International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hihc20

Estimating Level of Engagement from Ocular Landmarks

Zeynep Yücel , Serina Koyama , Akito Monden & Mariko Sasakura

To cite this article: Zeynep Yücel , Serina Koyama , Akito Monden & Mariko Sasakura (2020) Estimating Level of Engagement from Ocular Landmarks, International Journal of Human–Computer Interaction, 36:16, 1527-1539, DOI: 10.1080/10447318.2020.1768666

To link to this article: https://doi.org/10.1080/10447318.2020.1768666

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
2020, VOL. 36, NO. 16, 1527–1539
https://doi.org/10.1080/10447318.2020.1768666

Estimating Level of Engagement from Ocular Landmarks

Zeynep Yücel, Serina Koyama, Akito Monden, and Mariko Sasakura

Department of Computer Science, Division of Industrial Innovation Sciences, Okayama University, Okayama, Japan

ABSTRACT

E-learning offers many advantages like being economical, flexible and customizable, but also has challenging aspects such as lack of – social-interaction, which results in contemplation and sense of remoteness. To overcome these and sustain learners' motivation, various stimuli can be incorporated. Nevertheless, such adjustments initially require an assessment of engagement level. In this respect, we propose estimating engagement level from facial landmarks exploiting the facts that (i) perceptual decoupling is promoted by blinking during mentally demanding tasks; (ii) eye strain increases blinking rate, which also scales with task disengagement; (iii) eye aspect ratio is in close connection with attentional state and (iv) users' head position is correlated with their level of involvement. Building empirical models of these actions, we devise a probabilistic estimation framework. Our results indicate that high and low levels of engagement are identified with considerable accuracy, whereas medium levels are inherently more challenging, which is also confirmed by inter-rater agreement of expert coders.

1. Introduction

E-learning has progressed rapidly in the recent years, and became a popular choice of learning medium at schools (Arkorful & Abaidoo, 2015; Haßler et al., 2016) as well as corporate training and life-long learning (Beinicke & Bipp, 2018; Seta et al., 2014). The rapid propagation of e-learning is suggested to be due to a series of reasons including being economical (Piskurich, 2006) customizable (Liu et al., 2017) and rich in content (O'Donnell et al., 2015).

Despite these advantages, learners may experience some difficulties in using e-learning systems. Arkorful and Abaidoo (2015) point out to the lack of social interaction as one of the most important challenges of e-learning. Namely, lack of social interaction implies a bigger burden of motivation and requires better time management skills for overcoming contemplation and remoteness. In educational psychology, such commitment and involvement of students in learning activity is termed – in the broad sense – as "engagement" (Fredricks et al., 2004). In e-learning, where students often feel isolated and disconnected, sustaining learners' engagement arises as a serious challenge (Dixson, 2015). In particular, any effective design scheme or intervention to an e-learning session for enhancing engagement initially requires assessment of users' state.

For this purpose, this study proposes using visual feedback from users to estimate automatically their level of engagement. Namely, we focus on users' frontal view video footage and search for indications of decline in level of engagement on face images. These indications are chosen on the basis of the findings of several studies in the fields of cognitive science, affective computing, eye physiology, etc.

In particular, we exploit the following findings. Smilek et al. (2010) ascertain that perceptual decoupling from external stimuli is promoted by blinking during mental/cognitive fatigue. Besides, it is known that eye strain due to prolonged exposure to digital displays increase blinking rate (Rosenfield, 2011). Interestingly, Matthews and Desmond (1998) also establish the connection between blinking and visual fatigue due to increased task load. Namely, suppression of blinks due to increased task load eventually causes drying of the eyes and leads to a higher blinking rate. In addition to blinking, the aspect ratio of the eye is also a prominent factor in detection of subject's attentional state (Ji et al., 2004). Moreover, Asteriadis et al. (2011) show that users' head pose and position present significant correlation to their level of involvement in computer-based tasks.

Based on these findings, we propose several features, which are expected to provide relevant information on users' level of engagement. Subsequently, we confirm that there is a significant correlation between these features and level of engagement coded by professional teachers. By building a probabilistic method based on the empirical observations of such feature distributions, we prove that level of engagement can be estimated from video footage with significant accuracy.

The proposed approach has several advantages. First of all, it can be integrated into e-learning systems so as to provide continuous on-the-fly assessment of engagement. Therefore, it potentially enables stimulation of the user immediately upon detection of a decline in level of engagement by, for instance, providing motivational advice or interactive content (e.g., with an avatar) with possible adjustments to interpersonal variations

CONTACT Zeynep Yücel zeynep@okayama-u.ac.jp Department of Computer Science, Division of Industrial Innovation Sciences, Okayama University, 3-1-1 Tsushima-naka Kita-ku, Okayama 700-8530, Japan

in behavior. In addition to applications involving individual users, the proposed method has also the potential of being deployed in collaborative learning scenarios, where the socially shared regulation of learning poses several coordination problems, one particular aspect relating communication of the shared understanding through joint attention. Although this issue is often treated under the assumption that directing gaze to a target indicates sustained visual attention (Yücel et al., 2013), recent studies ascertain that such an approach oversimplifies human attentional process. In this respect, our study offers a potential to detect the declines in engagement level of a partner in collaborative learning.

2. Background and related work

In their seminal work, Fredricks et al. (2004) define "engagement" in broad terms as active commitment, willing participation and involvement of students in school activity. In particular, engagement is suggested to be a multidimensional phenomenon, governed by three fundamental elements as behavior, cognition, and emotion (Fredricks et al., 2016).

Behavioral engagement relates attendance, participation, completion of assignments, etc., in conventional classroom settings. On the other hand, in technology-mediated learning, behavioral engagement is quantified in terms of computer-recorded indicators such as frequency of logins, number and frequency of responses/views, time spent online and number of accessed resources (e.g., podcasts or screencasts). Obviously such indicators are quite easy to access in computer medium (Ben-Zadok et al., 2011). However, as pointed out by Arkorful and Abaidoo (2015), there are a number of issues with employing such metrics, one of the most important being the difficulty of incorporating them on-the-fly, i.e. in a live session.

Cognitive engagement is the focused effort to effectively understand the lesson and involves students' cognitive strategy/planning, and self-regulation. In this respect, unlike behavioral engagement, cognitive engagement may not always be observed or assessed in a quantitative way and require self-reporting (Greene, 2015). However, self-reports or questionnaires suffer from the varying implicit standards of respondents, heterogeneous frame of reference, social desirability bias, and memory call limitations (D'Mello et al., 2017). For this reason, there have been some attempts to quantify cognitive engagement fusing with behavioral indicators (e.g., time on task) or gauging cognitive processes such as reflection, interpretation, synthesis, or elaboration. However, it is shown that achieving a clear distinction between behavioral and cognitive elements of engagement is quite complex (Kong, 2011). Nevertheless, the interplay between cognitive engagement and various behaviors/actions is evident. For instance, numerous studies have shown that a large set of postural processes/behaviors is in close relation with cognitive engagement (Balaban et al., 2004; Bonnet & Baudry, 2016; Bonnet et al., 2017; Hunter & Hoffman, 2001), while others call attention to the interplay with eye movements (Ballenghein & Baccino, 2019; Kaakinen et al., 2018; Miller, 2015).

Emotional engagement includes positive or negative emotions toward learning, classmates, or instructors, etc., and may be observed through visible expressions (Henrie et al., 2015). D'Mello et al. (2017) discuss on the significant capabilities of advanced, analytical, and automated methods in measuring engagement. In particular, they point out to the potential of facial features, body movements, posture, eye gaze, and contextual cues, which relate the emotional dimension of engagement.[1] Since emotional engagement is more flexible to incorporate than behavioral engagement and more feasible to detect than cognitive engagement, it is often used in automatic assessment of e-learning users' state and various promising results have been reported over the years regarding emotion-aware interventions (Aslan et al., 2018; Eliot & Hirumi, 2019). Nevertheless, there is no a universal consensus on a concrete definition of (each aspect of) engagement, neither is there a well-accepted method of an effective measurement (Sinatra et al., 2015). In what follows, we list several works employing visual cues in estimating users' state/level of engagement.

Whitehill et al. use frontal videos of users cropped at varying time scales, which are classified at a frame-by-frame basis into two classes as high and low engagement through several feature and classifier combinations (e.g., Gabor features and SVM) (Whitehill et al., 2014). In that sense, they explore the possibility of using low level and abstract features (i.e., in no direct relation with the cause or effect of disengagement), and they make an effort to achieve comparable performance (in estimation) to human expert labeling.

