特別研究報告書

題目

Estimation of Students' Level of Engagement from Facial Landmarks

指導教員

報告者

小山 聖莉奈

岡山大学工学部 情報系学科

平成31年2月8日 提出

Abstract

This study focuses on e-learning systems, where a user utilizes a computer-based platform to learn certain subjects. Such platforms often have the capability of observing users' behavioral reactions, i.e. actions and gestures. This kind of reactions are known to involve important clues about users' engagement and teachers use such reactions in adjusting their teaching pace and method in conventional classroom scenarios. In this respect, this study makes an effort to introduce similar adaptation capabilities to e-learning systems by providing an estimation of level of engagement through such reactions.

To develop and test our estimation method, we record a video data set from various e-learning system users, carrying out various tasks (passive, semi-active or active) for a prolonged duration. We crop various video clips, each of which is 10 sec long, and present them to 2 coders, who hold professional teacher license and actively practice teaching languages. These coders rate each video clip on a scale from 1 to 5 according to the level of engagement of the user, which constitutes the ground truth of our experiments.

We analyze these video clips by resolving facial landmarks and study particularly the landmarks relating the eyes. We derive several features relating blink patterns such as duration and number of blinks as well as the eye aspect ratio and normalized eye size, which are all proven statistically to be correlated with the engagement labels.

We then compute empirical probability distributions relating each pair of features and engagement labels based on a kernel density estimate. These distributions are used to evaluate the probability of being engaged and it is shown that a satisfactory accuracy of estimation is achieved when the information from all proposed features are integrated. Specifically, the probability of being engaged ranges between 1 and 0.90, when the user is labeled to be fully engaged or medium engaged. On the other hand, for users labels with mild or complete lack of engagement, this probability drops drastically to below 0.6.

Contents

List of Figures

List of Tables

Section 1

Introduction

1.1 Motivation

E-learning is utilization of electronic technologies to access educational curriculum outside of a classroom. To that end, the courses and programs are delivered completely in computerized medium. Nowadays, e-learning has been introduced in various levels of education, i.e. from elementary to high schools, as well as universities. In addition to these (organized) education institutions, it is also a popular choice of learning medium for skill training (e.g. in companies) and for voluntary self-motivated pursuit of knowledge (i.e. lifelong learning).

The rapid diffusion of e-learning is suggested to be due to a series of reasons. First of all, users do not have to gather at the same time and place, making it flexible and particularly beneficial for corporate training and off-curriculum studying (e.g. as a hobby)[0]. Moreover, it saves the costs entailed to live classes such as hiring of rooms or professionals [0]. In addition, ease of access to a diverse range of materials makes it suitable to people of all ages, experiences, and interest [0]. Similar to the content, also the pace of learning is possible to regulate and customize, which makes it a convenient tool for learners with different abilities.

Despite these advantages, learners may experience some difficulties in using e-learning systems. Arkorful et al. point out to the lack of social interaction and relation as some of the most important challenges of e-learning [0], which imply a bigger burden of motivation and time management skills to overcome contemplation and remoteness. In educational psychology, such active commitment, willing participation and involvement of students in school activity is termed -in the broad sense- as “engagement” [0].

Keller et al. list several methods used in e-learning design to sustain or increase users' engagement such as introduction of incongruity or conflict, arousing of curiosity by mystery or unresolved problems as well as variability and changing of pace [0]. The incorporation of such stimuli initially requires assessment of users' state from an engagement point of view.

In this respect, this study proposes using visual feedback from individuals to infer about engagement. Namely, we focus on the frontal view video footage of the users and search for indications of decline in level of engagement. To that end, we propose several features derived from eye blink patterns and seek for a correlation between those and coded levels of engagement (by professional teachers). In this manner, we demonstrate that number and duration of blinks as well as the aspect ratio of the eyes present correlation with engagement level. 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 the e-learning system 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 of motivational advice or interactive content (e.g. with an avatar). Additionally, with the proposed method, it is possible to build person-specific estimators by a simple calibration of the fundamental models, which bears the potential to adjust to interpersonal variations in behavior (specifically, blink patterns).

