IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 June 2023, accepted 10 July 2023, date of publication 21 July 2023, date of current version 28 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297651

APPLIED RESEARCH

Analysis of Facial Expressions to Estimate the Level of Engagement in Online Lectures

RENJUN MIAO[1], HARUKA KATO[1], YASUHIRO HATORI[1,2], YOSHIYUKI SATO[2,3], AND SATOSHI SHIOIRI[1,2,3]

[1]Graduate School of Information Sciences, Tohoku University, Sendai, Miyagi 980-8577, Japan
[2]Research Institute of Electrical Communication, Tohoku University, Sendai, Miyagi 980-8577, Japan
[3]Advanced Institute for Yotta Informatics, Tohoku University, Sendai, Miyagi 980-8577, Japan

Corresponding author: Renjun Miao (miao.renjun.s1@dc.tohoku.ac.jp)

This work was supported in part by the Research Project Program of Research Center for 21st Century Information Technology (IT-21 Center), Research Institute of Electrical Communication (RIEC), Tohoku University; and in part by the Yotta Informatics Project by Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. The work of Satoshi Shioiri was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 19H01111.

ABSTRACT The present study aimed to develop a method for estimating students' attentional state from facial expressions during online lectures. We estimated the level of attention while students watched a video lecture by measuring reaction time (RT) to detect a target sound that was irrelevant to the lecture. We assumed that RT to such a stimulus would be longer when participants were focusing on the lecture compared with when they were not. We sought to estimate how much learners focus on a lecture using RT measurement. In the experiment, the learner's face was recorded by a video camera while watching a video lecture. Facial features were analyzed to predict RT to a task-irrelevant stimulus, which was assumed to be an index of the level of attention. We applied a machine learning method, light Gradient Boosting Machine (LightGBM), to estimate RTs from facial features extracted as action units (AUs) corresponding to facial muscle movements by an open-source software (OpenFace). The model obtained using LightGBM indicated that RTs to the irrelevant stimuli can be estimated from AUs, suggesting that facial expressions are useful for predicting attentional states while watching lectures. We re-analyzed the data while excluding RT data with sleepy faces of the students to test whether decreased general arousal caused by sleepiness was a significant factor in the RT lengthening observed in the experiment. The results were similar regardless of the inclusion of RTs with sleepy faces, indicating that facial expression can be used to predict learners' level of attention to video lectures.

INDEX TERMS Attention, affective computing, engagement, facial features, online lecture.

I. INTRODUCTION

Understanding students' engagement levels while studying is important for improving learning outcomes. To improve the quality of education, it is crucial to estimate learners' level of engagement with their studies. However, it is difficult for teachers to pay attention to all students, particularly in online classes. Automated measurement of engagement levels may be helpful for improving learning conditions. For online learning, webcams can be used to capture learners' facial expressions, which can be used to estimate their mental states [1], [2], [3]. For example, Shioiri et al. conducted image preference estimation from facial expressions and found that this information was useful for estimating subjective judgments of image preference. In education-related studies, Thomas and Jayagopi recorded students' face images in a classroom while they were studying with video material on a screen and estimated the level of engagement from students' facial expressions [4], [5]. The authors succeeded in predicting engagement, suggesting the usefulness of facial expressions for estimating the level of engagement. Heart rate has also been used to estimate mental states during learning. Darnell and Krieg showed that changes in heart rate are related to students' activity during a class [6]. Although previous studies have focused on engagement, which is assessed externally, this research has also been extended to

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

the measurement of internal states, which can be investigated by estimating internal states. In these studies, the mental state used as ground truth is based on subjective judgments [4], [7]. However, mental states involve factors other than those that can be evaluated subjectively. Unconscious processes, which cannot be estimated subjectively, may play more important roles than conscious processes. Thus, it is unlikely that subjective judgments are suitable for use as indexes of mental states. For example, heart rate change is reported to be a useful index of students' activity, and is not necessarily related to the subjective estimation of attention and engagement [6]. As such, it is important to develop methods involving objective measures for estimating the level of engagement. A previous study showed that facial features could be useful for estimating reaction time (RT) for mental calculations [8]. This result suggests that RT could be a good index of attention if it varies depending on focusing on the task as typically assumed in attention studies for simple detection, discrimination, or identification of visual stimuli. However, this type of measure is not available for lectures. Therefore, we attempted to use RT for task-irrelevant stimuli.

Although engagement is a term used with different meanings in different contexts [7], [9], it is often used in relation to attention [10], [11], [12], [13]. Attention to lectures, classes, and tasks is thought to be closely related to engagement. Here we use the term attention to refer to the facilitation of sensory processing by endogenous intention or salient exogenous stimulation, and consider it to be a major factor for engagement. It should be noted that engagement has also been used to indicate mental states of a longer duration in some previous studies, such as a whole lecture [6], [7], [14], [15], [16]. We measured levels of attention as an index of engagement during lectures in this study.

We designed an experiment in which participants were asked to detect an auditory target while watching a lecture video. The primary task of the experiment was to understand the lecture, and the secondary task was to detect the target. RT to the auditory target was used as an objective measure of attention level on the lecture videos. Here, we assumed that the time required to detect a target that was irrelevant to the primary task would be longer when the participant focused more on the primary task (i.e., watching video lectures in this experiment). Face images of participants were recorded while watching the videos, and facial expressions were analyzed after the experiment. The purpose of the study was to estimate the RT from facial expressions to develop a method for estimating engagement level from learners' face images.

Some of the results in this study with a smaller number of participants were published in a post-conference book as a preliminary report [14]. Here, we report analyses of facial expressions in more detail with data from a larger number of participants to consider the contributions of specific facial features, the effect of individual variation, and the effect of general arousal level or sleepiness.

II. EXPERIMENT

We conducted an experiment to investigate the relationship between the attention level and facial expression while watching video lectures. To estimate the level of attention in video lectures, we measured RT to an auditory target that was irrelevant to the lecture. We assumed that RT to an irrelevant stimulus would be longer when participants were focusing on the lecture compared with when they were not. The effect on brain responses to irrelevant stimuli has been suggested to be able to estimate attention to the primary task. For example, Kramer et al. conducted electroencephalography (EEG) measurements and reported that the event-related response (ERP) to a task-irrelevant stimulus changes with the difficulty of a primary task [18]. Similar changes were expected with RT measurements because both ERP and RT have been used to estimate attention in general [19]. In the current study, we attempted to use recorded face images to predict RT.

The auditory target we used was the disappearance of continuous white noise instead of the appearance of a sound stimulus, whereas previous experiments to measure attention have typically used a pulse stimulus [20], [21]. The reason for using the disappearance of sound was to avoid the influence of bottom-up attention to a salient stimulus, such as an auditory pulse. Bottom-up attention to a salient stimulus could be strong enough to mask the effect of attention to the lecture. Indeed, the effect of top-down attention cannot be detected when there is only one transient stimulation, while a target is discriminated by top-down attention among many transient stimuli [22], [23].

Fifteen participants (average age, 23.1 years) took part in the experiment. Participants had normal or corrected-to- normal vision and normal audition. Participants were instructed to watch a series of nine video lectures and to answer questions at the end of each video (Fig. 1). Participants were also instructed to press a key when they noticed the auditory target (the sudden disappearance of white noise) while watching the video lecture. Participants were instructed that the lecture was the primary task of the experiment while the detection of the target was a secondary task, and they were required to answer questions at the end of the experiment. RT to the target was measured to estimate participants' attention level at the time of the target presentation.

The learning materials were from an introductory course about a computer language, PHP, which was posted on YouTube [24]. The videos were shown on a computer display (MacBook Pro, Apple, California) with headphones (MDR-7506, Sony, Tokyo) in a room with office lighting (483.2 lx on the desk on which the computer was placed, and 211.8 lx at the location of the participant's face). The average loudness of the lecturer's voice was 70 db, and that of the white noise was 0.66 db.