Another very recent and interesting engagement estimation study aimed at technology-mediated learning belongs to Bosch and D'Mello, where upper-body movement, head pose, facial textures, facial action units and their temporal dynamics are employed in conjunction with support vector machines and deep neural networks for detecting *mind wandering* (Bosch & D'Mello, 2019; Smilek et al., 2010), which is some form of disengagement, where attention shifts from the immediate external environment to internal trains of thought (Killingsworth & Gilbert, 2010).

In mind wandering, "blinks" are suggested to have a particular role. In particular, spontaneous blinks are shown to help humans disengage from the outside stimuli in favor of the internal processing in two ways, (i) by physically closing the eyelid and thus interrupting the visual stimuli; and (ii) by applying a cortical suppression before and after lid closure (Smilek et al., 2010). In this respect, blinks can be considered as a particular means for the embodiment of "mind wandering" (Schooler et al., 2011).

In addition to these findings in universal settings, blinking is particularly important in human-computer interaction, where the interaction interface is in most cases an LCD display. Due to the intensity and frequency of the emitted light, users may feel visual fatigue (or eye strain) over a certain duration of time, which may affect their blinking pattern as well (Rosenfield, 2011). In addition to these physiological reasons, certain cognitive factors may result in similar consequences (i.e., increased blinking rate) in computerized tasks requiring mental workload. Namely, it is shown that task disengagement scales substantially

correlated with aspects of visual fatigue (Matthews & Desmond, 1998). In particular, blinks are suppressed in response to increased visual workload and this inhibition in turn may cause drying of the eyes followed by a higher blinking rate. In addition to blinking, the state of the eye can also be represented using percentage of eyelid closure (PERCLOS) to detect alertness or drowsiness. This kind of marker is particularly popular in driver monitoring systems and has been used in numerous works over the years (Ji et al., 2004; Mbouna et al., 2013).

3. Experiments and data set

The outline of experimentation and compilation of ground truth are as follows. We designed three tasks as explained in Section 3.1 and asked several participants to carry out each task for a time long enough to observe a wide spectrum of engagement levels. During experiments, we collected video data of participants' face, head, and upper torso. As detailed in Section 3.2, we asked two licensed teachers to label these videos and annotate their (ground truth) engagement levels. On these annotations, we carried out an inter-rater agreement analysis and verified that the assigned labels are sufficiently reliable as illustrated in Section 3.3. In what follows, we present the details of each of these steps.

3.1. Experiment tasks

We designed three sorts of tasks such that each one requires a different skill. However, denomination of the tasks is specified by the required level of user involvement. In particular, we use the terms *passive, semi-active*, and *active* (Koyama et al., 2019).

Specifically, in the passive task, the participants watch a slide show of images, which are selected from a benchmark saliency data set (Borji & Itti, 2015). The subjects are told that they will be given a memory test afterward as a motivation to attend the images but no test is given. This task is considered to be similar to passive online learning in being linear and straightforward.

In the semi-active task, the participants listen to the narration of a story in English accompanied by illustrations, requiring listening comprehension and inference skills.[2] At the end of each story, a multiple choice question with a single keyed answer is displayed for a limited duration.[3] We consider the narration part of this task to be passive, where the participants need to comprehend the information, and the subsequent test to be active, which requires reasoning, deduction, and inference (Mayer, 2017).

The active task is Wisconsin card sorting (Heaton et al., 1993), which is a common tool in neuropsychology for examining the functioning of the frontal lobe. The test requires users to match a stimulus with one of the four options based on an undisclosed rule, which changes at uneven steps so that the participants need to discover the new rule by trial and error, entailing the necessity of keeping continuous focus on previous and subsequent rules as well as refuted ones (see (Yücel, 2020) for implementation details). This task requires strategic planning and organized search skills, as well as utilization of feedback, modulation of impulsive response, and directing behavior toward a goal.

Since the three tasks require different skills, we expect to observe a wide spectrum of engagement levels. Moreover, for eliminating individual behavioral variations, we asked five people[4] to perform all three tasks.[5] We call the specific implementation of a certain task carried out by a certain participant *a session*. Thereby, we implemented a total of 15 sessions, where each session takes approximately 2 hours. In this manner, we gathered a data set composed of roughly 30 hours of video recordings.

3.2. Coding process and ground truth

At each session, we used a notebook computer for presenting the task to the participant and recording his/her behavior. Namely, we collected a video footage depicting the face, head, and upper torso using the built-in webcam of the computer (see Figure 1).[6]

The ground truth for the level of engagement is obtained by manual annotation of video footage by licensed (and practicing) teachers. At this point, we would like to address several works on teachers' reliability regarding assessment of learners' affective states. In particular, inter-rater reliability is shown to be higher within teachers than between teacher-learner pairs (X. Wu et al., 2013). In addition, teachers are considered to estimate certain affective states such as confusion better than others (e.g., boredom) (Graesser & D'Mello, 2012), independent of the rating method (Urhahne & Zhu, 2015; Zhu & Urhahne, 2014). Since teachers are shown to have consistent judgments between each other and their level of judgment does not depend on practical specifics of coding, we consider them to be a reliable source of ground truth.

Nevertheless, the substantial duration of the footage makes it hard for the teachers to view and label the entire data. A common approach in dealing with this issue is to crop segments of the video footage (henceforth referred as *clips*), which are expected to involve representative cues on the mental/cognitive state of a participant within a certain period (Thomas & Jayagopi, 2017). By this means, the amount of coding work is reduced to a reasonable level and the labeling process becomes more affordable for the annotators. In addition, the set of clips summarizes the entire term of the

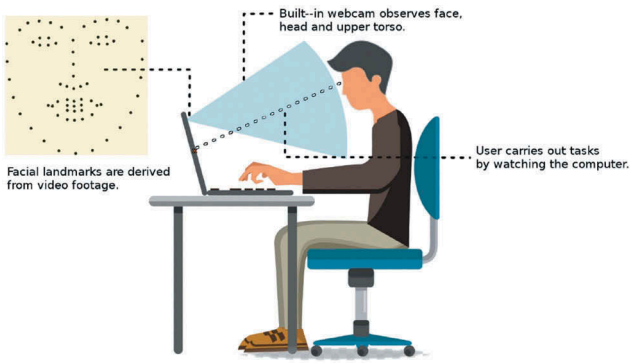

Figure 1. Experimental setup. User's face, head and upper torso are observed using the built-in webcam, as he/she performs tasks on the computer. Facial landmarks are derived from the video footage.

experiments and provides a comprehensive overview of participants' behavior.

Specifically, we cropped 15 clips from the raw footage of a single session. To that end, we defined a set of time instants, which describe the initial time (or equivalently initial frame number) of every clip.[7] We then cropped the portion of the raw footage starting from each element of this set for a duration of 10 sec. This time scale is regarded as reassuring by (Whitehill et al., 2014), since level of engagement is not subject to a serious variation throughout such observation window.

Two licensed teachers (henceforth referred as *coders*) evaluated the clips according to the apparent level of engagement of the participants. The coders practice regularly teaching foreign languages and have vast experience of interacting with students in conventional classroom settings. In that respect, they are not provided any guidelines or sets of learner actions/behaviors for judging the level of engagement, but are rather asked to decide relying on their experience and feelings without overthinking. Specifically, they assign each clip an engagement label e on a Likert scale from 1 to 5, where $e = 1$ denotes "disengaged" and $e = 5$ represents "engaged."[8] As a result of the coding process, a total of 189 clips are labeled by each of the two expert coders and considered in the forthcoming analysis on estimation of engagement.[9]

3.3. Assessment of inter-rater agreement

In order to judge the consensus of the expert coders in the evaluation of engagement levels, we examined the assigned labels in a qualitative manner as well as in a quantitative manner. In qualitative analysis, we considered the dependence of assigned labels on two factors: (i) the type of task (i.e., passive, semi-active, or active) and (ii) assigned labels.

By examining the distribution of e with respect to task type, the coders are observed to agree that the active task is performed with a relatively higher rate of engagement in general, followed by semi-active and passive tasks.

By examining the dependence of engagement on assigned labels, it is observed that for extremities (i.e., "disengaged" and "engaged") the coders very often agree on *side* of the spectrum. Namely, if one coder labels a certain clip as "engaged" ($e = 5$), then even when the other coder does not assign the exact same label, he/she often labels the same clip as "moderately engaged" ($e = 4$) or fairly engaged ($e = 3$), but only very rarely as "poorly engaged" ($e = 2$) or "disengaged" ($e = 1$).