1.2 Background and Related Work

In educational psychology the term “engagement” refers in the broad sense to active commitment, willing participation and involvement of students in school activity [0]. In particular, (student) engagement is regarded to be governed by three variables as behavior, cognition and emotion. In what follows, we provide an overview of these variables [0].

In conventional classroom settings, behavioral engagement relates attendance, participation, completion of assignments etc. 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 [0]. However, as pointed out by [0], there are a number of problems with such metrics. First of all, analysis of logged data is not easy to incorporate on-the-fly, i.e. in a live session of e-learning. In addition, since tests and assessments are frequently supervised by proxy, it may be difficult, if not impossible, to control activities such as cheating or piracy.

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 externally visible and require self-reporting. Even though there have been some attempts to quantify it by either behavioral indicators (e.g. time on task) or gauging cognitive processes such as reflection, interpretation, synthesis, or elaboration, a clear distinction between behavioral and cognitive engagement could not be achieved [0].

Emotional engagement includes positive or negative emotions towards learning, classmates, or instructors, etc. and may at times be seen through visible expressions of positive emotion in addition to self-reporting [0]. For this reason, it is often used in assessment engagement state of e-learning users. Monkaresi et al. use computer vision techniques to extract several features from videos, such as heart rate and animation units. However, their method needs a physiological device to calibrate the remote HR monitor. Kamath et al. introduce a system for automatic recognition of students engagement levels during e-learning sessions using a crowd-sourced discriminative learning approach [0], while Altrabsheh et al. use real students' feedback in trainining [0]. Grafsgaard et al. introduce an automated analysis of fine-grained facial movements that occur during computer-mediated tutoring [0]. They track fine-grained facial movements consisting of eyebrow raising (inner and outer), brow lowering, eyelid tightening, and mouth dimpling. Kaur et al detect eye gaze and head pose features using OpenFace and feed them to a Deep Multi-Instance Network [0]. Chickerur et al propose recognizing facial expression using 3D models of Kinect [0] for recognizing emotional states such as normal, happy, sad, surprised and angry.

Similar to these studies, we consider students' state of engagement to be evident in their facial features and propose a method based on eye blink patterns.

Section 2

Data Set

In this chapter, we introduce the systems and methods used in collecting our data set.

2.1 Experiment Task and Setup

In order to collect a data set, which enables a continuous observation of evolution (i.e. decrease) of engagement levels, we designed three kinds of tasks, which require different levels of user involvement as passive, semi-active and active tasks.

In the passive task, users are required to watch a slide show of images, which are selected from a benchmark saliency data set involving indoor and outdoor images from 20 categories [0]. Each image is displayed for a duration of 5 seconds separated by a reset image and a blank screen, displayed for 2 seconds and 1 second, respectively.

In the semi-active task, the users listen to the narration of a story in English accompanied by illustrations, requiring listening comprehension skills. At the end of each story a multiple choice question is displayed together with 4 options. We consider the narration part to be a passive task, where the user needs to comprehend the information, and the subsequent test to be an active task, which requires reasoning, deduction, and inference.

The active task is Wisconsin card sorting [0], which is a common tool in neuropsychology for examining the functioning of the frontal lobe [0]. 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 user needs 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.

As experimental equipment, we used a Lenovo Ideapad Y700 notebook PC running on Intel Core i7 processor with 16 GB DDR memory. Its internal web camera affords 1.0MP HD recording and is used to record user's upper torso motion while he performs the tasks. The recorded 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 products.

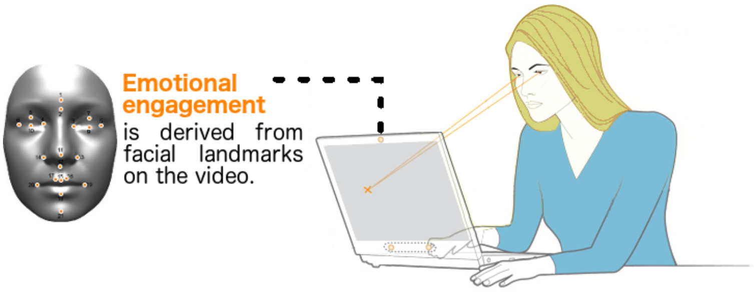


Figure 2.1 Experimental setup.