The white noise occasionally disappeared, which was the target for the secondary task. The interval between two

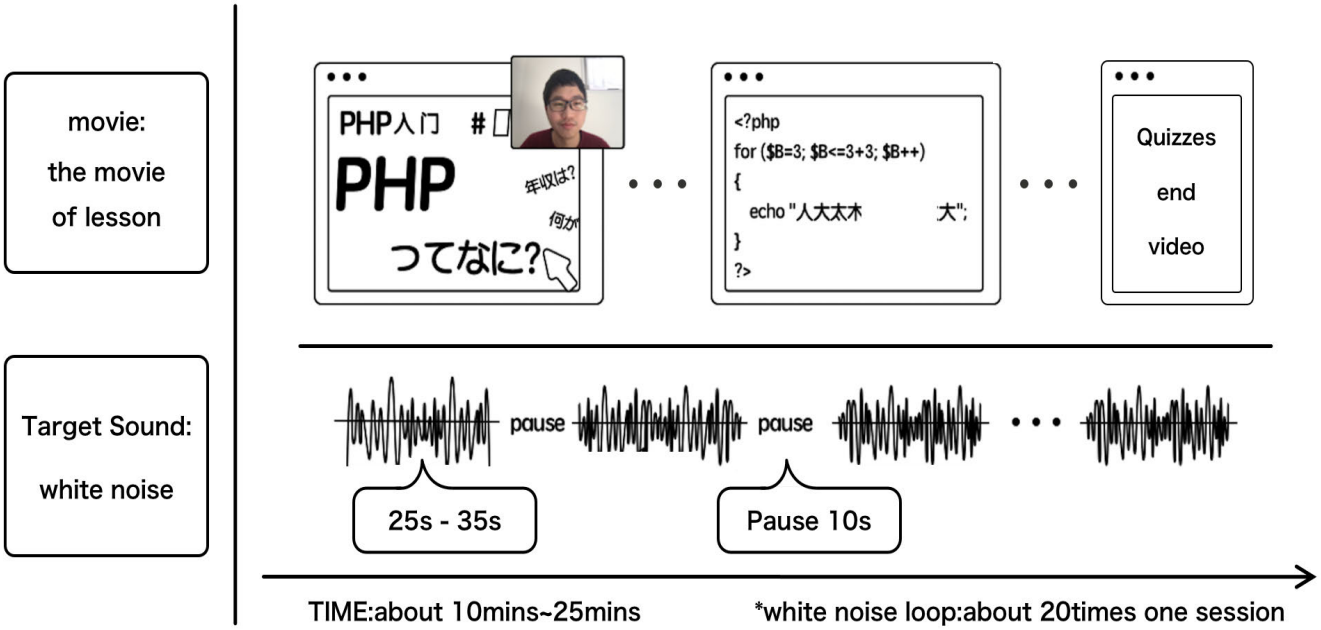

FIGURE 1. Experimental design. While watching a video lecture of an introductory PHP course in a session, auditory signals of white noise were added to the original auditory track of the lecturer video. At the end of a session after watching the video, participants answered to several quizzes about the lecture.

targets, which was a period of white noise presentation, was randomly selected between 25 and 35 seconds. The white noise started again immediately after the key press to indicate detection, or after a period of 10 seconds if no key press had been performed. Each lecture lasted between 10 and 20 minutes, depending on the content. At the end of each lecture, eight questions were provided in a google form format. For each question, participants selected one of four choices as their answer.Watching one lecture is a session of the experiment. There were nine lecture sessions.

In addition to the lecture sessions, there were two control sessions to measure RT for the detection task without paying attention to the lecture, so that the total number of session was eleven. In the control sessions, two videos from the same video lectures were used so that the participants knew the content and had little or no reason to be attracted to the content. Participants were asked to focus on the white noise and told that they did not have to pay attention to the content of the video on the display. The first control session was conducted as the 6th session with the first lecture video, and the second control session was conducted as the 11th session with the 6th lecture video used at the 7th session. The experiment was conducted over 2 days. Five lecture sessions and the first control session were performed on the first day, and the rest of the sessions (four lectures and one control session) were performed on the second day. The interval between the first and second days was within 1 week. The total duration of the experiment, 11 sessions, was approximately 130 minutes.

III. FACIAL FEATURE ANALYSIS

Participants' faces were recorded while watching lecture videos, and their facial features were analyzed after the experiment. We analyzed face images recorded in the 3 seconds before the target presentation (disappearance of white noise), using OpenFace [25] to extract the facial features. To perform facial expression analysis using OpenFace, the first step is to gather facial images or video data. From each video frame, OpenFace detects a face (multiple faces can be detected while there was only one face in our experiment) and locate it in the frame. Then, it makes facial appearance as face orientation and makes facial landmarks such as boundaries of eyes, eyebrows, and mouth. By analyzing the position changes of the facial landmarks and facial appearance, OpenFace evaluates the degree of facial muscle activity as action units (AUs). AUs are assigned to muscle movements related to facial expressions based on the Facial Action Coding System (FACS) [26]. For example, AU1 indicates the raising of the inner eyebrows, AU4 indicates the lowering of the eyebrows, and AU5 indicates the raising of the upper lids (Table 1). OpenFace offers several research advantages for facial analysis. Firstly, leveraging deep learning techniques, particularly convolutional neural networks (CNNs), OpenFace achieves high accuracy in facial recognition and feature extraction tasks. This is crucial for research projects that require precise identification and comparison of facial features. Secondly, OpenFace not only enables facial recognition but also facilitates the extraction of facial features such as expressions and poses. This broadens its applications in research areas such as facial

TABLE 1. Meanings of AUs are also listed.

Action Unit	Description	Facial Muscle
AU1	Inner Brow Raiser	Frontalis, pars medialis
AU2	Outer Brow Raiser	Frontalis, pars lateralis
AU4	Brow Lowerer	Depressor Glabellae, Depressor Supercilli, Currugator
AU5	Upper Lid Raiser	Levator palpebrae superioris
AU6	Cheek Raiser	Orbicularis oculi, pars orbitalis
AU7	Lid Tightener	Orbicularis oculi, pars palpebralis
AU9	Nose Wrinkler	Levator labii superioris alaquae nasi
AU10	Upper Lip Raiser	Levator Labii Superioris, Caput infraorbitalis
AU12	Lip Corner Puller	Zygomatic Major
AU14	Dimpler	Buccinator
AU15	Lip Corner Depressor	Depressor anguli oris (Triangularis)
AU17	Chin Raiser	Mentalis
AU20	Lip stretcher	Risorius
AU23	Lip Tightener	Orbicularis oris
AU25	Lips part	Depressor Labii, Relaxation of Mentalis (AU17), Orbicularis Oris
AU26	Jaw Drop	Masetter; Temporal and Internal Pterygoid relaxed
AU28	Lip Suck	Orbicularis oris
AU45	Blink	Relaxation of Levator Palpebrae and Contraction of Orbicularis Oculi, Pars Palpebralis.

emotion recognition, facial tracking, and facial attribute analysis. Thirdly, being an open-source toolkit, OpenFace allows researchers to modify and customize it according to their specific needs. This flexibility enables adjustments and improvements tailored to individual research objectives and various application scenarios. Fourthly, OpenFace supports processing large datasets of facial images and videos. This is particularly valuable for research projects that involve handling extensive data, such as facial recognition in video surveillance systems or the establishment of facial image databases. All of these advatages are important to us particulary when to apply research achievements to practical occasions. We arbitrarily chose the period of time between 3 sec and 0 sec before target presentation as the time window during which the effect of attention might be reflected in target detection, but the uses of 1 or 5 seconds showed similar results (see Fig. 5).