Subsequent to these qualitative observations, we carried out a methodological analysis quantifying the degree of agreement between the coders. In literature, various statistical methods are available for analyzing such inter-rater agreement. Most methods consider "agreement" as improvement in joint probability over expected agreement due to chance. Some commonly adopted metrics such as Cohen's κ coefficient or Fleiss' κ coefficient apply only to qualitative (i.e., nominal) items and thus are not feasible for our case, where the labels have a gradual relationship (i.e., increasing progressively from "disengaged" to "engaged") (Cohen, 1960; Fleiss, 1971). Therefore, in this study we choose using Krippendorff's α coefficient, which is a powerful metric applicable to labels from various measurement scales (e.g., ordinal and interval) (Krippendorff, 2004). Since our labels have a ranked relation, we computed Krippendorff's α for ordinal variables and found an inter-rater agreement rate of $\alpha = 0.78$, which is considered to be sufficient based on the remarks of Krippendorff (2004). Having coders' agreement confirmed, we arbitrarily chose one coder and based our computations on her labels.

4. Method

The outline of the proposed approach is as follows. Based on the ocular landmarks derived from the videos in Section 4.1, we define a set of features in Section 4.2, which are shown in Section 4.3 to be in close relation with coded levels of engagement. Subsequently, the probability density functions of these features are constructed based on a kernel density estimation scheme in Section 4.4. By ascertaining the independent nature of the features in Section 4.5, the distributions presented in Section 4.6 are built. Finally, from a given set of observations, engagement level is estimated within a probabilistic framework as explained in Section 4.7.

4.1. Estimation of landmarks and detection of blinks

Facial landmarks are the set of points marking the locations (or boundaries) of facial components such as eyes, nose, jawline, etc., (Wang et al., 2018). Over the years, numerous landmark estimation methods have been proposed. These methods consider a variety of templates (or maps), which potentially involve a different number of markers (see Figure 2a for a sample template). However, essentially the same facial features are addressed by all templates. In other words, varying templates describe similar facial features, actions, or expressions at varying resolutions. For instance, W. Wu et al. (2018) consider a total of 98 landmarks, whereas Uřičář et al. (2016) consider 68 landmarks and Kasinski et al. (2008) consider 30 landmarks. Since this study focuses on the eyes, we take a closer look at the landmarks describing the eyes (henceforth referred as *ocular landmarks*) in the aforementioned studies.[10] We notice that in all three studies the descriptors of the eyes are constituted by lateral canthus, medial canthus, and palpebral fissure (Neog, 2018).

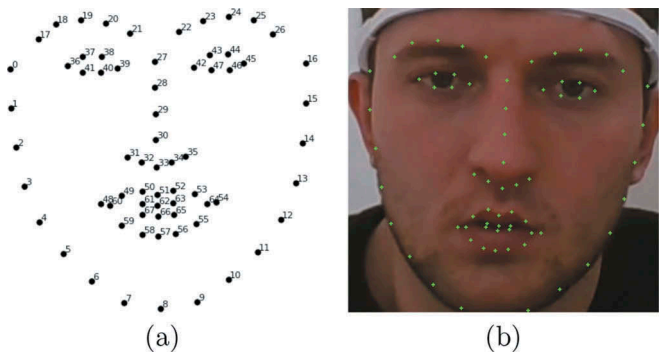

Figure 2. (a) Landmarks detected by Dlib and (b) landmarks on a sample frame.

Basically, lateral and medial canthi are represented by a single marker in all approaches, whereas the number of markers over the palpebral fissure is the varying factor. Since the contraction of the palpebral fissure mostly does not involve any lateral movements and the relative positions of the concerning landmarks are expected to be quite stable and dependent, we consider descriptions of the palpebral fissure using different number of markers to have virtually the same effect in representing the state of the eye, provided that the accuracy of landmark estimation is comparable.

Therefore, in choosing the landmark estimation method and the template entailed by that, we regard speed and accuracy to be the determining factors and opt for the method proposed by Kazemi and Sullivan (2014), which is considered to be the state-of-the-art in terms of both factors (Y. Wu & Ji, 2019). Pursuant to this choice, we employ the Dlib toolbox in deriving the landmark locations (King, 2018), which is based on the principles of Kazemi and Sullivan (2014). In this manner, using a shape predictor pre-trained on the iBUG 300-W data set (Sagonas et al., 2016), a set of 68 landmarks are obtained as seen in Figure 2.

Using these landmarks, we study spontaneous blink patterns (Cruz et al., 2011).[11] In blink detection, we employ the simple and yet powerful real-time blink detection method proposed by Soukupová and Cech (2016). Specifically, this method is based on the changes in aspect ratios of the right and the left eye, r_L and r_R, respectively. Namely, concerning the landmarks in Figure 2a, r_L is defined as,

$$r_L = \frac{|p_{37} - p_{41}| + |p_{38} - p_{40}|}{2|p_{36} - p_{39}|}. \tag{1}$$

Moreover, blinks being symmetric (to the right and the left eyes), it is plausible to use the average of the two eye aspect ratios r_L and r_R in detection of blinks.

Since the time course of -average- eye aspect ratio can be subject to individual variations on the speed of closing and opening, or the degree of squeezing the eye, etc., Soukupová and Cech (2016) account for the effect of these individual variations by training an SVM classifier on sample blinking and non-blinking patterns from several data sets (Duda et al., 2012). In addition, a detailed analysis on the estimation accuracy on two standard eye blink data sets with ground truth annotations (Drutarovsky & Fogelton, 2014; Pan et al., 2007) reveals that this approach yields an almost perfect estimation accuracy.

As result of this detection process, we obtain (i) the locations of the 12 ocular landmarks (for the both eyes) at each frame and (ii) a binary function $b[i]$ which describes the blinking state of the eyes at every frame i with a 0 for blink and 1 for non-blink.

4.2. Derivation of features

Subsequent to the detection of ocular landmarks (see Figure 3) and blinks, we define a set of four features derived thereof as duration of blinks t_b, frequency of blinks f_b, average aspect ratio for open eyes $\bar{r}_o$, and interocular breadth d_{io}. Note that although the current study considers these features as reflections of emotional engagement, several other studies consider similar features to represent directly cognitive engagement, since their tasks and probes are designed to assure a firm relation. (Ballenghein & Baccino, 2019; Kaakinen et al., 2018). In what follows, we describe the features in detail.

The average duration of blinks t_b is considered basically as the ratio of the number of frames at a blinking state to the number of blinks. Using the binary function $b[i]$ defining the state of the eyes at frame i with a 0 for blink and 1 for non-blink, t_b is found as,

$$t_b = \frac{N_f - \sum_i b[i]}{f_b N_f}, \tag{2}$$

where f_b is the frequency of blinks and N_f stands for the total number of frames of the input video.[12] For determining the frequency of blinks f_b, we first count how many times a blink is initialized by the closing of the eye (i.e., blink onset).[13] The frequency of blinks is found as,

$$f_b = \frac{\sum_i H(-b[i+1] + b[i])}{N_f}, \tag{3}$$

where H is the left-continuous Heaviside step function (Abramowitz et al., 1988). Note that, t_b and f_b are particularly beneficial in representing the degree of perceptual decoupling due to blinks (Smilek et al., 2010).

Average aspect ratio $\bar{r}_o$ for open eyes is the average of the aspect ratios of both eyes over the time interval while the eyes are open,

$$\bar{r}_o = \left\{ * \overline{r_{L,R}[i]} b[i] = 1 \right\}. \tag{4}$$

In this sense, $\bar{r}_o$ presents similarity to percentage of eyelid closure, which is a common indicator in identifying drowsiness in driver monitoring systems (Ji et al., 2004).

Interocular breadth d_{io} is in close relation to the depth of the user, i.e., distance between the user and screen, which is a common feature in evaluation of engagement (Asteriadis et al., 2011). In this study, rather than the explicit value of depth, we consider interocular breadth d_{io}, which can be considered to be inversely proportional to the depth of the user. On the landmark map depicted in Figure 2a, d_{io} corresponds to $|p_{39} - p_{42}|$.[14]

Interocular breadth is a *de facto* feature that does not depend on the blinking state, whereas f_b and t_b account for

(a) (b)

Figure 3. Eye regions and corresponding landmarks as eyes are (a) open, and (b) closed.

the time intervals while the eyes are closed due to blinking, whereas the remaining time intervals (i.e., when eyes are at a non-blinking state) are accounted for by $\bar{r}_o$. Henceforth, we denote the set of all features $\{t_b, f_b, \bar{r}_o, d_{io}\}$ with Σ, an arbitrary feature in Σ with σ and a subset of several features with Σ'.

Since duration of the clips (i.e., 10 sec) is chosen so as to enable assuming stability of engagement (Whitehill et al., 2014), a set of features concerning a particular clip can be considered to represent the level of engagement assigned to that clip by the two expert coders.

