimodal learning dataset with a total of 51GB, including learners' video streams, eye movement logs and mouse dynamics logs. After subsampling the collected data from these 46 subjects, we obtained 7224 learning performance instances with amended labels.

B. FACIAL EXPRESSION MODEL EVALUATION

We evaluated our emotional state recognition methods and models on ImageNet and USTC-NVIE databases. To select an optimal model for facial expression recognition, we compared four models, VGG16 without LSTM, Inception-ResNetV2 without LSTM, VGG16 with LSTM, and Inception-ResNetV2 with LSTM, from the following aspects: time need to train, execution or response time and occupied memory while running. To obtain a precise execution time on GPU and CPU, we repeated the measure ten times to calculate the 95% confidence interval ($z_{0.025} = 1.96$ from the Standard Normal Distribution Table [66]). The results of this evaluation are shown in Table 4. It can be seen that the average execution time of four models on CPU exceeds 3 seconds, while less than 60 milliseconds on GPU. Considering 15 frames per second of the sample rate from webcam, the execution time of model should be no more than 66 milliseconds, so that all four models on GPU can make recognition in real-time. From the training time aspect, the Inception-ResNetV2 model is more complex than VGG16, and the former consume more memory than the later.

Table 5 indicates the recognition accuracies of these four models. It is obvious that the models with LSTM have a higher accuracy than ones without LSTM. This means the combination of spatial and temporal features is beneficial for facial expressions classification. The best performing model is VGG16 with LSTM, which has a recognition accuracy of 76.08%. Meanwhile, there is no much difference in classifying capability of VGG16 and Inception-ResNetV2, with or without LSTM.

TABLE 5. Accuracy of the four models.

Model	Accurancy(%)
VGG16 without LSTM	71.54
Inception-ResNetV2 without LSTM	72.32
VGG16 with LSTM	76.08
Inception-ResNetV2 with LSTM	75.47

TABLE 6. Dataset size on different segmentation time lengths.

Segmentation Time Length	Dataset Size
2s	3305
3s	2214
4s	1667
5s	1336
6s	1118
7s	961
8s	847

C. EYE BEHAVIORS EVALUATION

Each learner's eye tracking time series in Eye Movement Database lasts for about 5 minutes, in order to accurately identify eye movement patterns, we firstly segment those eye movement time series data. In this experiment, we set 7 different segment length values that are 2s, 3s, 4s, 5s, 6s, 7s and 8s, to uniformly divide the original time series data, and test the identification performance on different time segmentation granularity. Table 6 presents the dataset size after segmenting.

The experiments for feature dimensions selection, filtering and reduction are conducted on the above segmented dataset. Here, we take the case when the segmentation time is 2 seconds, to illustrate the feature filtering procedure. Fig. 8a demonstrate the BY procedure of {(Watching,(Reading,Typing))} binary classification problem, and the right hand side figure of Fig. 8a zooms in the intersection of P value sequence and the rejection line (FDR = 0.05). From these two graphs, we find that 1588 feature dimensions are extracted from two defined features ($S_{average}$ and R_h) and forty time series features. Among these features, 970 dimensions are selected and remaining features are ignored. Likewise, 1040 and 1100 features are selected in {(Reading,(Watching,Typing))} and {(Typing,(Watching,Reading))} binary classification