The features of the facial expressions were extracted as AUs from the video taken for each target presentation using OpenFace as well as the positions and angles of the head and eyes. The meanings of AUs are shown in Table 1. Two types of AU indexes are available from OpenFaces: a continuous value between 0 and 5 for 17 AUs (referred to as AUr) and a binary value of 0 or 1 (absence or presence) for 18 AUs (referred to as AUc), which are 17 AUs and the AU28 for Lip Suck. Because we collected data for a 3-sec period for each target, we used statistical features of the time-varying values: minimum, maximum, mean, standard deviation, and three levels of percentiles (25%, 50%, and 75%) for AUr and mean and standard deviation for AUc. The number of parameters was 155 in total numbers of variables in total. There are perhaps better statistical features of sequential data rather than what we used here. However, they were sufficient to show the usefulness of the AUs to predict RT (see later). For better prediction in the future, we could investigate more complex temporal features.

To investigate the relationship of facial expressions with the RT of target detection, we attempted to predict RT from AUs using a machine learning method called LightGBM [27]. LightGBM is a gradient boosting model, which operates quickly and exhibits relatively accurate performance in general. LightGBM is a decision tree model with gradient boosting, in which the node of trees grows to minimize the residuals. Since training data with large residuals are used preferentially, thus learning proceeds efficiently, which is a powerful machine learning technique that can be used for both regression and classification tasks. It works by combining multiple weak learners (simple decision trees) into a strong learner, which is able to make accurate predictions on new data. In this study, two different methods were tested for RT predictions of AUs. One method was to train a model with pooled data of all participants (pooled data model), and the other was to train a model with all but one participant and test with the remaining participant (across individual test models). The latter method was to investigate individual differences. If individual differences are small, the model built with other participants should be able to predict RTs of the participant tested. However, individual variations may prevent the building of a general model that can be used for anyone whose data are not used to build the model.

For the evaluation of the models, a 15-fold cross-validation method was used. All data were divided into 15 groups randomly for the pooled data model, 14 of which were used for training and the remaining group was used for testing. The process was repeated 15 times, one test for each group, and the average was used as the model performance. For the across-individuals test model, data for 14 of 15 participants were used for training, and data for the remaining participant were used for testing. The process was repeated 15 times, with one test for each participant. The average of the 15 test scores was used as the model performance. Prediction performance was assessed by the root mean square error (RMSE) of the prediction against the data and by the Pearson's correlation coefficient between the data and the prediction.(Fig. 2)

IV. RESULTS

Target presentations without responses within 10 sec were excluded from the reaction time (RT) analysis. Such target presentations occurred on 5.5% of trials on average across all participants. The average RT over all sessions of all participants was 1.1 sec, with a standard deviation of 2.3 sec. Because average RT varied among participants, we normalized RT as Z-scores after taking the logarithm. We took the logarithm of RT to minimize the effects of asymmetrical distribution (usually a heavy tail for longer

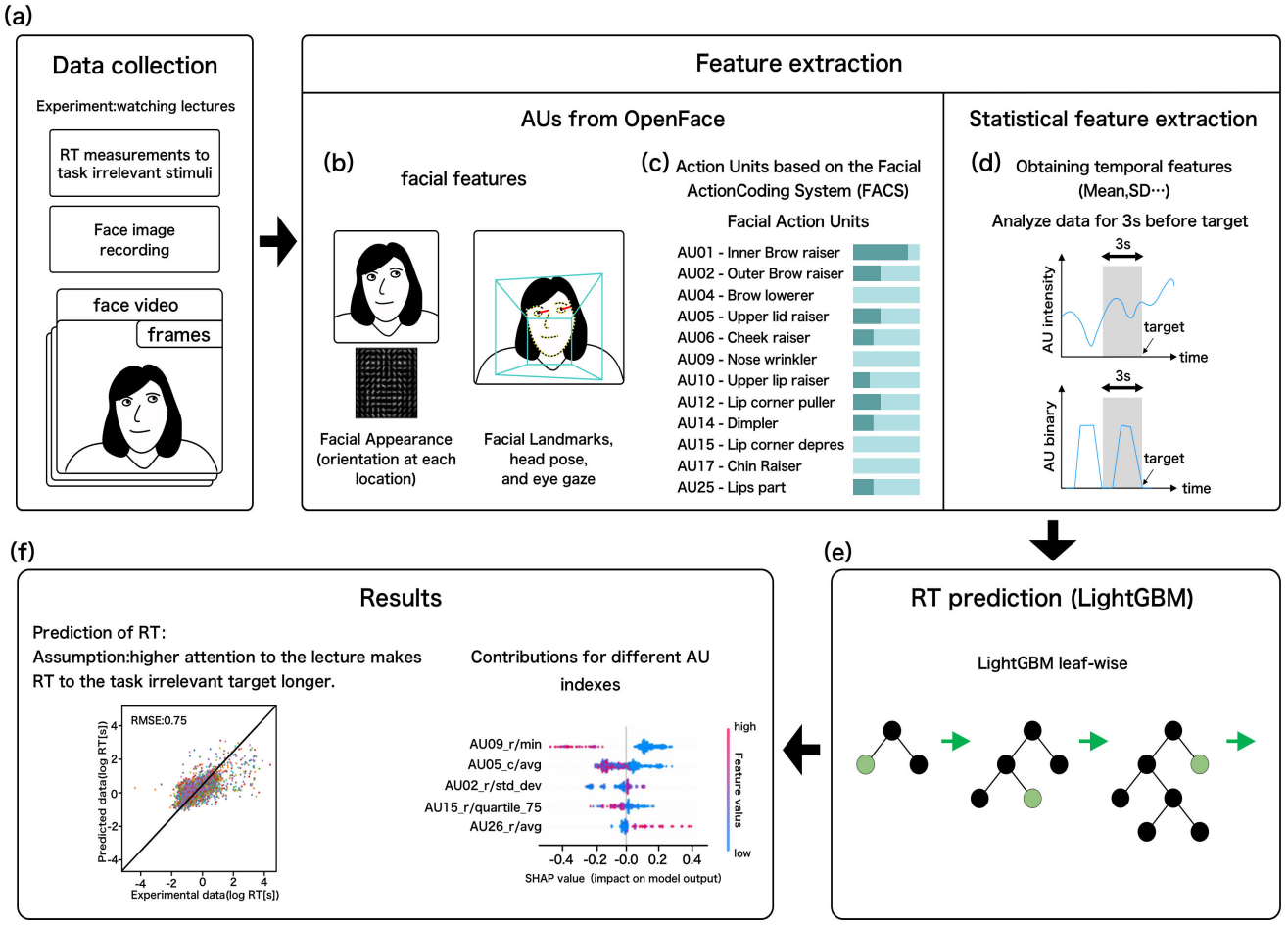

FIGURE 2. The framework of the analysis: (a) Video recording of participants' faces while watching online lectures and recording of reaction time of target detection measured as the time from the target presentation (disappearance of white noise and the key press for the detection). (b) Orientation information at each location as facial appearance and landmarks on a face, such as the eyes, nose, mouth, and chin are detected for all video frames through each session by OpenFace. Facial appearance and landmarks are used to obtain AUs based on the Facial Action Coding System(FACS). Head pose and eye gaze are also detected. Head pose is an important factor to analyzed face images as normalized fashion. (c) OpenFace extracted Action units (AUs) from facial landmarks and appearance for each frame. (d) We used several statistical measures of sequential AU values from a time window (3 s for main analysis and 1s and 5s were also used) before each target presentation. Used statistical measures were average, standard deviation, minimum, maximum, and percentiles of 25th, 50th, and 75th for intensity indexes. Only average and standard deviation were used for binary indexes. (e) The statistical measures from all AUs were used to predict reaction time using a machine learning method, LightGBM. LightGBM constructs a tree-type model with leaf-wise tree growth, choosing the leaf with max delta loss to grow. (f) We compared predicted RTs with measured RTs, showing their correlation. Higher correlation indicates that the LightGBM model can predict RTs to task-irrelevant stimulus well, so that the model can predict the attention level at the time of target presentation under the assumption that higher attention to the lecture makes RT to the task irrelevant target longer. We also analyzed strength of contribution using a method called Shapley additive explanations (SHAP). SHAP shows relationship between contribution values (strength to contribute the prediction) and each of feature indexes.