4.3. Verifying relevance of features

Due to the inferences from previous literature, it is plausible to assume that the features defined in Section 4.2 are expected to be correlated with the subjective evaluations of engagement level e (Asteriadis et al., 2011; Matthews & Desmond, 1998; Rosenfield, 2011; Smilek et al., 2010). In order to confirm that, we first present the descriptive statistics of the variables and then determine in a quantitative manner (through polyserial correlation) the extent of this correlation.

Table 1 reports descriptive statistics for three states of engagement ($e = 1$ disengaged, $e = 3$ fairly engaged, $e = 5$ fully engaged). It can be observed from this table that for the duration of blinks t_b, the values for $e = 5$ are lower than those relating $e = 3$, which are lower than the ones of $e = 1$, all in line with the expectations. Moreover, f_b is lower for $e = 5$ (fully engaged) as compared to $e = 3$ and $e = 1$, which means that they blink more often for these values. Nevertheless, there is not a monotonic decrease. For the aspect ratio of the eyes $\bar{r}_o$, eyes are more widely open for higher levels of engagement and the values of $\bar{r}_o$ are smaller for users with lowers levels of engagement. In addition, interocular breadth d_{io} gets lower as e decreases, which means that participants prefer staying further away from the monitor as they lose their engagement. Although the general tendency is mostly monotonic for the four variables, by taking a closer look at the standard deviations, we can see that the distinction between variable values is lower for similar states of engagement (i.e., closer values of e).

In addition, we compute the polyserial correlation coefficient ρ. In particular, polyserial correlation defines the correlation between a quantitative variable and an ordinal variable. It is based on the assumption that the joint distribution of the quantitative variable and a latent continuous variable underlying the ordinal variable is bivariate normal.

In our case, we consider the values of the proposed features described in Section 4.2 relating each clip as the numerical variables. The corresponding ordinal variables are the labels assigned by the expert coder. For estimating their correlation, we opt for a maximum likelihood approach, which maximizes the bivariate-normal likelihood with respect to the ordinal variable.[15]

Computing polyserial correlation coefficient ρ for duration and frequency of blinks, we demonstrate that they have a mild- negative correlation with engagement, namely $\rho(t_b) = -0.29$ and $\rho(f_b) = -0.31$. Specifically, these values ascertain that when the level of engagement decreases, the average duration of blinks gets longer and frequency of blinks increases. This finding is in line with the suggestion of (Smilek et al., 2010), which state that blinking helps humans disengage from the outside stimuli, in favor of the other cognitive processing (i.e., mind wandering); as well as the fact that increased workload causes a higher blinking rate in the long term (Matthews & Desmond, 1998).

On the other hand, the interocular breadth d_{io} and eye aspect ratio $\bar{r}_o$ have a somewhat stronger positive correlation with the apparent level of engagement, namely $\rho(d_{io}) = 0.58$ and $\rho(\bar{r}_o) = 0.66$. In other words, as the level engagement increases, interocular breadth and normalized eye size increase as well. This indicates that when the user is concentrated on the task, his/her face is closer to the screen and; he/she looks at the screen with eyes wider open.

Table 1. Descriptive statistics for the values of the variables regarding $e = 1$, $e = 3$ and $e = 5$. The values in the table are organized as $\mu \pm \sigma(\varepsilon)$, where μ is the mean, σ is the standard deviation and ε is the standard error.

	$e = 1$	$e = 3$	$e = 5$
t_b	0.26 ± 0.06 (0.04)	0.16 ± 0.11 (0.02)	0.07 ± 0.06 (0.02)
f_b	0.12 ± 0.03 (0.02)	0.16 ± 0.08 (0.02)	0.07 ± 0.07 (0.02)
$\bar{r}_o$	0.22 ± 0.01 (0.01)	0.28 ± 0.02 (0.00)	0.29 ± 0.02 (0.01)
d_{io}	0.21 ± 0.00 (0.00)	0.25 ± 0.03 (0.00)	0.29 ± 0.03 (0.01)

4.4. Deriving probability distributions

Subsequent to verifying that the proposed features present credible correlation with the assigned levels of engagement, we propose a method to probabilistically assess the level of engagement. To that end, we derive probability density function (pdf) of the features described in Section 4.2 from their respective empirical observations. In doing that, we utilize kernel density estimation (KDE), which is one of the most popular non-parametric methods in estimating the underlying pdf of a set of observations (Alpaydin, 2016).

Let x be a random variable and $(x_1, x_2, \ldots, x_n)$ be a set of n samples drawn from a distribution with an unknown density f. KDE of f can be expressed as,

$$\hat{f}(x|h) = \frac{1}{nh}\sum_{i=0}^{n} F\left(\frac{x - x_i}{h}\right), \tag{5}$$

where F is the kernel (a non-negative function) and $h > 0$ is the smoothing hyper-parameter (i.e., bandwidth). Regarding the kernel F, Gaussian distribution is considered to give satisfactory results in most cases and we adopt this convention. On the other hand, the selection of the smoothing hyper-parameter h in Equation (5) emerge as one sensitive point of KDE, which often bears a bias-variance trade-off (Heidenreich et al., 2013). Since the bandwidth estimate selected by the least squares cross-validation is known to be subject to large sample variation, we use grid search over a given interval at evenly spaced points (VanderPlas, 2016).

In addition to bandwidth selection, KDE needs to be handled carefully also against curse of dimensionality. Namely, the principles explained using Equation (5) based on a single variable can in theory be extended easily to a multivariate case. However, in practice, multivariate kernel

density estimation is usually restricted to 2-D due to the curse of dimensionality. Similar to most other applications, also in our case, operating in the full (4-D) variable space Σ potentially yields an overwhelmingly large number of bins, and thus the space is sparsely populated by data points. Therefore, we prefer using a set of 1-D variable spaces. However, this choice needs a justification for conditional independence of observations, which is elaborated on in Section 4.5.

4.5. Verification of conditional independence of feature distribution

To verify that the features are conditionally independent, we adopt an information theoretic approach and use relative entropy distance. Principally, entropy distance of two random variables as Θ and Δ is defined as

$$D_H(\Theta, \Delta) = H(\Theta, \Delta) - I(\Theta; \Delta), \tag{6}$$

where $H(\Theta, \Delta)$ and $I(\Theta, \Delta)$ are, respectively, the joint entropy and mutual information of these variables (MacKay, 2017).

In explicit terms, joint entropy is

$$H(\Theta, \Delta) = -\sum_{i,j} p(\theta_i, \delta_j) \log_2(p(\theta_i, \delta_j)), \tag{7}$$

whereas the mutual information is defined as

$$I(\Theta; \Delta) = \sum_{i,j} p(\theta_i, \delta_j) \log_2\left(\frac{p(\theta_i, \delta_j)}{p(\theta_i)p(\delta_j)}\right). \tag{8}$$

On the other hand, the relative entropy distance, which is defined as

$$D(\Theta, \Delta) = \frac{D_H(\Theta, \Delta)}{H(\Theta, \Delta)}, \tag{9}$$

is useful for our purposes since it is a *true* metric. Namely, as it is elaborately explained in (Li et al., 2003), it is non-negative, symmetric and it satisfies the triangle inequality. In addition, it is bounded to the interval $[0, 1]$, which makes it easy to interpret. Specifically, for uncorrelated variables, the relative entropy distance $D(\Theta, \Delta)$ should be 1, and closer to 0 for correlated ones.

Table 2 shows relative entropy distance values between each pair of variables.[16] From the values presented in Table 2, we can infer that there is a reasonable degree of independence between all pairs.[17] As a result, we can claim that the curse of dimensionality can be overcome using a decomposition of individual density distributions.

4.6. Resulting 1-D feature distributions

The outcomes of the estimation process described in Section 4.4 applied to each individual feature are presented in Figure 4. For the sake of simplicity, Figure 4 provides a comparison between the two extremities of engagement (i.e., fully engaged and disengaged) regarding the features defined in Section 4.2.

From Figure 4a, it is clear that as e decreases, there is a trend of observing longer blinks. On the other hand, the frequency of blinks has the tendency to be higher for disengaged participants as presented in Figure 4b. These findings support the conclusions of (Schooler et al., 2011) relating the effect of blinks on perceptual decoupling. Regarding $\bar{r}_o$ presented in Figure 4c, as e decreases, eye aspect ratio decreases. This observation is also in line with the findings of Section 4.3 and (Koyama et al., 2019). From Figure 4d, it can be seen that the interocular breadth is smaller for $e = 1$, indicating that as the level of engagement decreases participants prefer staying further away from the display (Asteriadis et al., 2011).