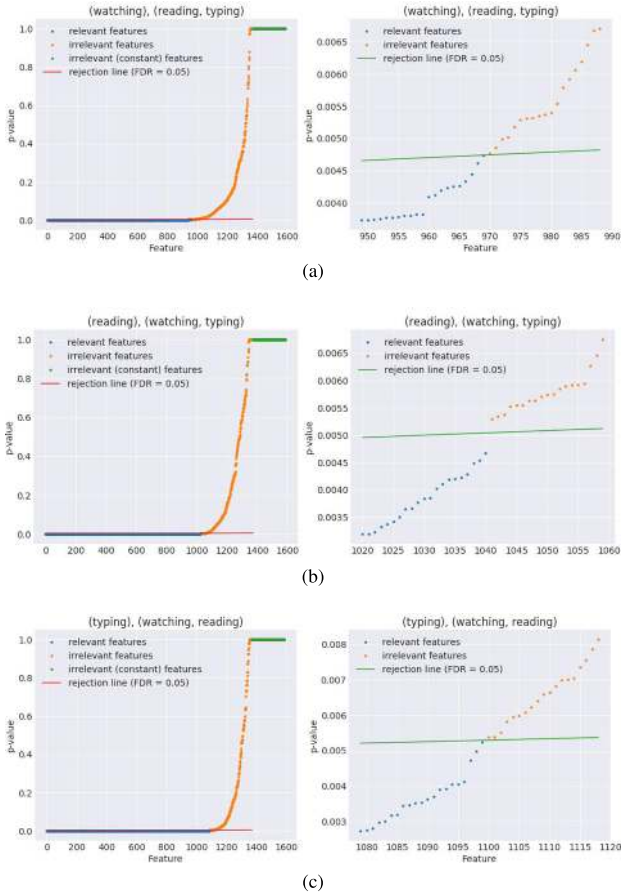


FIGURE 8. Feature selection on three different classification task. (a) Features selection of P_W. (b) Features selection of P_R. (c) Features selection of P_T.

TABLE 7. Number of sifted and reduced features on different segmentation time length.

Segmentation Time Length	Sifted Features	Reduced Features
2s	1164	371
3s	848	290
4s	778	262
5s	744	226
6s	695	199
7s	626	184
8s	609	179

problems respectively, which can be seen in Fig. 8b and 8c. Finally, we obtain 1164 features in the union set of the three binary problems, and 371 features are retained after using PCA method with a 95% threshold value.

Similarly, the features selection results of other segmentation time lengths are calculated and listed in Table 7.

We adopt three supervised models CART (Classification and Regression Trees), Random Forest, and GBDT (Gradient Boosted Decision Tree) to classify learner's eye movement using the selected features, and Table 8 illustrates the classification accuracy of these three classifiers. We find that the GBDT has best performance to classify the three eye movement with a 0.81 accuracy of recognition capability as compared to Random Forest and the CART. The highest

TABLE 8. Classification accuracy on different segmentation time length.

Segmentation Time Length	CART	Random Forest	GBDT
2s	0.60	0.70	0.73
3s	0.65	0.71	0.75
4s	0.65	0.73	0.76
5s	0.68	0.80	0.81
6s	0.66	0.74	0.77
7s	0.64	0.72	0.76
8s	0.62	0.71	0.76

TABLE 9. Retained features of three channel source data.

Channel Source	Retained Features After Using PCA(0.95)
Video or image sequence	273
Eye movement	89
Mouse dynamics	292

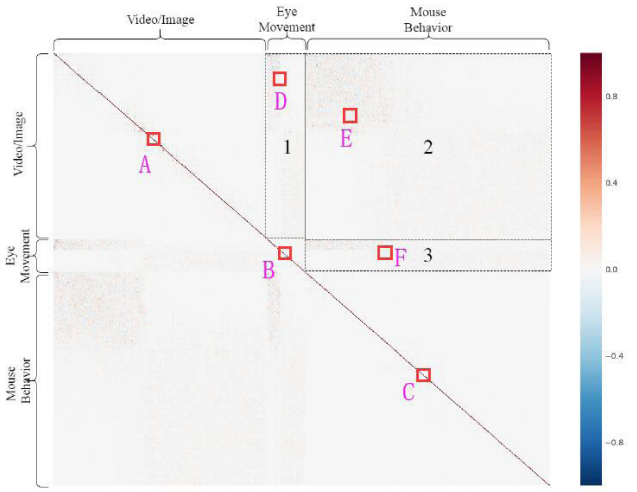


FIGURE 9. Correlation analysis on channel features.

accuracy is obtained when the segmentation time length is 5 seconds.