RTs). We also used normalized values of AUs by Z-scoring to avoid the effects of individual variations of facial features. We expected that variations of AUs after normalization were related to changes in mental processes, whereas the absolute AU values include facial differences among different individuals. We then applied LightGBM to model the relationship between RT and facial expressions, and tested the model using a 15-fold cross-validation method. Fig. 3a shows the prediction results of the pooled data model. The horizontal axis shows RT measured in the experiment and the vertical axis shows the prediction from LightGBM. Each point represents each target presentation from all sessions of all participants and different colors indicate different training-test combinations (15 different combinations with different colors). The RMSE of data deviation from the predictions (or the deviation of predictions from the data) was 0.75. The average of the RT data is zero, with a unit standard deviation after Z scoring by definition. Thus, the RMSE of model prediction (0.75, which is smaller than 1) indicates that the model can at least partially explain the data variation (25% in this case). The Pearson's correlation coefficient between data and prediction was 0.66. A statistical test of no correlation showed that the correlation was statistically significant ($p < 0.001$, $t(2412) = 11$). We used a test to examine whether the Pearson's correlation coefficient is not significantly different from zero and showed the assumption

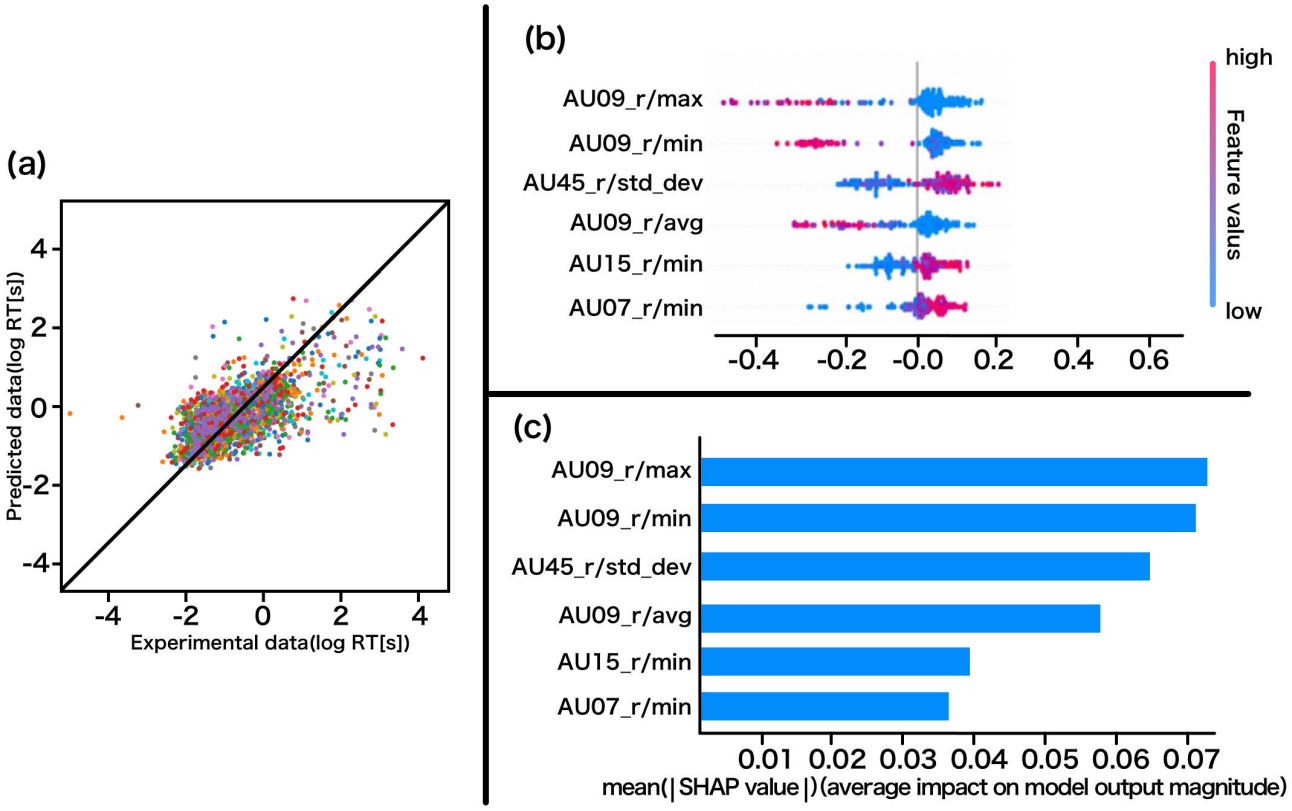

FIGURE 3. (a) Correlation between measured reaction time (RT) and the predicted RT of the model. Each point represents the RT of each target presentation from all sessions of all participants. Different colors indicate different training-test combinations (15 combinations). (b) Indexes are arranged according to the level of contribution to the prediction obtained using the Shapley additive explanations (SHAP) method. Each point is from each RT, as in the correlation figure of figure3 (a), and the color (red or blue) indicates a positive or negative contribution. The horizontal axis indicates the level of contribution to the prediction of the RT by the model (c) The absolute value that corresponds to the contribution of each index to the prediction estimated by SHAP.

of not different was rejected with a level of 5%.In addition to the statistical significance of correlation coefficient, we also used a statistical test of RMSE to show that our prediction is better than chance. We compared RMSE of the model prediction and that of data, which is one after Z-scoring, using a t-test ($p < 0.001$, $t(14) = 16.62$). The present analysis successfully predicted RT to task-irrelevant targets, which we assumed to vary depending on attention states. This prediction of RT, in turn, predicted the attention state at the time some seconds before the target presentation during learning. We concluded that facial features and movements of the head and eyes contain information about attention.

Further analysis revealed the level of contribution of each index to the prediction (i.e., the importance of each index for the prediction) using a method called Shapley additive explanations (SHAP) [28]. SHAP provides the value that corresponds to the contribution of each input feature to the prediction (Fig. 3 c). AU9 (nose wrinkler), AU45 (blink), AU15 (lip corner depressor), and AU7 (lid tightener) were the best five contributors among all AUs. The analysis also provides the degree of contribution of each input feature for predicting each event of target detection, as shown by the dots in Fig. 3 (b). Red dots indicate high values of facial feature indexes and blue dots indicate low values. The patterns of dot distribution of data points in red and blue show, for example, that AU9 negatively contributes to RT. Higher values (red dots) were distributed toward the negative direction of the horizontal axis, indicating that shorter RT was associated with more nose wrinkling, which, in turn, suggests that less attention was paid to the lecture when more nose wrinkling was exhibited. We will discuss the effect of these AUs in more detail in the Discussion section.

We performed control sessions to confirm that watching a lecture video influences RT to the auditory target. In the control condition, participants were asked to detect the target without paying any attention to the video lecture. RTs in this condition are considered to reflect full attention to the auditory target. The average RT for the two control sessions across all participants was 0.7 sec. This RT duration is clearly shorter than the average RT in the lecture sessions, which was 1.1 sec. and the Pearson's correlation coefficient between the

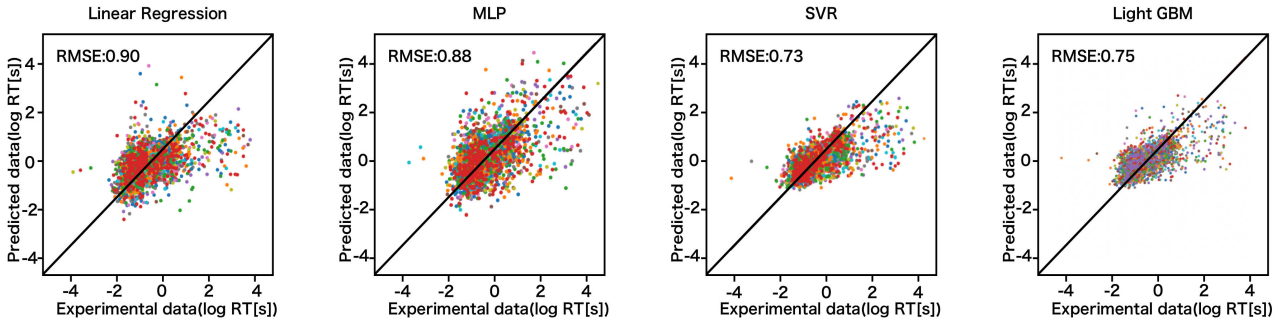

FIGURE 4. Comparison of four different models: Support Vector Regression (SVR), Multilayer perceptron (MLP), Linear Regression and LightGBM.

experiment and control sessions was statistically significant ($p < 0.05$, $t(14) = 2.74$), indicating that RT to the target was an appropriate measure of attention to the lectures. We attempted to predict RTs of the control conditions with the same procedure used for the lecture session. The results revealed that the RMSE of the predictions was 0.89, and the Pearson's correlation coefficient between data and prediction was 0.45, which was not statistically significant ($p = 0.092$, $t(517) = 3.1$).