Table 2. Relative entropy distances for all pairs of variables.

	t_b	f_b	$\bar{r}_o$	d_{io}
t_b	0.00	0.95	0.93	0.92
f_b		0.00	0.91	0.91
$\bar{r}_o$			0.00	0.90
d_{io}				0.00

4.7. Probabilistic determination of engagement level

Potentially, the procedure described in Section 4.4 can be applied on all values of $e \in [1, 5]$ and five pdfs can be derived for each feature. In that regard, given the observed value of an arbitrary feature σ and its concerning pdfs, it is trivial to estimate the level of engagement, where the most intuitive approach would be to evaluate this value in pdfs relating all $e \in [1, 5]$ and consider the engagement level with the highest likelihood as the final (discrete) estimation result. Note that it is still necessary to blend together the discrete estimations (relating each individual feature) into one decisive estimation.

However, a qualitative comparison of the KDEs reveals that neighboring engagement levels, namely $e = j$ and $e = j + 1$ present a quite similar behavior, which is not completely surprising, given the subjective nature of these labels. Nevertheless, the difference between the two extremities, $e = 1$ and $e = 5$, is obviously larger than the parameter variation. For this reason, we consider the distributions representing $e = 1$ (disengaged) and $e = 5$ (fully engaged) as benchmarks in estimation of engagement level. In relation to that, we also opt for a probabilistic estimation method rather than a discrete one.

Without loss of generality, consider initially the case concerning an arbitrary σ. By evaluating the observed value of σ in its pdf relating $e = 1$, we compute the likelihood that this observation belongs to a disengaged person, $L_d(\sigma)$,

$$L_d(\sigma) = p(\sigma | e = 1). \tag{10}$$

Similarly, by evaluating it in its pdf concerning $e = 5$, the likelihood that σ comes from a distribution relating fully engaged users is computed as, $L_e(\sigma)$,

$$L_e(\sigma) = p(\sigma | e = 5). \tag{11}$$

Then, we can derive in an empirical way, the probability of being engaged p_e and the probability of being disengaged p_d, where

$$p_e = \frac{L_e(\sigma)}{L_e(\sigma) + L_d(\sigma)}, \tag{12}$$

and p_d is simply the complementary probability, $p_d = 1 - p_e$.

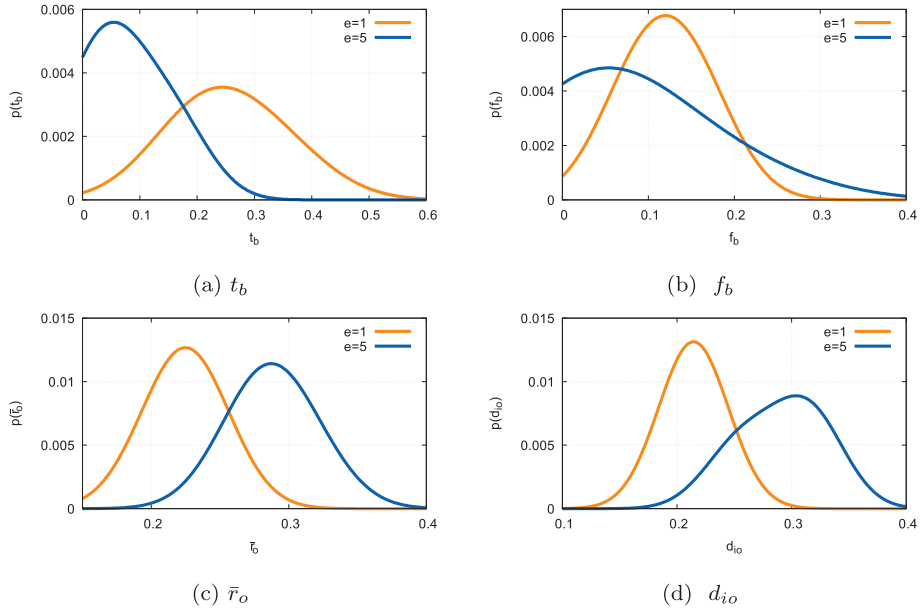

Figure 4. Probability distributions for (a) t_b, (b) f_b, (c) $\bar{r}_o$, and (d) d_{io} for two extremities of engagement level ($e = 1$ disengaged, and $e = 5$ fully engaged).

Clearly, it is possible to apply the above procedure on any $\sigma \in \Sigma$ as well as on the entire set of features Σ or a subset of several features $\Sigma' \subseteq \Sigma$. In other words, by exploiting the property of conditional independence of features illustrated in Section 4.5, we can compute the likelihood that the user is engaged as,

$$L_e(\Sigma') = \prod_{\sigma \in \Sigma'} L_e(\sigma). \tag{13}$$

Upon determining $L_d(\Sigma')$ in a similar fashion, we can use the same idea as in Equation (12) (i.e., complementarity of being engaged and disengaged) and compute the probability of engagement p_e based on Σ'.

5. Results and discussion

In investigating the efficacy of the proposed features and estimation method, we adopt the following assessment approach. First of all, for evaluating the effectiveness of each particular feature $\sigma \in \Sigma$, we apply the probabilistic approach presented in Equation (12) to individual features. Next, by applying the estimation method on the set of all features Σ as in Equation (13), we determine the potentially optimum performance.

Figure 5 demonstrates the probability of being engaged based on each of the four features described in Section 4.2. From this figure, it is clear that p_e is monotonically increasing for growing values of e for most σ, although the rate of increase varies between the features. Nevertheless, the overall

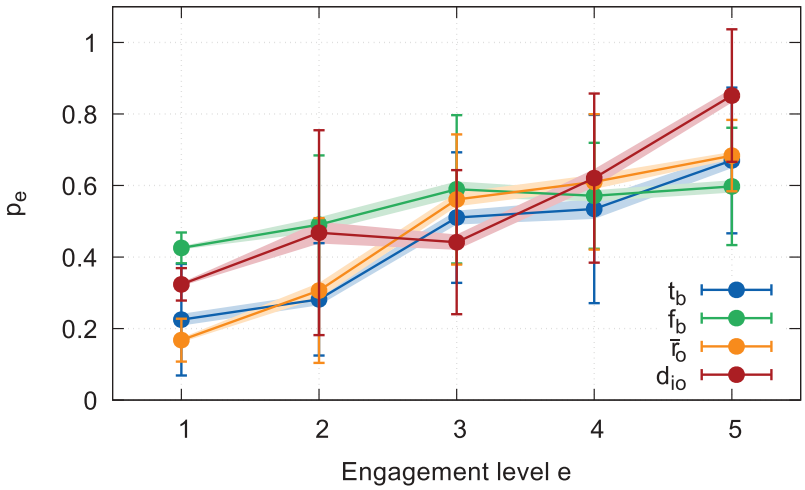

Figure 5. The probability of being engaged p_e computed using each individual feature. Error bars represent standard deviation and shaded region represents standard error.

tendency still presents supporting evidence for the efficacy of the proposed features, particularly for d_{io} and $\bar{r}_o$, which are in line with the conclusions based on polyserial correlation values given in Section 4.3.

Moreover, it is not surprising that standard deviation values are smaller for extremities, since the extremities are adopted as benchmarks. On the other hand, standard deviations are higher for the intermediate values of e. However, from qualitative observations, we know that these values of e are the ones, which the coders agree less, thus it is expected that there is a larger deviation on those. In addition, due to the large number of samples, the standard error is much smaller than standard deviation for all values of e.

As presented in Figure 6, we obtain a clear improvement in estimation of engagement by integrating the information from all the features, i.e., employing Σ. As expected, the values of p_e obtained by the integration of all σ is monotonically increasing. Also, it yields a considerable separation between values of p_e relating $e \in [1, 2]$ (i.e., disengaged or poorly engaged) and $e \in [4, 5]$ (i.e., moderately engaged to fully engaged). In particular, we see that when the user is not engaged p_e is almost 0, whereas it increases steeply as he/she reaches fair or higher levels of engagement. These findings suggest that by estimating p_e with the proposed method and setting a threshold at some value around 0.50, we can detect the engagement levels below average and above average with a satisfactory accuracy.