D. MULTI-FEATURES FUSION AND LEARNING PERFORMANCE EVALUATION

We have extracted time series features from the three channel source data collected as mentioned in the subsection A of Section IV, and applied the methods described in the subsection D of Section III to reduce and select features of each channel. At last, 654 dimensional features are retained, shown in Table 9. The correlation intensity of each two channel features is illustrated as heat map in Fig. 9, in which area 1 displays the correlated intensity between video image sequence channel and eye movement channel, area 2 shows the correlated intensity between video image sequence and mouse dynamics channel, and area 3 depicts the correlated intensity between eye movement channel and mouse dynamics channel. If the color gets closer to dark red or dark blue, it means the positive correlation or negative correlation intensity is stronger. When the color is closer to white, the intensity is weaker. Fig. 10 demonstrates six representative local heat maps, to better visualize the correlation intensity of three channels features. Fig. 10a, Fig. 10b and Fig. 10c indicate the intensity between local features from each channel itself,

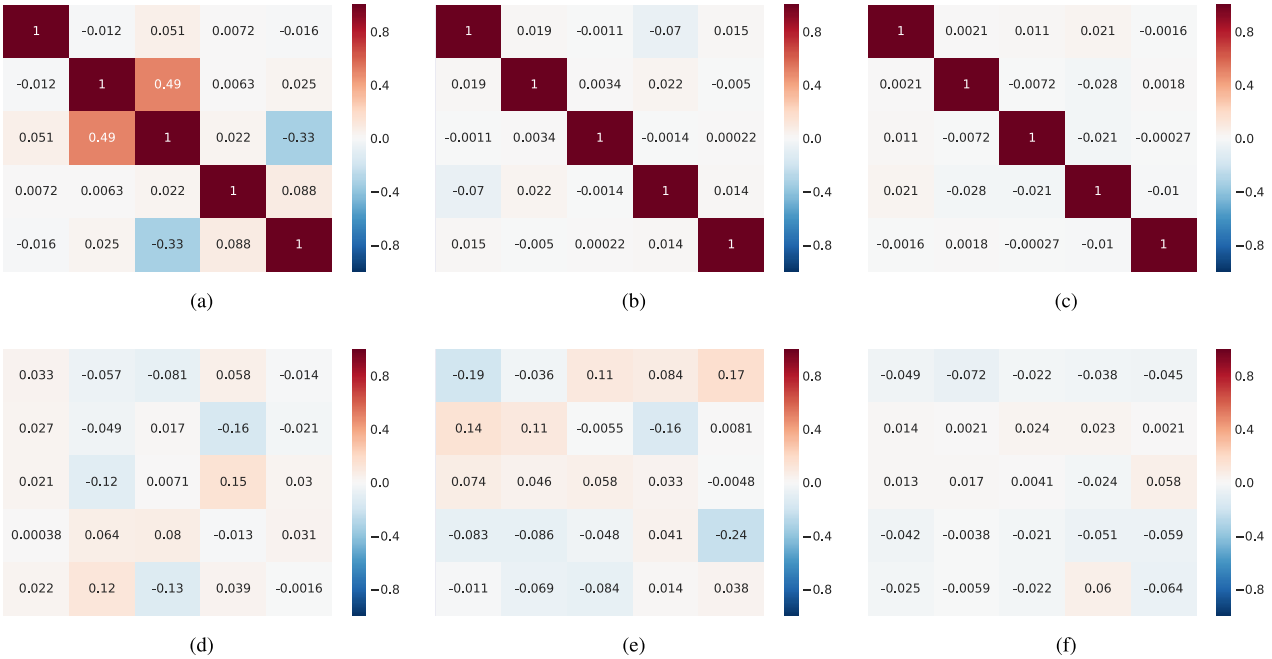


FIGURE 10. Six local area heat maps to visualize correlation intensity between channel features. (a) Intensity between video channel local features. (b) Intensity between eye gaze channel local features. (c) Intensity between mouse channel local features. (d) Intensity between video channel local features and eye gaze channel local features. (e) Intensity between video channel local features and mouse channel local features. (f) Intensity between eye gaze channel local features and mouse channel local features.

TABLE 10. Regression model performance comparison in different combined channels.