There are three issues to examine before accepting the results. The first is whether the results depend on the choice of the machine learning methods, the second is whether they depend on the selection of time windows and the third is whether they depend on individual variations. First, we used three different models other than LightGBM as a comparison: Support Vector Regression (SVR), Multilayer perceptron (MLP), and Linear Regression. The results showed that accuracy of lightGBM is similar to that of SVR, which is better than MLP and Linear Regression (Fig.4), and that the time required to analyze was the shortest for lightGBM among the four methods.

Second, there is no theoretical reason to select a certain period of time for facial feature extraction to estimate RTs. We used 1 and 5 second windows in addition to 3 second windows to see the effect of the time on the analysis. The results are similar for the three cases (fig. 5). A t-test of RMSE of the model prediction with one showed statistical significance both for 1- and 5-second windows ($p < 0.001$, $t(14) = 15.32$ for 1s and $p < 0.001$, $t(14) = 18.08$ for 5s).

Third, we tested whether a model built with other individuals' data (across individual models) can predict the data of another individual. Figure 6 shows the results of the predictions. Surprisingly, the results revealed no successful prediction across participants. Thus, a model that was based on a group of individuals could not be used to predict the attention level of an individual in the group. The face information related to attention appeared to vary from participant to participant.

V. DISCUSSION

In the current study, we measured RT to task-irrelevant targets as an index of attentional level. With the RTs, we developed a method for predicting engagement to video lectures using a machine learning technique. Our approach was to predict the response time under the assumption that the response time would become longer when more attention was paid to the lecture, reducing attention to a target that was irrelevant to the lecture. The model built for the prediction provided information about the facial features that contributed most to the prediction, which were as follows: AU9 (nose wrinkler), AU45 (blink), AU15 (lip corner depressor), and AU7 (lid tightener). Here, we discuss possible explanations for the importance of these factors in predicting RTs. AU9 was negatively related to RT. Longer RT was associated with more attention to the lecture, suggesting that AU9 was negatively related to the amount of attention paid to the lecture. Increased nose wrinkling was associated with deviation of attention from the lecture. On the hand, the results suggest that AU45 (blink), AU15 (lip corner depressor), and AU7 (lid tightener) were positively related to level of attention paid to the lecture. Thus, more depression of the lip corner, more frequent blinking, and more tightening of the eyelids are expected when a person pays more attention to lectures. Lip corner depression may be related to situations in which a learner has difficulty in understanding the lecture. This may lead the learner to try and attend more to lectures, and to exhibit a serious facial expression. Tightening the eyelids and blinking are similar facial actions, and both may be related to making an effort to understand the content of lectures by opening the eyes wider. However, more blinks and tightening eyelids may also be related to sleepiness. When a person is sleepy, they would be likely to not attend either to lectures or to any task-irrelevant stimuli, which would result in longer RT to the target even without a high level of attention being paid to the lecture. Although the present experiment was designed assuming only two attention states, attending to the lecture or to the task irrelevant target, attention level could potentially be reduced by sleepiness, resulting in longer RTs for the target with decreased attention to the lecture. We attempted to estimate the effect of sleepiness during lectures and re-analyzed the data.

To exclude the possible influence of sleepiness on the results, we re-analyzed the data after removing data with sleepy faces. To identify times at which a participant appeared

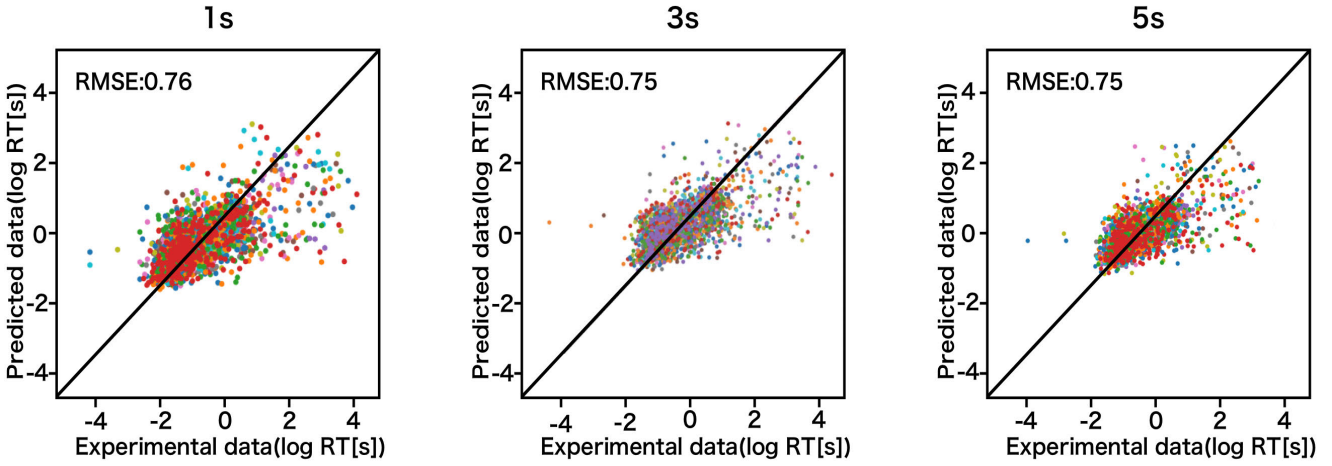

FIGURE 5. We applied 1, 3 and 5 second time windows to see the effect of the time to analyze facial features.

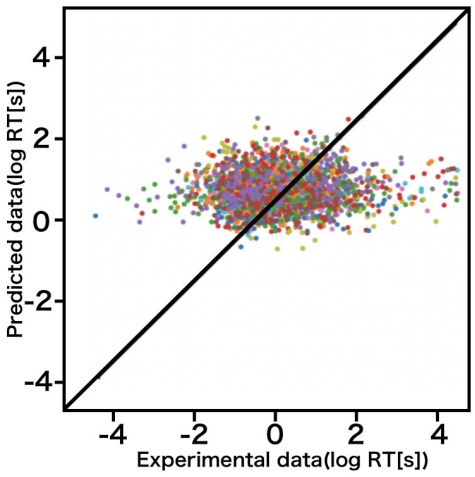

FIGURE 6. Correlation between measured and predicted RTs, using the across individual test model. Configurations are the same as in Fig. 3 (a) .

to be sleepy, we used eye movement data and subjective evaluation of sleepiness in videos. A previous study reported that the eyes become stationary when sleepy [28]. We attempted to detect when learners were sleepy using the gaze data. We calculated the standard deviation of gaze positions, obtained through OpenFace analysis, for 3 sec before each target presentation. The histogram in Fig. 7 (a) shows the distribution of standard deviation of gaze locations, which reflects eye movement activity. The horizontal axis shows the logarithmic scale of the visual angle in radians, showing data with small values clearly. The distribution results can be described as standard deviation values following a single peak distribution with a peak at approximately -0.75 in log deg. However, there appeared to be a peak at very small values at approximately -1.34 in log deg. The eye movements for the video images that were judged subjectively as sleepy exhibited a standard deviation less than -1.13 log deg. Thus, we defined the video faces with standard deviation of gaze location smaller than -1.13 log deg as faces that reflected sleepiness. Note that this analysis is not based on accurate eye movement measurements, but on rough estimation by image processing using OpenFace, by which we estimated that the spatial resolution was higher than 2 radians. Despite the low precision of this method, gaze stability could be evaluated on the basis of the distribution shown in Fig. 7.