2.2 Assessment of Engagement

We prepared 10 second long video clips extracted from each session at various stages, which is a common evaluation approach in many other engagement estimation studies [0]. A total of 15 such short clips are prepared from each session. This enables a comprehensive overview of user behavior spanning the entire term of the experiments and also provides convenience to the expert coders in terms of time and effort.

As expert coders (henceforth, Coder-1 and Coder-2), we ask two licensed teachers to annotate (i.e. code) the short video clips. These coders practice regularly teaching languages and have vast experience of interacting with students in conventional classroom settings. They give an engagement label e to each clip on a Likert scale from 1 to 5, where $e = 1$ represents the highest level of engagement and $e = 5$ denotes complete lack of focus.

By examining the distribution of labels given by Coder-1 and Coder-2 for the three kinds of tasks , we observe that the coders agree that the active task is performed with a relatively higher rate of engagement in general, whereas the passive task causes an apparent lack of engagement. In addition to these qualitative remarks, we carried out a methodological analysis of inter-rater agreement based on Krippendorf's α coefficient, and found out that the coders have a satisfactory level of agreement.

Section 3

Method

The locations of certain points around facial components and contour are important in capturing the rigid and non-rigid facial deformations due to facial expressions. Such points are often referred as "facial landmarks". Over the years, many facial landmark detection algorithms have been developed to automatically detect those key points. One of the recent works in this field is by Kazemi et al. [0], which is accepted as state-of-the-art in terms of speed and accuracy [0]. The principles of [0] constitute the basis for the landmark detection tools of the Dlib toolbox [0], which is is a C++ toolkit containing machine learning algorithms for creating complex software in C++ to solve real world problems.

In this study, we employed Dlib in deriving the facial landmark points in our videos. Dlib considers a set of 68 landmark points as depicted in Figure . These 68 point mappings were obtained by training a shape predictor on the labeled iBUG 300-W data set. Among the 68 landmarks points derived by Dlib, we focus on two sets of landmarks concerning the right and the left eye, each of which involves 6 points (see Figure). There are several features that can potentially be derived from landmark points, which are known to contain relevant information regarding user engagement.

3.1 Detection of Blinks

Blink is the fast closing and reopening of human eye. Blinking can roughly be categorized into three as (i) spontaneous blinks that occur as a motor process without external stimuli or internal effort (ii) reflex blinks due to an external stimulus, and (iii) voluntary blinks [0].

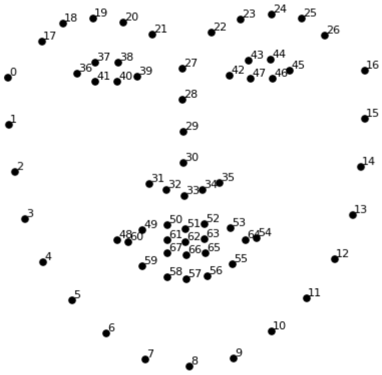
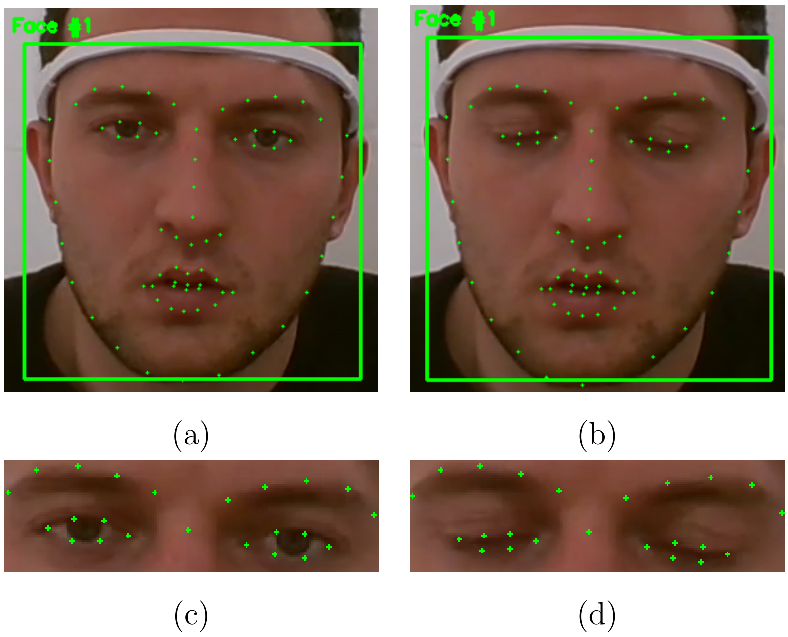


Figure 3.1: The set of 68 landmark points detected by the pre-trained Dlib shape-predictor-68.