Channel Source	CART	Random Forest	GBDT
Video/image sequence	0.874	0.932	0.917
Eye movement	0.412	0.491	0.504
Mouse dynamics	0.805	0.924	0.929
Video/Image sequence + Eye movement	0.873	0.954	0.962
Eye Movement + Mouse Dynamics	0.871	0.957	0.967
Video/Image sequence + Mouse Dynamics	0.867	0.954	0.962
Video/Image sequence + Eye movement + Mouse Dynamics	0.919	0.982	0.984

and these three heat maps are locate in area A, B and C area from Fig. 9. For example, the area A represents the correlation intensity between video channel local features. Likewise, the heat maps of area D, E and F represents the correlation intensity between each two channel local features, which can be seen in Fig. 10d, Fig. 10e and Fig. 10f. For example, the area D represents the correlation intensity between video channel local features and eye gaze channel local features. From these six local heat maps, we can see that almost all areas' color are white, light red or light blue, except for diagonal area. This concludes an extremely weak correlation between each two channel features, which proves the independence of the three channel features and satisfies the prerequisites of multi-source data fusion methodology. Therefore, three channel features can be considered as independent of the each other, i.e., these features are able to provide complementary information to increase the robustness of the regression models.

Finally, we got 7224 learning performance instances with labels from 46 subjects. Unlike eye behavioral classification experiment, we took the cognitive state recognition as a regression problem. In this experiment, we chose metric R^2 to evaluate the prediction performance of three models: CART, Random Forest and GBDT. This metrics value ranges from 0 to 1, and the higher value of the metric indicates better performing model. We adopted the 10-fold cross validation method to train models on the course performance labeled data. Table 10 summarizes the results of R^2 metric value of different models on seven feature combinations. It can be observed that the performance of features-fused models outperform single channel features models, which demonstrates the effectiveness of our fusion method. Particularly, fusing video, eye movement and mouse dynamics features achieves the best prediction performance, and the metrics values of three models exceeds 0.9. We also reach a similar conclusion with eye behaviors classification experiment that has a small differences between Random Forest and GBDT, but both have much better performance than CART. When considering single channel model, the models relying on image sequence or mouse dynamics performs much better than that rely on eye behaviors for which R^2 metric values are no more than 0.51. This implies that the eye movement channel provides less useful information to predict cognitive state. This could be attributed to that the eye movement features extracted from time series do not contribute well in course performance predication, which needs more effective features.

FIGURE 11. Visually recognizing learner's engagement from three facets. (a)Monitoring learner's facial expressions, Eye movement behaviors in real-time. (b)Monitoring learner's performance on each short video course.

E. VISUALLY MONITORING LEARNING ENGAGEMENT

Finally, to select the most appropriate models or parameters for monitoring learner's facial expressions, eye movement behaviors and short video course performance, we carried out a series of comparative experiments. In the emotional state recognition experiment, we compared the performance of four different combinations of models, and selected Inception-ReNet-V2 with LSTM model for real-time detection of the learner's facial expressions. In the eye behavioral state recognition experiment, we contrasted the classification effect on 7 different time segmentation lengths with 3 models to obtain an optimal classification model with 5 seconds as the segment time. For cognitive state prediction, we fused video sequence, eye movement and mouse dynamics to predict learner's course performance. Fig. 11a exhibits the probabilities of seven facial expressions and display of eye movement behavior of a learner at the current learning moment. The learner's whole course performance on each short video is shown after completing learning, which is presented in Fig. 11b.

V. CONCLUSION AND DISCUSSION

In order to address the ''lack of supervision'' problem in e-learning environments, we propose a framework to measure multi-dimensional engagement of learners based on multi-channel data. In our research, the learning engagement is a multidimensional structure that includes emotional state, eye behavioral state and cognitive state. To measure these states of learners, three channel data streams are captured by low cost devices, including a camera, an eye-tracker and a mouse. We adopt a transfer learning strategy to fine-tune parameters and to improve our facial expression recognition model accuracy with insufficient number of Asian images. We also propose a new method to make the self-reported labels more reliable. In addition, a feature-level fusion method is designed to combine the three different channel data. The experimental results show that we have obtained 76.08% in facial expressions recognition accuracy, 81% in eye movement behavior classification precision and 0.98 of R_2 metric value of course performance prediction. These results demonstrate the effectiveness of the proposed approach and methods for feature selection, reduction and fusion. Particularly, a prototype computer-based learning environment has been developed for students to participate in a MOOC course and collect their multimodal data. More importantly, the learner's facial expressions and eye movement behaviors can be detected in real-time, and his or her course performance prediction results are shown at the end of the course video learning.