We re-analyzed the data after removing data associated with sleepiness using a threshold of the standard deviation of gaze location lower than -1.13 log deg. The results without sleepy faces revealed that the RMSE of the predictions was 0.77 (see Fig. 7 b), which was smaller than the baseline RMSE of 1.0. The Pearson's correlation coefficient between data and prediction was 0.67. and the correlation was statistically significant ($p < 0.001$, $t(2298) = 11$), we also used a statistical test of RMSE to show that our prediction is better than chance. We compared RMSE of the model prediction and that of data, which is one after Z-scoring, using a t-test ($p < 0.001$, $t(14) = 15.07$). These results confirm that facial expressions can be used to predict attention states while watching a lecture. Figure 7 (c) shows the contribution level of each AU to the prediction by SHAP. Similar to the original analysis (Fig. 3), AU9 (nose wrinkler) was found to be the largest contributor, and AU45 (blink) and AU15 (lip corner depressor) were the second and third largest contributors, respectively. However, AU7 (lid tightener), which was in the top five contributors in the original analysis, was no longer included in the top five. These results indicate that wrinkling the nose, blinking, and depressing the lip corners are major factors in predicting attention to lectures.

Individual differences in the relationship between internal concentration state and facial expression, which have not been captured in previous studies that used subjective ratings [14], [15], [16], were found in the present study. We consider several possible reasons for these results. One possibility is that individual identification affected the

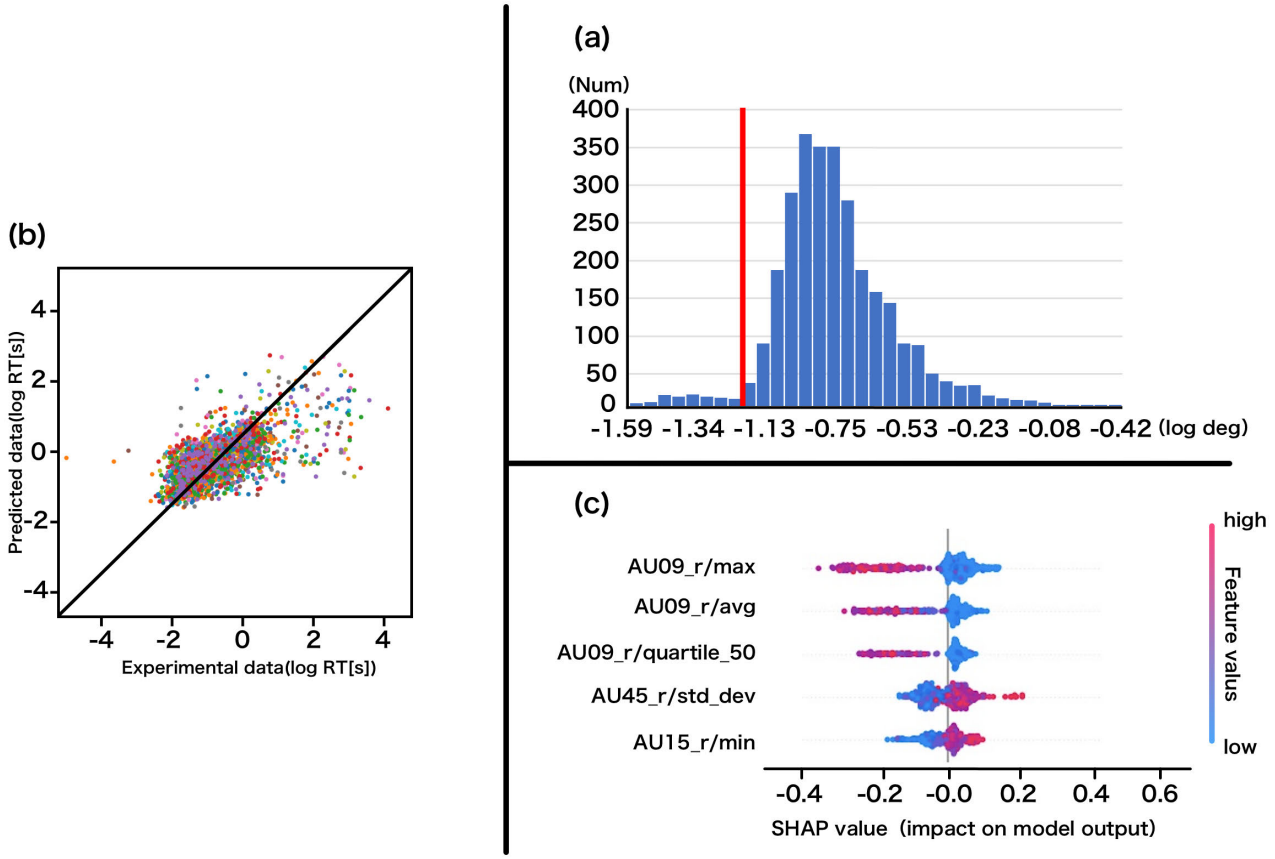

FIGURE 7. (a) Histogram of standard deviation of gaze movements (gaze SD). The gaze SDs before target presentations with sleepy faces estimated subjectively were smaller than the red line, and we assumed that RTs with gaze SDs larger than the red line were not influenced by sleepiness. (b) Correlation between measured and predicted RTs for data without the influence of sleepiness. Configurations are the same as those in Fig. 3 (a). (c) Indexes are arranged according to the level of contribution to the prediction obtained using SHAP. Configurations are the same as those in Fig. 3 (b).

findings. Because AUs themselves may contain information about the facial features of individual participants, the AU analysis might identify individuals. If there is substantial individual variation in RTs in the present experiment, identification of individuals by facial features could potentially predict RT results with some level of accuracy because there is a correlation between facial features and RT for individuals. However, because we used normalized values of RTs and AUs for each participant, the averages of each parameter did not exhibit any correlations among the parameters. In other words, the individual differences we found could be explained by individual differences in contributions of facial features to RT estimation.

To investigate the effects of individual variability, we first conducted the same analyses for data from each participant. Because the amount of data for each participant is relatively small, we performed a 5-fold (instead of 15-fold) cross-validation analysis on each participant's own data. The average RMSE of the prediction against the data for all participants was very close to baseline 1.01 (Fig. 8 c). The RMSE is as poor as the that across individual models (shown in Fig. 6), likely because of the small amount of data used for each model even with 5-fold cross-validation. We, then, examined the effect of the size of the data set on prediction accuracy, and we found that approximately 20% of all data were required to obtain a training effect with RMSE of about 0.8 (Fig. 8 f). To keep the proportion of the data set larger than 20%, we compared the predictions between within and across participants using data sets of three or five groups of participants, instead of datasets of individual participants. Better predictions in the within-group analysis compared with those in the across-group analysis were expected if there were large individual variations in feature expressions related to engagement with the lecture. In the case of three-group division, two of three groups were used for training and the third group was used for a test for across group analysis, while four groups were used for training with the fifth one as a test in the case of five-group division. For within-group analysis, data were divided into three or five sets, selecting equal number of data from each group (each data set had one third of first group, one third of the second group and one third of the third group in the case of three groups). These three or five datasets were used for three- or five-fold validation testing.