In addition to this integration approach (i.e., employing the set of all features Σ), we can also apply a "differential approach," where we remove one feature at a time from the input set, $\Sigma' = \Sigma - \sigma$, as another means to evaluate the sensitivity of the method to each individual feature. The curves in Figure 7 present the outcomes of the differential approach. In agreement with the previous inferences, d_{io} is found to make the largest contribution to performance, since its removal causes a larger degradation. Yet this degradation is no more than 0.1 on the average. It is interesting to note that removing d_{io} we achieve a slight degradation at the extremities but also a smaller deviation at intermediate levels, which indicates

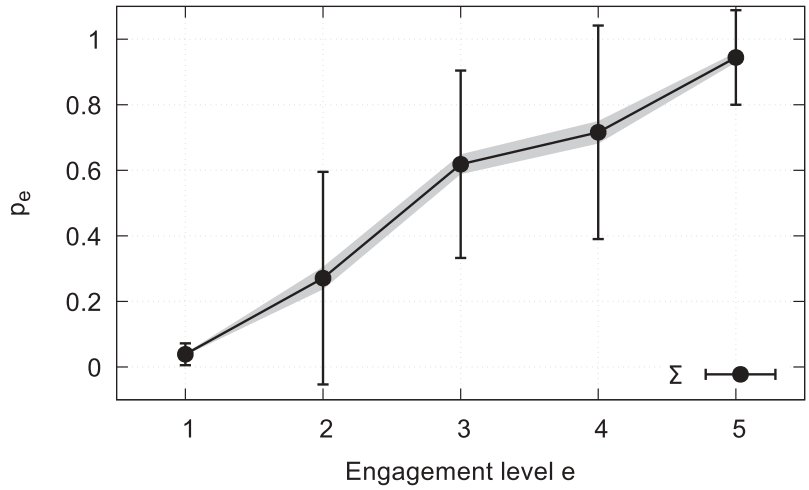

Figure 6. The probability of being engaged p_e computed integrating all features. Error bars represent standard deviation and shaded region represents standard error.

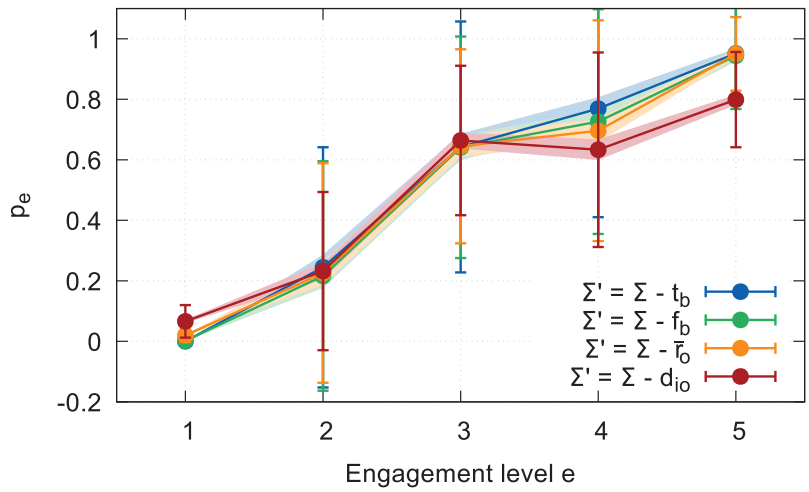

Figure 7. The probability of being engaged p_e computed by removing one feature at a time from the feature set, $\Sigma' = \Sigma - \sigma$. Error bars represent standard deviation and shaded region represents standard error.

a trade-off. Nevertheless, the small standard error values show that provided that a larger sample size is achieved, the effect of this deviation will diminish.

In addition to the above-mentioned differential approach on input variables, we also examined the dependence of p_e on the type of users' task. We observed that the active task receives slightly higher values regarding $e = 3$, and quite similar values for the remaining levels of engagement. Nevertheless, the behavior at $e = 3$ is not significant once it is compared to the variation of p_e values for other values of e, in particular for $e \in \{2, 4\}$.

In addition, we would like to mention several limitations of our approach. The first issue relates the generalization of implications due to certain actions. Namely, some actions such as perceptual decoupling may indicate disengagement in certain settings (e.g., closing the eyes to take a break from studying), while they may indicate to engagement in others (e.g., closing the eyes to stop distracting stimuli). Although the learning tasks used in this study are designed in a way that such contradictory implications will not be observed, a generalized version of this study addressing a larger variety of learning tasks needs to account for such issues, possibly by considering sets or sequences of actions (of the eyes or the mouth, etc.) rather than isolated actions (e.g., of only the eyes).

Another limitation is that only the semi-active task has ecological validity in actual learning settings. In addition, the passive task can be considered to be somewhat similar to passive online learning, since it is linear and straightforward. Nevertheless, tasks such as the active one do not have direct correspondence in e-learning.

Note that several kinds of performance data are collected involving participants' subjective evaluations as well as objective measures such as user activity logs in the active task, or answers to the questions in the semi-active task. However, we could not find a smooth way to incorporate them with the behavioral variables derived from the video-clips. The main reason for this is that the video-clips span a too short time window to assess performance in a reliable way. In other words, either the performance values inherently concern the entire duration/segment of the task (longer than the annotated 10 s window) or the number of data points (e.g., derived from the computer registered values) are too small to establish a direct relation between the video-clip of interests and task performance.

6. Conclusion

Although conventional approaches in joint attention rely on the assumption that gaze indicates sustained visual attention (Yücel et al., 2013), elaborate studies on brain activity show that humans may seem to attend but lose engagement over time (Eldenfria & Al-Samarraie, 2019; Szafir & Mutlu, 2012). In that sense, this study focuses particularly on technology-mediated learning systems and offers a method to estimate users' level of engagement in a probabilistic manner.

In order to detect declines in levels of engagement, we choose a set of features relying on the findings of several works from the research fields of cognitive science, affective computing, eye physiology. In particular, we exploit the following findings: (i) mental or cognitive fatigue is in close connection with perceptual decoupling, which is enhanced by blinking (Smilek et al., 2010); (ii) involvement in computer-based tasks correlates significantly with depth of the users (Asteriadis et al., 2011); (iii) visual fatigue due to exposure to digital displays is correlated with blinking rate and this correlation is expected to grow if the user is under high task load (Matthews & Desmond, 1998; Rosenfield, 2011), (iv) aspect ratio of the eyes present significant information on attentional state of users (Ji et al., 2004). Based on these results, we define a set of four features, all derived from ocular landmarks and search for subtle cues of dis/engagement.

Specifically, we compute facial landmarks from videos of human subjects carrying out various tasks and derive several features from the landmarks describing the eyes. Namely, we derive the frequency and duration of spontaneous blinks; interocular breadth and eye aspect ratio. These features are confirmed to involve valuable information about labels of engagement assigned by expert coders (Koyama et al., 2019). Building concerning pdfs from empirical observations through kernel density estimation, we propose a probabilistic method for engagement level. We show the features and the method achieve a considerable performance in estimation of engagement. In addition, we evaluate the performance using each individual feature and the set of all features, as well as subset of features (removing a single feature at a time).

The proposed approach has several advantages as (i) capability of on-the-fly assessment of engagement, (ii) potential stimulation of the user with motivational advice, etc., immediately upon detection of disengagement, (iii) enabling a customization to person-specific factors by a simple calibration of the fundamental models and adjustment to interpersonal variations in behavior, and (iv) a possibility of detection of dis/engagement (or its embodiment) by interaction agents.

Notes

1. D'Mello et al. call attention to biological or physical indicators such as skin conductance or mouse pressure, etc., as well. But these require a specific sensory configuration and can not be easily incorporated with existing systems.
2. The participants speak English as a foreign language, have followed a similar academic curricula and, thus, are assumed to have a similar level of proficiency in English. In addition, NASA TLX surveys carried out following the semi-active task, as well as investigation of correctness of participants' answers to the questions on the narrations, reveal that they do not experience any problems due to any insufficiency in English proficiency.
3. The participants listen to 35 stories narrated on the average for 251 ± 47 sec. After each narration, they are given 15 sec to answer a question on the story.
4. There are one male and four female participants with age 26.8 ± 2.7.
5. The participants performed only a single task on each day and finished all tasks within a time window of 2 weeks.
6. The video footage has a resolution of 1280×720 and a frame rate of 30 fps, which are in line with the specifications of most off-the-shelf recording products or built-in computer hardware.