Finally, some limitations need to be considered. Firstly, this research only concentrates on e-learning or distance learning environments that students follow video courses from PC not their mobile devices, and it is practically impossible to put the eye tracker on all users' computers. A typical application scenario where our proposed work could be used in is the blended learning classroom, such as using MOOC mode in on-campus education, to provide a new insight to enhance students' learning efficiency and improve the teaching effectiveness. Secondly, it can't be ignored that engagement within e-learning has more facets than summarized and discussed in this paper, and some facets would be useful even when mentioned briefly. Specifically, engagement is not solely a matter of how learners feel about a subject or perform on a test. Engagement is also increasingly measured through how learners interact with others, both in structure and in nature [67]. Engagement is also driven by various incentives [68]. In our future work, we will put efforts on improving our method to make it more applicable. At the same time, we will combine new facets of engagement and incentives, to more precisely and comprehensively monitor learner's engagement. What's more, we also plan to expand our multi-channel dataset by inviting more participants, and we will attempt to improve generalization of face expression recognition models by employing GAN (Generative Adversarial Networks) to augment database with limited samples. Furthermore, we strive for more effective eye movement features to improve prediction performance.

REFERENCES

[1] R. GhasemAghaei, A. Arya, and R. Biddle, "A dashboard for affective E-learning: Data visualization for monitoring online learner emotions," in *Proc. EdMedia+ Innovate Learn.*, 2016, pp. 1536–1543.

[2] Q. Zheng, Y. Qian, and J. Liu, "Yotta: A knowledge map centric e-learning system," in *Proc. IEEE 7th Int. Conf. E-Bus. Eng.*, Nov. 2010, pp. 42–49.

[3] M. Zhou, "Chinese university students' acceptance of MOOCs: A self-determination perspective," *Comput. Edu.*, vol. 92, pp. 194–203, Jan./Feb. 2016.

[4] T. Daradoumis, R. Bassi, F. Xhafa, and S. Caballé, "A review on massive E-learning (MOOC) design, delivery and assessment," in *Proc. 8th Int. Conf. P2P, Parallel, Grid, Cloud Internet Comput.*, Oct. 2013, pp. 208–213.

[5] T. R. Liyanagunawardena, P. Parslow, and S. Williams, "Dropout: Mooc participants perspective," in *Proc. EMOOCs 2nd MOOC Eur. Stakeholders Summit*, Feb. 2014, pp. 95–100.

[6] W. Dai, "Analysis on learning behaviors in mooc for college students," *Argos*, vol. 35, no. 68, pp. 50–55, 2018.

[7] R. Sujatha and D. Kavitha, "Learner retention in MOOC environment: Analyzing the role of motivation, self-efficacy and perceived effectiveness," *Int. J. Edu. Develop. Using (ICT)*, vol. 14, no. 2, pp. 62–74, 2018.

[8] K. S. Hone and G. R. El Said, "Exploring the factors affecting MOOC retention: A survey study," *Comput. Edu.*, vol. 98, pp. 157–168, Jul. 2016.

[9] A. L. Reschly and S. L. Christenson, "Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct," in *Handbook of Research on Student Engagement*. Berlin, Germany: Springer, 2012, pp. 3–19.

[10] M. Ez-Zaouia and E. Lavoué, "EMODA: A tutor oriented multimodal and contextual emotional dashboard," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, Mar. 2017, pp. 429–438.

[11] N. Bosch, "Detecting student engagement: Human versus machine," in *Proc. Conf. User Modeling Adaptation Personalization*, Jul. 2016, pp. 317–320.

[12] D. Leony, P. J. Muñoz-Merino, J. A. Ruipérez-Valiente, A. Pardo, and C. D. Kloos, "Detection and evaluation of emotions in massive open online courses," *J. Universal Comput. Sci.*, vol. 21, no. 5, pp. 638–655, 2015.