Figure 8 shows the results of both within- and across-group analyses for the three and five groups in addition to the

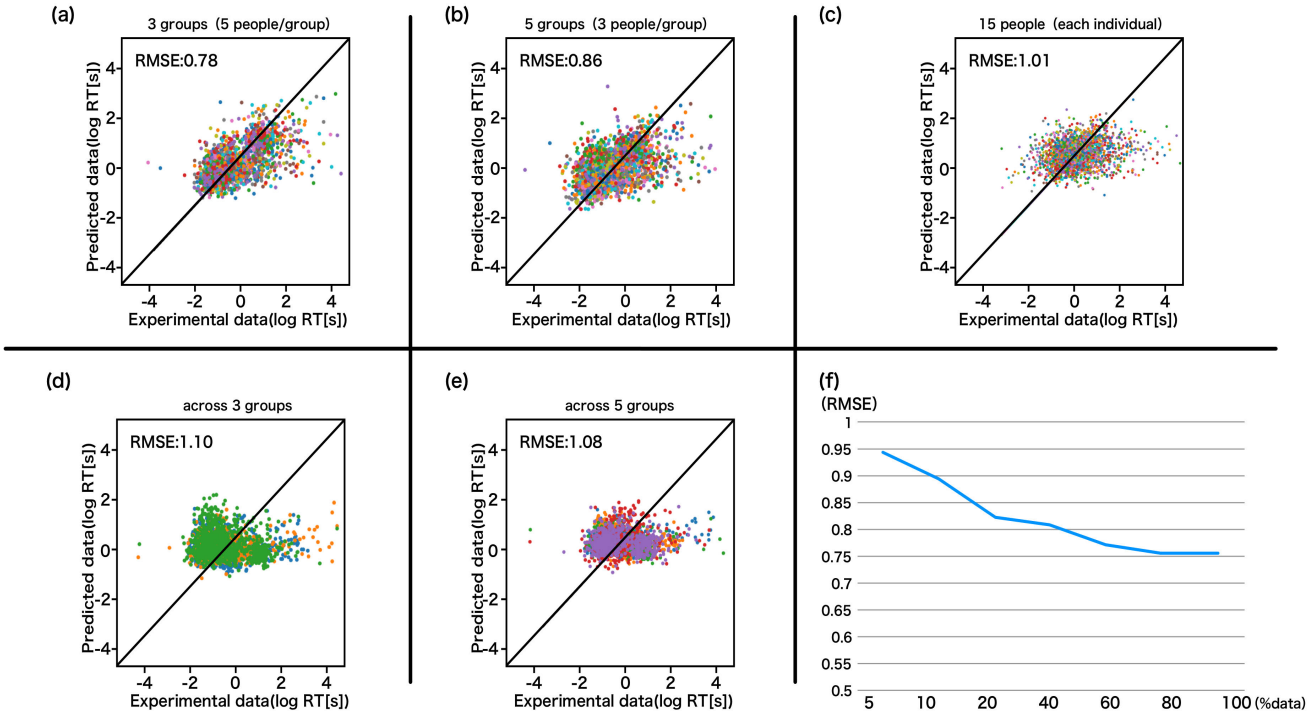

FIGURE 8. (a) the results of the three within groups, (b) five within groups, (c) each individual, (d) three across groups and (e) five across groups. (f) the prediction performance as a function of the data size.

averaged individual predictions. The prediction accuracy was better for within-group analyses compared with that for across-group analyses. RMSE values were 0.86 and 0.78 for three and five within-group analyses, respectively, and 1.10 and 1.08 for three and five across-group analyses, respectively. A t-test of RMSE of the model prediction with one showed statistical significance both for 3- and 5-groups ($p < 0.001$, $t(14) = 7.11$ for 3-group and $p < 0.001$, $t(14) = 11.19$ for 5-group). These results indicate nontrivial individual variations in the relationship between facial expressions and engagement. These variations do not mean that there is no common factor shared by some individuals because pooling data from many participants was shown to improve the prediction (compare Fig. 7 and 8). SHAP values for the three- and five-group analyses showed that AU9 and AU2 were among the best five features in both groups. AU9 was also included in the original analysis with all data. This result suggests that these features are important for all individuals, while other features that differ substantially across individuals could impair the across-participant predictions. Although individual variation limits to use the model without doubt, it is possible to construct a model for a group of individuals with similar properties.

The results suggest that individual variation is substantial, and appears to be a disadvantage in general when the present technique is applied to a supporting system, using a model trained with different individuals. However, the model can be customized to each individual and models constructed for particular individuals may be more precise. Although individual variation should be investigated further to understand the essential factors, the technique developed here can be used for applications in actual education conditions.

Although psychophysical studies used sound stimuli as a probe to measure attention level [30], [31], [32], such approach is not practical in the actual learning situations. Therefore, we investigated whether facial images are sufficient to provide indexes of attention level. The model performance depends on OpenFace performance. Although Baltrusaitis et al. [33] reported that the accuracy of the OpenFace is better than other methods, it is obvious that its performance is not perfect, and it depends on recording conditions of faces. Our estimation of RT from AUs, therefore, includes estimation errors of facial features at a certain amount. We believe that this analysis is useful to obtain information of a learner's conditions (mental states) at each time to make appropriate feedbacks. For example, 70% of correct detection of less attention to a lecture should be useful to provide a warning signal to the leaners and/or the lecturer. Three times of erroneous warnings out of ten should not be problem if the warning signal used does not disturb the class much. Also, the detection rate becomes higher than 99% if there are more than five learners who loose attention to the class even that is 70% for one learner.

VI. CONCLUSION

In conclusion, we revealed that facial expressions can be used to predict learners' level of attention to video lectures, which serves as an index of student engagement. Facial features captured by a video camera can predict reaction times (RTs), which are assumed to be indicative of attentional states. Specific facial features, such as nose wrinkling, blinking, and lip corner depression, appear to be associated with attention during video lectures. The application of facial expression technology has the potential to enhance the quality of teaching. However, before implementing it in actual teaching conditions, a few considerations should be taken into account. Firstly, the underlying mechanisms behind the contributions of these features are not yet understood, which is essential for generalization. Secondly, significant individual differences have been observed. Customizing the model may be one possible solution. In future research, we will focus on exploring individual differences and the physiological relationship between engagement and facial expressions during learning.

REFERENCES

[1] S. Shioiri, Y. Sato, Y. Horaguchi, H. Muraoka, and M. Nihei, "Quali-informatics in the society with Yotta scale data," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2021, pp. 1–4.

[2] Y. Sato, Y. Horaguchi, L. Vanel, and S. Shioiri, "Prediction of image preferences from spontaneous facial expressions," *Interdiscipl. Inf. Sci.*, vol. 28, no. 1, pp. 45–53, 2022.

[3] Y. Horaguchi, Y. Sato, and S. Shioiri, "Estimation of preferences to images by facial expression analysis," *IEICE Tech. Rep.*, vol. 120, no. 306, pp. 71–76, 2020.

[4] C. Thomas and D. B. Jayagopi, "Predicting student engagement in classrooms using facial behavioral cues," in *Proc. 1st ACM SIGCHI Int. Workshop Multimodal Interact. Educ.*, Glasgow, U.K., Nov. 2017, pp. 33–40.

[5] N. K. Mehta, S. S. Prasad, S. Saurav, R. Saini, and S. Singh, "Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement," *Appl. Intell.*, vol. 52, no. 12, pp. 13803–13823, 2022, doi: 10.1007/s10489-022-03200-4.

[6] D. K. Darnell and P. A. Krieg, "Student engagement, assessed using heart rate, shows no reset following active learning sessions in lectures," *PLoS ONE*, vol. 14, no. 12, Dec. 2019, Art. no. e0225709, doi: 10.1371/journal.pone.0225709.

[7] D. M. Bunce, E. A. Flens, and K. Y. Neiles, "How long can students pay attention in class? A study of student attention decline using clickers," *J. Chem. Educ.*, vol. 87, no. 12, pp. 1438–1443, Dec. 2010, doi: 10.1021/ed100409p.

[8] H. Kato, K. Takahashi, Y. Hatori, Y. Sato, and S. Shioiri, "Prediction of engagement from temporal changes in facial expression," in *Proc. World Conf. Comput. Educ.*, Hiroshima, Japan, Aug. 2022.