7. The start instants of the video clips are determined (from the beginning of the task) in minutes as follows: $[5, 10, 15, 25, 30, 40, 50, 60, 70, 80, 85, 90, 100, 105, 110]$.
8. In particular, we consider $e = 5$ as "fully engaged," $e = 4$ as "moderately engaged," $e = 3$ as "fairly engaged," $e = 2$ as "poorly engaged," and $e = 1$ as "disengaged." However, while contrasting the extremities, we use the terms "engaged" and "disengaged" for the sake of brevity.
9. The last 35 clips coded by the teachers are found to have unreliable labels most probably due to a confusion of one of the coders; and one clip is found to involve no learning task and, thus, is discarded.
10. Taking a closer look at the most popular landmark estimation methods, one may notice that it is quite common to use templates involving around 60 points.
11. In other words, we exclude any reflex or voluntary blinks. The reason for this exclusion is two-fold. First of all, since tactile stimuli (to the face or other body parts) are not present in our experiments, and the degree of optical or auditory stimuli is not to a significant degree or subject to large variations, we assume no reflex blinks take place. In addition, since the participants are not aware that we study their blinking patterns and are neither instructed to blink intentionally, they are assumed not to perform any voluntary blinks.
12. In our specific set, since frame rate is 30 fps and clip duration is 10 sec, $N_f = 300$.
13. Obviously, one may as well opt for replacing the blink onset with blink offset. Since the duration of the clips (i.e. 10 sec) is considered to assure a uniform level of engagement over this course, even though the value associated with a particular clip may change (i.e. increase or decrease by 1), the distribution of number of blinks relating a particular level of engagement is expected not to be affected by this choice. In addition, the integration of f_b with the other features is regarded to improve resiliency and stability of estimation.
14. The biocular breadth, i.e. the distance between the two landmarks representing the lateral canthi, can also be used to replace d_{io}.
15. For optimization, a general-purpose method based on Nelder-Mead algorithm is used (Nelder & Mead, 1965). In implementation, we used rpy2 package, which is a back-end for R programming language to Python (Gautier, 2008).
16. Since the matrices in Table 2 are symmetric, only the upper triangular part is presented.
17. There is no guideline to assess independence based on relative entropy distance but various studies consider values over 0.90 to indicate independence to a sufficient degree (Zanlungo et al., 2017).

Acknowledgments

We would like to thank our volunteer participants for their help in the experiments. We would also like to thank Dr. Francesco Zanlungo for his invaluable discussion.

Disclosure of potential conflict of interest

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Japan Society for the Promotion of Science KAKENHI Grant Number J18K18168.

ORCID

Zeynep Yücel http://orcid.org/0000-0003-3404-4485
Akito Monden http://orcid.org/0000-0003-4295-207X

References

Abramowitz, M., Stegun, I. A., & Romer, R. H. (1988). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. AAPT.

Alpaydin, E. (2016). *Machine learning: The new AI*. MIT press.

Arkorful, V., & Abaidoo, N. (2015). The role of e-learning, advantages and disadvantages of its adoption in higher education. *International Journal of Instructional Technology and Distance Learning*, *12*(1), 29–42. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.694.3077&rep=rep1&type=pdf#page=33

Aslan, S., Alyuz, N., Okur, E., Mete, S. E., Oktay, E., & Esme, A. A. (2018). Effect of emotion-aware interventions on students' behavioral and emotional states. *Educational Technology Research and Development*, *66*(6), 1399–1413. https://doi.org/10.1007/s11423-018-9589-7

Asteriadis, S., Karpouzis, K., & Kollias, S. (2011). The importance of eye gaze and head pose to estimating levels of attention. In *Proceedings of international conference on games and virtual worlds for serious applications* (pp. 186–191). IEEE. http://dx.doi.org/10.1109/VS-GAMES.2011.38

Balaban, C. D., Cohn, J., Redfern, M. S., Prinkey, J., Stripling, R., & Hoffer, M. (2004). Postural control as a probe for cognitive state: Exploiting human information processing to enhance performance. *International Journal of Human-Computer Interaction*, *17*(2), 275–286. https://doi.org/10.1207/s15327590ijhc1702_9

Ballenghein, U., & Baccino, T. (2019). Referential processing during reading: Concurrent recordings of eye movements and head motion. *Cognitive Processing*, *20*(3), 371–384. https://doi.org/10.1007/s10339-018-0894-1

Beinicke, A., & Bipp, T. (2018). Evaluating training outcomes in corporate e-learning and classroom training. In *Vocations and learning* (pp. 1–28). Springer. http://dx.doi.org/10.1007/s12186-018-9201-7

Ben-Zadok, G., Leiba, M., & Nachmias, R. (2011). Drills, games or tests? Evaluating students' motivation in different online learning activities using log file analysis. *Journal of E-Learning and Learning Objects*, *7*(1), 235–248. http://dx.doi.org/10.28945/1522

Bonnet, C. T., & Baudry, S. (2016). A functional synergistic model to explain postural control during precise visual tasks. *Gait & Posture*, 50, 120–125. https://doi.org/10.1016/j.gaitpost.2016.08.030

Bonnet, C. T., Szaffarczyk, S., & Baudry, S. (2017). Functional synergy between postural and visual behaviors when performing a difficult precise visual task in upright stance. *Cognitive Science*, *41*(6), 1675–1693. https://doi.org/10.1111/cogs.12420

Borji, A., & Itti, L. (2015). *Cat2000: A large scale fixation dataset for boosting saliency research*. arXiv:1505.03581

Bosch, N., & D'Mello, S. (2019). Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing*, 1. https://doi.org/10.1109/TAFFC.2019.2908837

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cruz, A. A., Garcia, D. M., Pinto, C. T., & Cechetti, S. P. (2011). Spontaneous eyeblink activity. *The Ocular Surface*, *9*(1), 29–41. https://doi.org/10.1016/S1542-0124(11)70007-6

D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist*, *52*(2), 104–123. https://doi.org/10.1080/00461520.2017.1281747

Dixson, M. D. (2015). Measuring student engagement in the online course: The online student engagement scale (ose). *Online Learning*, *19*(4), n4. https://doi.org/10.24059/olj.v19i4.561

Drutarovsky, T., & Fogelton, A. (2014). Eye blink detection using variance of motion vectors. In *Proceedings of European conference on computer vision* (pp. 436–448). http://dx.doi.org/10.1007/978-3-319-16199-0_31

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern Classification*. John Wiley & Sons.

Eldenfria, A., & Al-Samarraie, H. (2019). Towards an online continuous adaptation mechanism (OCAM) for enhanced engagement: An EEG

study. *International Journal of Human–Computer Interaction*, *35*(20), 1–15. http://dx.doi.org/10.1080/10447318.2019.1595303

Eliot, J. A., & Hirumi, A. (2019). Emotion theory in education research practice: An interdisciplinary critical literature review. In *Educational technology research and development* (pp. 1–20). Springer. http://dx.doi.org/10.1007/s11423-018-09642-3

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378. https://doi.org/10.1037/h0031619

Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. In J. Vermunt (Ed.), *Learning and Instruction* (pp. 1–5). Elsevier.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Gautier, L. (2008). *rpy2: A simple and efficient access to R from Python*. Retrieved January 26, from http://rpy.sourceforge.net/rpy2.html

Graesser, A. C., & D'Mello, S. (2012). Moment-to-moment emotions during reading. *The Reading Teacher*, *66*(3), 238–242. https://doi.org/10.1002/TRTR.01121

Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist*, *50*(1), 14–30. https://doi.org/10.1080/00461520.2014.989230

Haßler, B., Major, L., & Hennessy, S. (2016). Tablet use in schools: A critical review of the evidence for learning outcomes. *Journal of Computer Assisted Learning*, *32*(2), 139–156. https://doi.org/10.1111/jcal.12123

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *WCST: Wisconsin card sorting test*. Psychological Assessment Resources.

Heidenreich, N.-B., Schindler, A., & Sperlich, S. (2013). Bandwidth selection for kernel density estimation: A review of fully automatic selectors. *AStA Advances in Statistical Analysis*, *97*(4), 403–433. https://doi.org/10.1007/s10182-013-0216-y

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, *90*, 36–53. https://doi.org/10.1016/j.compedu.2015.09.005

Hunter, M. C., & Hoffman, M. A. (2001). Postural control: Visual and cognitive manipulations. *Gait & Posture*, *13*(1), 41–48. https://doi.org/10.1016/S0966-6362(00)00089-8

Ji, Q., Zhu, Z., & Lan, P. (2004). Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE Transactions on Vehicular Technology*, *53*(4), 1052–1068. https://doi.org/10.1109/TVT.2004.830974

Kaakinen, J. K., Ballenghein, U., Tissier, G., & Baccino, T. (2018). Fluctuation in cognitive engagement during reading: Evidence from concurrent recordings of postural and eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(10), 1671. http://dx.doi.org/10.1037/xlm0000539

Kasinski, A., Florek, A., & Schmidt, A. (2008). The PUT face database. *Image Processing and Communications*, *13*(3–4), 59–64. https://www.researchgate.net/profile/Adam_Schmidt/publication/232085001_The_PUT_face_database/links/09e4150d0bf1e5080f000000/The-PUT-face-database.pdf