[13] J. Li, G. Ngai, H. V. Leong, and S. C. Chan, "Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics," *ACM SIGAPP Appl. Comput. Rev.*, vol. 16, no. 3, pp. 37–49, Sep. 2016.

[14] Z. Ren, H. Rangwala, and A. Johri, "Predicting performance on MOOC assessments using multi-regression models," 2016, *arXiv:1605.02269*. [Online]. Available: https://arxiv.org/abs/1605.02269

[15] M. H. Baturay, "An overview of the world of MOOCs," *Procedia Social Behav. Sci.*, vol. 174, pp. 427–433, Feb. 2015.

[16] D. G. Glance, M. Forsey, and M. Riley, "The pedagogical foundations of massive open online courses," *First Monday*, vol. 18, no. 5, May 2013.

[17] C. R. Henrie, L. R. Halverson, and C. R. Graham, "Measuring student engagement in technology-mediated learning: A review," *Comput. Educ.*, vol. 90, pp. 36–53, Dec. 2015.

[18] B. M. Booth, T. J. Seamans, and S. S. Narayanan, "An evaluation of EEG-based metrics for engagement assessment of distance learners," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 307–310.

[19] D. Di Mitri, M. Scheffel, H. Drachsler, D. Börner, S. Ternier, and M. Specht, "Learning pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, Mar. 2017, pp. 188–197.

[20] N. J. Pienta, "Studying student behavior and chemistry skill using browser-based tools and eye-tracking hardware," *Química Nova*, vol. 40, no. 4, pp. 469–475, May 2017.

[21] S. D'Mello, E. Dieterle, and A. Duckworth, "Advanced, analytic, automated (AAA) measurement of engagement during learning," *Educ. Psychologist*, vol. 52, no. 2, pp. 104–123, Feb. 2017.

[22] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.

[23] K. R. Koedinger, S. D'Mello, E. A. McLaughlin, Z. A. Pardos, and C. P. Rose, "Data mining and education," *Wiley Interdiscipl. Rev., Cogn. Sci.*, vol. 6, no. 4, pp. 333–353, Jul./Aug. 2015.

[24] C.-H. Wu, Y.-M. Huang, and J.-P. Hwang, "Review of affective computing in education/learning: Trends and challenges," *Brit. J. Educ. Technol.*, vol. 47, no. 6, pp. 1304–1323, Nov. 2016.

[25] C. Cunha-Pérez, M. Arevalillo-Herráez, L. Marco-Giménez, and D. Arnau, "On incorporating affective support to an intelligent tutoring system: An empirical study," *IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje*, vol. 13, no. 2, pp. 63–69, May 2018.

[26] O. C. Santos, "Emotions and personality in adaptive e-learning systems: An affective computing perspective," in *Emotions and Personality in Personalized Services*. Springer, 2016, pp. 263–285.

[27] A. Landowska, "Affective computing and affective learning-methods, tools and prospects," *Stara Strona Magazynu Eduakcja*, vol. 5, no. 1, pp. 1–16, Nov. 2013.

[28] M. Magdin, M. Tur áni, and L. Hudec, "Evaluating the emotional state of a user using a Webcam," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 4, no. 1, pp. 1–8, Sep. 2016.

[29] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learn. Instruct.*, vol. 22, no. 2, pp. 145–157, Apr. 2012.

[30] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, I. Estévez-Ayres, and C. D. Kloos, "Sentiment analysis in MOOCs: A case study," in *Proc. IEEE Global Eng. Edu. Conf. (EDUCON)*, Apr. 2018, pp. 1489–1496.

[31] V. Tze, L. M. Daniels, E. Buhr, and L. Le, "Affective profiles in a massive open online course and their relationship with engagement," *Frontiers Edu.*, vol. 2, p. 65, Dec. 2017.

[32] S. A. Narayanan, M. Kaimal, K. Bijlani, M. Prasanth, and K. S. Kumar, "Computer vision based attentiveness detection methods in E-learning," in *Proc. Int. Conf. Interdiscipl. Adv. Appl. Comput.*, Oct. 2014, p. 51.