[9] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 1, pp. 50–69, Jan. 2010.

[10] A. M. Leiker, A. T. Bruzi, M. W. Miller, M. Nelson, R. Wegman, and K. R. Lohse, "The effects of autonomous difficulty selection on engagement, motivation, and learning in a motion-controlled video game task," *Hum. Movement Sci.*, vol. 49, pp. 326–335, Oct. 2016, doi: 10.1016/j.humov.2016.08.005.

[11] L. S. Pagani, C. Fitzpatrick, and S. Parent, "Relating kindergarten attention to subsequent developmental pathways of classroom engagement in elementary school," *J. Abnormal Child Psychol.*, vol. 40, no. 5, pp. 715–725, Jul. 2012, doi: 10.1007/s10802-011-9605-4.

[12] M. N. Nguyen, S. Watanabe-Galloway, J. L. Hill, M. Siahpush, M. K. Tibbits, and C. Wichman, "Ecological model of school engagement and attention-deficit/hyperactivity disorder in school-aged children," *Eur. Child Adolescent Psychiatry*, vol. 28, no. 6, pp. 795–805, Jun. 2019, doi: 10.1007/s00787-018-1248-3.

[13] M. Kinnealey, B. Pfeiffer, J. Miller, C. Roan, R. Shoener, and M. L. Ellner, "Effect of classroom modification on attention and engagement of students with autism or dyspraxia," *Amer. J. Occupational Therapy*, vol. 66, no. 5, pp. 511–519, 2012, doi: 10.5014/ajot.2012.004010.

[14] Ö. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1012–1027, Apr./Jun. 2023, doi: 10.1109/TAFFC.2021.3127692.

[15] 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 Trans. Affect. Comput.*, vol. 8, no. 1, pp. 15–28, Jan. 2017, doi: 10.1109/TAFFC.2016.2515084.

[16] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, Jan. 2014, doi: 10.1109/TAFFC.2014.2316163.

[17] R. Miao, H. Kato, Y. Hatori, Y. Sato, and S. Shioiri, "Analysis of facial expressions for the estimation of concentration on online lectures," in *Proc. World Conf. Comput. Educ.*, Hiroshima, Japan, Aug. 2022.

[18] A. F. Kramer, L. J. Trejo, and D. Humphrey, "Assessment of mental workload with task-irrelevant auditory probes," *Biol. Psychol.*, vol. 40, nos. 1–2, pp. 83–100, May 1995.

[19] A. Pfefferbaum, J. M. Ford, W. T. Roth, and B. S. Kopell, "Age differences in P3-reaction time associations," *Electroencephalogr. Clinical Neurophysiol.*, vol. 49, pp. 257–265, Aug. 1980, doi: 10.1016/0013-4694(80)90220-5.

[20] M. I. Posner, "Orienting of attention," *Quart. J. Exp. Psychol.*, vol. 32, pp. 3–25, Feb. 1980.

[21] S. A. Hillyard, R. F. Hink, V. L. Schwent, and T. W. Picton, "Electrical signs of selective attention in the human brain," *Science*, vol. 182, no. 4108, pp. 177–180, Oct. 1973, doi: 10.1126/science.182.4108.177.

[22] S. Shioiri, M. Ogawa, H. Yaguchi, and P. Cavanagh, "Attentional facilitation of detection of flicker on moving objects," *J. Vis.*, vol. 15, no. 14, p. 3, Oct. 2015, doi: 10.1167/15.14.3.

[23] S. Shioiri, H. Honjyo, Y. Kashiwase, K. Matsumiya, and I. Kuriki, "Visual attention spreads broadly but selects information locally," *Sci. Rep.*, vol. 6, no. 1, p. 35513, Oct. 2016, doi: 10.1038/srep35513.

[24] (2023). *@Fuku-Programming*. Accessed: Feb. 14, 2023. [Online]. Available: https://www.youtube.com/watch?v=uVaOzQLxXt0

[25] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[26] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. San Francisco, CA, USA: Consulting Psychologists Press, 1978.

[27] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–9.

[28] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–10.

[29] T. Abe, T. Nonomura, Y. Komada, S. Asaoka, T. Sasai, A. Ueno, and Y. Inoue, "Detecting deteriorated vigilance using percentage of eyelid closure time during behavioral maintenance of wakefulness tests," *Int. J. Psychophysiol.*, vol. 82, no. 3, pp. 269–274, Dec. 2011, doi: 10.1016/j.ijpsycho.2011.09.012.

[30] M. A. Bedard, F. El Massioui, B. Pillon, and J. L. Nandrino, "Time for reorienting of attention: A premotor hypothesis of the underlying mechanism," *Neuropsychologia*, vol. 31, no. 3, pp. 241–249, Mar. 1993, doi: 10.1016/0028-3932(93)90088-h.

[31] G. Rhodes, "Auditory attention and the representation of spatial information," *Perception Psychophys.*, vol. 42, no. 1, pp. 1–14, Jan. 1987, doi: 10.3758/bf03211508.

[32] J. R. Simon, E. Acosta, and S. P. Mewaldt, "Effect of locus of warning tone on auditory choice reaction time," *Memory Cognition*, vol. 3, no. 2, pp. 70–167, Mar. 1975, doi: 10.3758/BF03212893.

[33] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 59–66, doi: 10.1109/FG.2018.00019.

RENJUN MIAO was born in Wenzhou, Zhejiang, in 1986. He received the bachelor's degree in mechanical automation engineering from the Zhejiang University City College, in 2008, and the master's degree in information engineering from Tohoku University, Japan, in 2012, where he is currently pursuing the Ph.D. degree, with a focus on affective computing, mainly on detecting the quality of students' online education through the change of facial expressions.

From 2010 to 2012, his main research focus was on signal processing of color and shape in brain visual neurology. He has been an Engineer since graduation and an Education SAAS Development Supervisor, in 2017.

HARUKA KATO received the B.S. degree in engineering and the M.S. degree in information engineering from Tohoku University, in 2021 and 2023, respectively.

From 2021 to 2023, her main research was affective computing mainly on detecting engagement of student while studying through the change of electroencephalography and facial expressions. Her research interests include engagement, attention while studying, and facial expressions.

YASUHIRO HATORI received the B.S. degree in information engineering and the M.S. and Ph.D. degrees in engineering from the University of Tsukuba, in 2007, 2009, and 2014, respectively.

From 2014 to 2016, he was a Postdoctoral Fellow with the Research Institute of Electrical Communication, Tohoku University. From 2016 to 2018, he was a Postdoctoral Fellow with the National Institute of Advanced Science and Technology. Since 2018, he has been an Assistant Professor with Tohoku University. His research interests include eye movement, visual attention, and multisensory integration.

YOSHIYUKI SATO received the B.S. degree from Kyoto University, in 2004, and the M.S. and Ph.D. degrees from The University of Tokyo, Japan, in 2006 and 2009, respectively.

From 2010 to 2016, he was an Assistant Professor with the University of Electro-Communications, Japan. From 2012 to 2013, he was a Visiting Professor with Northwestern University, USA. From 2016 to 2018, he was a Project Researcher with The University of Tokyo. Since 2018, he has been a specially-appointed Assistant Professor with Tohoku University, Japan. His research interests include mathematical and machine learning modeling of human behaviors, including perception, cognition, attention, motor functions, and communications.

SATOSHI SHIOIRI received the B.S. degree in mechanical engineering and the M.S. and Ph.D. degrees in physical information engineering from the Tokyo Institute of Technology, in 1981, 1983, and 1986, respectively.

From 1986 to 1989, he was a Postdoctoral Fellow with the University of Montreal. From 1989 to 1990, he was a Postdoctoral Fellow with the Advanced Telecommunications Research Institute International, Kyoto. He was an Assistant Professor, an Associate Professor, and a Professor with Chiba University, from 1990 to 2004. He has been a Professor with Tohoku University, since 2004. His research interests include motion perception, depth perception, color perception, visual attention, eye movement, and vision for action.