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 1867–1874). IEEE. http://dx.doi.org/10.1109/CVPR.2014.241

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*(6006), 932. https://doi.org/10.1126/science.1192439

King, D. (2018). *Dlib C++ Library*. Retrieved March 11, 2019, from http://dlib.net/

Kong, S. C. (2011). An evaluation study of the use of a cognitive tool in a one-to-one classroom for promoting classroom-based dialogic interaction. *Computers & Education*, *57*(3), 1851–1864. https://doi.org/10.1016/j.compedu.2011.04.008

Koyama, S., Yücel, Z., & Monden, A. (2019). Quantitative evaluation of the relation between blink features and apparent task engagement. In *Proceedings of European conference on visual perception* (pp. 732). SAGE.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communications Research*, *30*(3), 411–433. http://dx.doi.org/10.1093/hcr/30.3.411

Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. (2003). The similarity metric. In *Proceedings of the ACM-SIAM symposium on discrete algorithms* (pp. 863–872). http://dx.doi.org/10.1109/TIT.2004.838101

Liu, M., McKelroy, E., Corliss, S. B., & Carrigan, J. (2017). Investigating the effect of an adaptive learning intervention on students' learning. *Educational Technology Research and Development*, *65*(6), 1605–1625. https://doi.org/10.1007/s11423-017-9542-1

MacKay, D. J. (2017). *Information theory, inference and learning algorithms*. Cambridge University Press.

Matthews, G., & Desmond, P. A. (1998). Personality and multiple dimensions of task-induced fatigue: A study of simulated driving. *Personality and Individual Differences*, *25*(3), 443–458. https://doi.org/10.1016/S0191-8869(98)00045-2

Mayer, R. E. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning*, *33*(5), 403–423. https://doi.org/10.1111/jcal.12197

Mbouna, R. O., Kong, S. G., & Chun, M.-G. (2013). Visual analysis of eye state and head pose for driver alertness monitoring. *IEEE Transactions on Intelligent Transportation Systems*, *14*(3), 1462–1469. https://doi.org/10.1109/TITS.2013.2262098

Miller, B. W. (2015). Using reading times and eye-movements to measure cognitive engagement. *Educational Psychologist*, *50*(1), 31–42. https://doi.org/10.1080/00461520.2015.1004068

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

Neog, D. R. (2018). *Measurement and animation of the eye region of the human face in reduced coordinates* [Unpublished doctoral dissertation]. University of British Columbia.

O'Donnell, E., Lawless, S., Sharp, M., & Wade, V. P. (2015). A review of personalised elearning: Towards supporting learner diversity. *International Journal of Distance Education Technologies*, *13*(1), 22–47. https://doi.org/10.4018/ijdet.2015010102

Pan, G., Sun, L., Wu, Z., & Lao, S. (2007). Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *Proceedings of IEEE international conference on computer vision* (pp. 1–8). IEEE. http://dx.doi.org/10.1109/ICCV.2007.4409068

Piskurich, G. M. (2006). Online learning: E-learning. Fast, cheap, and good. *Performance Improvement*, *45*(1), 18–24. https://doi.org/10.1002/pfi.2006.4930450105

Rosenfield, M. (2011). Computer vision syndrome: A review of ocular causes and potential treatments. *Ophthalmic and Physiological Optics*, *31*(5), 502–515. https://doi.org/10.1111/j.1475-1313.2011.00834.x

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, *47*, 3–18. https://doi.org/10.1016/j.imavis.2016.01.002

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences*, *15*(7), 319–326. https://doi.org/10.1016/j.tics.2011.05.006

Seta, L., Kukulska-Hulme, A., & Arrigo, M. (2014). What have we learnt about mobile LifeLong Learning (mLLL)? *International Journal of Lifelong Education*, *33*(2), 161–182. https://doi.org/10.1080/02601370.2013.831954

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist*, *50*(1), 1–13. https://doi.org/10.1080/00461520.2014.1002924

Smilek, D., Carriere, J. S., & Cheyne, J. A. (2010). Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering. *Psychological Science*, *21*(6), 786–789. https://doi.org/10.1177/0956797610368063

Soukupová, T., & Cech, J. (2016). Real-time eye blink detection using facial landmarks. In *Proceedings of computer vision winter workshop* (pp. 1–8). http://cvww2016.vicos.si/index.html%3Fp=13.html

Szafir, D., & Mutlu, B. (2012). Pay attention!: Designing adaptive agents that monitor and improve user engagement. In *Proceedings of SIGCHI conference on human factors in computing systems* (pp. 11–20). ACM. http://dx.doi.org/10.1145/2207676.2207679

Thomas, C., & Jayagopi, D. B. (2017). Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of ACM SIGCHI international workshop on multimodal interaction for education* (pp. 33–40). ACM. http://dx.doi.org/10.1145/3139513.3139514

Urhahne, D., & Zhu, M. (2015). Accuracy of teachers' judgments of students' subjective well-being. *Learning and Individual Differences*, *43*, 226–232. https://doi.org/10.1016/j.lindif.2015.08.007

Uřičář, M., Franc, V., Thomas, D., Sugimoto, A., & Hlaváč, V. (2016). Multi-view facial landmark detector learned by the structured output SVM. *Image and Vision Computing*, *47*, 45–59. https://doi.org/10.1016/j.imavis.2016.02.004

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. O'Reilly Media Inc.

Wang, N., Gao, X., Tao, D., Yang, H., & Li, X. (2018). Facial feature point detection: A comprehensive survey. *Neurocomputing*, *275*, 50–65. https://doi.org/10.1016/j.neucom.2017.05.013

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86–98. https://doi.org/10.1109/TAFFC.2014.2316163

Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., & Zhou, Q. (2018). Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of IEEE International conference on computer vision and pattern recognition* (pp. 2129–2138). IEEE. http://dx.doi.org/10.1109/CVPR.2018.00227

Wu, X., Anderson, R. C., Nguyen-Jahiel, K., & Miller, B. (2013). Enhancing motivation and engagement through collaborative discussion. *Journal of Educational Psychology*, *105*(3), 622. https://doi.org/10.1037/a0032792

Wu, Y., & Ji, Q. (2019). Facial landmark detection: A literature survey. *International Journal of Computer Vision*, *127*(2), 115–142. Springer. http://dx.doi.org/10.1007/s11263-018-1097-z

Yücel, Z. (2020). *Implementation of Wisconsin card sorting task in Java*. Retrieved February, 2020, fromhttps://github.com/yucelzeynep/WCST

Yücel, Z., Salah, A. A., Meric Li, C., Meric Li, T., Valenti, R., & Gevers, T. (2013). Joint attention by gaze interpolation and saliency. *IEEE Transactions on Cybernetics*, *43*(3), 829–842. https://doi.org/10.1109/TSMCB.2012.2216979

Zanlungo, F., Yücel, Z., Brščić, D., Kanda, T., & Hagita, N. (2017). Intrinsic group behaviour: Dependence of pedestrian dyad dynamics on principal social and personal features. *PLoS One*, *12*(11), e0187253. https://doi.org/10.1371/journal.pone.0187253

Zhu, M., & Urhahne, D. (2014). Assessing teachers' judgements of students' academic motivation and emotions across two rating methods. *Educational Research and Evaluation*, *20*(5), 411–427. https://doi.org/10.1080/13803611.2014.964261

About the Authors

Zeynep Yücel is an assistant professor at Okayama University, Japan. She obtained her B.S. degree from Bogazici University, Istanbul, Turkey, and her M.S. and Ph.D. degrees from Bilkent University, Ankara, Turkey in 2005 and 2010, all in electrical engineering. She was a postdoctoral researcher at ATR labs in Kyoto, Japan for 5 years, before being awarded a JSPS fellowship in 2016. Her research interests include robotics, signal processing, computer vision, and pattern recognition.

Serina Koyama received a B.E. degree in Information Technology from Okayama University in 2016. Her research interests include pattern recognition with applications in affective computing.

Akito Monden is a professor in the Graduate School of Natural Science and Technology at Okayama University, Japan. He received the B.E. degree (1994) in electrical engineering from Nagoya University, and the M.E. and D.E. degrees in information science from Nara Institute of Science and Technology (NAIST) in 1996 and 1998, respectively. His research interests include software measurement and analytics, and software security and protection. He is a member of the IEEE, ACM, IEICE, IPSJ and JSSST.

Mariko Sasakura is an assistant professor at Okayama University, Japan. Her research interests include visualization, especially visualizing program structures, animation systems, human computer interaction on mobile devices, origami simulator and migration simulator systems.