[33] L. Wang, "Attention decrease detection based on video analysis in e-learning," in *Transactions on Edutainment XIV*. New York, NY, USA: Springer, 2018, pp. 166–179.

[34] B. M. Booth, A. M. Ali, S. S. Narayanan, I. Bennett, and A. A. Farag, "Toward active and unobtrusive engagement assessment of distance learners," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 470–476.

[35] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," *Comput. Intell. Neurosci.*, vol. 2018, Oct. 2018, Art. no. 6347186.

[36] T. Phan, S. G. McNeil, and B. R. Robin, "Students' patterns of engagement and course performance in a massive open online course," *Comput. Edu.*, vol. 95, pp. 36–44, Apr. 2016.

[37] S. Guo and W. Wu, "Modeling student learning outcomes in moocs," in *Proc. 4th Int. Conf. Teach., Assessment, Learn. Eng.*, 2015, pp. 1305–1313.

[38] W. Zhang, X. Huang, S. Wang, J. Shu, H. Liu, and H. Chen, "Student performance prediction via online learning behavior analytics," in *Proc. Int. Symp. Educ. Technol. (ISET)*, Jun. 2017, pp. 153–157.

[39] B. A. Greene, "Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research," *Educ. Psychologist*, vol. 50, no. 1, pp. 14–30, Jan. 2015.

[40] M. Vuorela and L. Nummenmaa, "Experienced emotions, emotion regulation and student activity in a web-based learning environment," *Eur. J. Psychol. Edu.*, vol. 19, no. 4, pp. 423–436, Dec. 2004.

[41] V. Carchiolo, A. Longheu, M. Previti, and G. Fichera, "Monitoring students activities in CS courses," in *Proc. 15th RoEduNet Conf., Netw. Edu. Res.*, Sep. 2016, pp. 1–6.

[42] D. Canedo, A. Trifan, and A. J. Neves, "Monitoring students' attention in a classroom through computer vision," in *Proc. Int. Conf. Practical Appl. Agents Multi-Agent Syst.* Springer, 2018, pp. 371–378.

[43] F. Tian, P. Gao, L. Li, W. Zhang, H. Liang, Y. Qian, and R. Zhao, "Recognizing and regulating e-learners' emotions based on interactive Chinese texts in e-learning systems," *Knowl. Based Syst.*, vol. 55, pp. 148–164, Jan. 2014.

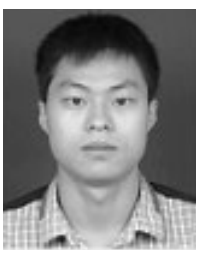
[44] A. Hershkovitz and R. Nachmias, "Learning about online learning processes and students' motivation through Web usage mining," *Interdiscipl. J. E-Learn. Learn. Objects*, vol. 5, no. 1, pp. 197–214, Jan. 2009.

[45] S. Saha, S. Datta, A. Konar, and R. Janarthanan, "A study on emotion recognition from body gestures using Kinect sensor," in *Proc. Int. Conf. Commun. Signal Process.*, Apr. 2014, pp. 056–060.

[46] C. Weigle and D. C. Banks, "Analysis of eye-tracking experiments performed on a Tobii T60," *Proc. SPIE*, vol. 6809, Jan. 2008, Art. no. 680903.

[47] A. Kołakowska, "Recognizing emotions on the basis of keystroke dynamics," in *Proc. 8th Int. Conf. Hum. Syst. Interact. (HSI)*, Jun. 2015, pp. 291–297.