In our experiments, tactile stimulus is not present, whereas the degree of optical or auditory stimuli is not to a significant degree to cause reflex blinks. Moreover, since the users are not aware about the observation of their blinking patterns and are neither instructed to blink intentionally, they are assumed not to perform any voluntary blinks. Therefore, only spontaneous blinks are assumed to take place.

Various empirical studies demonstrated the relation between spontaneous blinks and engagement. In particular, blinking is shown to help humans disengage from the outside stimuli in favor of the inertial processing [0]. Therefore, it can be considered as an embodiment of “mind wandering” [0].

In addition to these findings in universal settings, blinking is particularly important in e-learning scenarios, since the learning material is presented though a monitor, which is in most cases an LCD display. Due to the intensity and frequency of the emitted light, the user may feel visual fatigue (or eyestrain) over a certain duration of time, which may affect his blinking pattern as well [0]. Obviously, visual fatigue may potentially arise from the same reasons that cause disengagement, i.e. visual and mental workload [0]. In particular, it is shown that task disengagement scales substantially, correlated with aspects of visual fatigue [0]. Namely, blinks are be suppressed in response to increased visual workload and this inhibition in turn may cause drying of the eyes followed by a higher blinking rate. On the other hand, blinking can contribute to the emergence of disengagement as well [0]. Nonetheless, the relation between blinking and engagement is established under various visually or mentally demanding conditions. In this study, we follow the real-time blink



(a) (b)

(c) (d)

Figure 3.2: Facial landmarks as eyes are (a) open, and (b) closed; and (c,d) corresponding eye regions.

detection method proposed by Soukupova et al. [0]. This study presents a very simple and yet powerful principle to detect blinks. Namely, eye aspect ratio is computed using the landmark points. Eye aspect ratio associated with the left eye in Figure is computed as follows,

$$r_L = \frac{|p_{37} - p_{41}| + |p_{38} - p_{40}|}{2|p_{36} - p_{39}|} \tag{3.1}$$

Eye aspect ratio concerning the right eye, r_R, is computed in a similar manner to Equation 1. Obviously, this value is independent of the roll angle of the head. Although it depends on yaw and pitch angles, we assume extreme rotations do not take place, since users continuously watch the screen while performing their task.

It is known that each individual has a somewhat different pattern of blinks with varying duration, speed of closing and opening, and degree of squeezing the eye. Soukupova et al. account this variability by training an SVM classifier on examples of blinking and non-blinking patterns. Here, we simply consider cumulative distributions of r_L and r_R and adopt the local minima ρ of these distributions as decision threshold. Specifically, we first compute the histogram of blink thresholds and then apply a smoothing operation with a Hanning window and compute the local minima over this smoothed curve. Figure depicts