[48] G. Tsoulouhas, D. Georgiou, and A. Karakos, "Detection of learner affective state based on mouse movements," *J. Comput.*, vol. 3, no. 11, pp. 9–18, Nov. 2011.
[49] H. Qi, X. Wang, S. S. Iyengar, and K. Chakrabarty, "Multisensor data fusion in distributed sensor networks using mobile agents," in *Proc. 5th Int. Conf. Inf. Fusion*, Aug. 2001, pp. 11–16.
[50] Y. Gogia, E. Singh, S. Mohatta, and V. Sreejith, "Multi-modal affect detection for learning applications," in *Proc. IEEE Region Conf. (TENCON)*, Nov. 2016, pp. 3743–3747.
[51] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," 2004, *arXiv:cs/0410027*. [Online]. Available: https://arxiv.org/abs/cs/0410027
[52] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *Proc. 5th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2010, pp. 375–382.
[53] P. Ekman, "An argument for basic emotions," *Cognit. Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992.
[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[55] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 682–691, Nov. 2010.
[56] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1749–1756.
[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[58] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[59] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400
[60] W. Liu, Z. Fan, F. Liu, J. Xu, and W. Cheng, "Gaze based behavior analysis of online learners," in *Proc. 5th Int. Conf. Inf. Edu. Technol.*, Jan. 2017, pp. 44–48.
[61] *Overview on Extracted Features*. Accessed: May 2018. [Online]. Available: https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html
[62] F. J. Massey, Jr., "The Kolmogorov-Smirnov test for goodness of fit," *J. Amer. Statist. Assoc.*, vol. 46, no. 253, pp. 68–78, 1951.
[63] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Stat.*, vol. 29, no. 4, pp. 1165–1188, 2001.
[64] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," 2016, *arXiv:1610.07717*. [Online]. Available: https://arxiv.org/abs/1610.07717
[65] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[66] *Standard Normal (z) Table*. Accessed: Jun. 2019. [Online]. Available: http://www.sjsu.edu/faculty/gerstman/EpiInfo/z-table.htm
[67] A. A. Tawfik, T. D. Reeves, A. E. Stich, A. Gill, C. Hong, J. McDade, V. S. Pillutla, X. Zhou, and P. J. Giabbanelli, "The nature and level of learner–learner interaction in a chemistry massive open online course (MOOC)," *J. Comput. Higher Edu.*, vol. 29, no. 3, pp. 411–431, Dec. 2017.
[68] T. D. Reeves, A. A. Tawfik, F. Msilu, and I. im ek, "What's in it for me? Incentives, learning, and completion in massive open online courses," *J. Res. Technol. Edu.*, vol. 49, nos. 3–4, pp. 245–259, 2017.



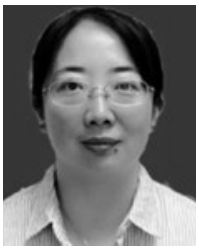
JIA YUE received the B.S. degree in automation from Xidian University, in 2014. He is currently pursing the Ph.D. degree with the Satellite-Terrestrial Network Technology Research and Development Key Laboratory, Xi'an Jiaotong University, focusing on multimodal learning.

FENG TIAN has been with Xi'an Jiaotong University, since 2004, where he is currently with the National Engineering Laboratory of Big Data Analytics and with the Systems Engineering Institute, as a Professor. He is a member of the Satellite-Terrestrial Network Technology Research and Development Key Laboratory, Shaanxi. His research interests include big data analytics, learning analytics, system modeling and analysis, and cloud computing.

KUO-MIN CHAO is currently a Professor of computing with Coventry University, U.K. He has over 150 refereed publications. His research interests include cloud computing, big data, and their applications.

NAZARAF SHAH is currently a Senior Lecturer with Coventry University, U.K. He has over 50 publications in various international conferences and journals. His research interests include intelligent agent and cloud computing.

LONGZHUANG LI is currently a Professor with Texas A&M University-Corpus Christi. His research interests include data integration and data mining and has been supported by the National Science Foundation and Air Force Research Laboratory.



YAN CHEN is currently an Associate Professor and a Graduate Supervisor with the Department of Computer Science and Technology, Xi'an Jiaotong University. Her research interests include educational data mining and learning analytics.

QINGHUA ZHENG received the B.S. degree in computer software, the M.S. degree in computer organization and architecture, and the Ph.D. degree in system engineering from Xi'an Jiaotong University, China, in 1990, 1993, and 1997, respectively.

He did post-doctoral research at Harvard University, in 2002, and was a Visiting Professor of Research with The University of Hong Kong, from 2004 to 2005. He received the First Prize for National Teaching Achievement, State Education Ministry, in 2005, and the First Prize for Scientific and Technological Development of Shanghai City and Shaanxi Province, in 2004 and 2003, respectively.