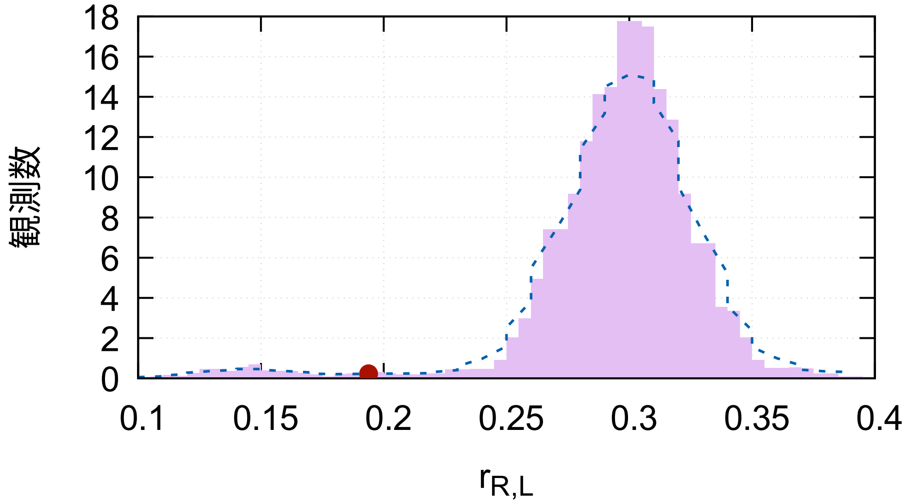


Figure 3.3: Histogram of $r_{R,L}$ for a sample user, corresponding smoothed curve and local minima.

this process for a sample user. Moreover, the benefit in determining blink thresholds in a person-specific manner proves useful, when we compare the distribution of different users.

3.2 Derivation of Features

Using the landmarks derived from the video clips, we define a set of features as follows.

- Interocular distance d_{io} and biocular distance d_{bo}: Depth of the user is defined by how far he is from the screen. A strong indicator of depth is the interocular distance[1]. Numerous studies use depth as a marker of engagement such as [0], and they show that inter-ocular distance is inversely correlated with engagement.

  In this study, we take a similar approach and rather than measuring explicitly depth of the user, we employ inter-ocular distance, d_{io}. Using the landmark map depicted in Figure , we can define inter-ocular distance as $d_{io} = |p_{39} - p_{42}|$. Moreover, in order to increase stability, we also employ bi-ocular distance, $d_{bo} = |p_{36} - p_{45}|$.

[1] In our specific experimental setting, depth is assumed to independent of head pose. Since users need to watch the screen to carry out their task, and we consider extreme yaw and pitch rotations do not take place and roll angle clear does not affect inter-ocular distance. Similarly, inter-ocular distance is not affected by facial expression either.

- Number of blinks n_b: Assuming that $r(t)$ represents the average aspect ratio for the both eyes at time t
$$r(t) = \frac{r_L(t) + r_R(t)}{2},$$
we basically count how many times a blink is initialized by the closing of the eye,
$$n_b = 0.5 \left(1 - \frac{\mathrm{d}\left(\mathrm{sign}(r(t) - \rho)\right)}{\mathrm{d}t} \right).$$

- Duration of blinks t_b: When the user blinks, we measure how long the blink lasts and consider the average duration of blinks in each video clip as a feature. At first, we compute the number of the frames N_b when eyes are closed,
$$N_b = 0.5 \left(1 + \mathrm{sign}(r(t) - \rho)\right).$$
Dividing N_b by the number of blinks n_b, t_b is regarded as the average duration blinks,
$$t_b = \frac{N_b}{n_b}.$$

- Aspect ratio $\bar{r}_o$ for open eyes: Determining the blink threshold, the time instants when the eyes are open t_o and the time instants when the eyes are closed t_c, are found as,
$$t_o = \{t |\ r(t) > \rho\},$$
$$t_c = \{t |\ r(t) < \rho\}.$$
The time interval(s) when the eyes are closed t_c, is treated by the features relating the duration and frequency of blinks, t_b and N_b. For the remaining time intervals(s), we propose using the average aspect ratio $\bar{r}_o$ is the average of the set r_o, where
$$r_o = \{r(t) |\ r(t) > \rho\}.$$

- Normalized eye size a_o: We compute the areas of the right and left eyes as the area of the hexagon composed of 6 landmarks points defining each eyes. However, this value is obviously not independent of the depth of the user (i.e. his distance to the screen). Thus, we divide it by the square of inter-ocular distance. Assuming a_L and and a_R are the areas of the polygons defining the left and the right eyes, respectively, a_o is,
$$a_o = \frac{a_L + a_R}{2d_{io}^2}.$$

The features listed above are assumed to have a correlation with the engagement labels e given by the coders. To verify this assumption, in what follows, we present a statistical analysis based on polyserial correlation.

Table 3.1 Polyserial correlation values for the proposed features.

Feature	User-1		User-2		User-3	
	Coder-1	Coder-2	Coder-1	Coder-2	Coder-1	Coder-2
d_{io}	-0.45	-0.62	-0.61	-0.58	-0.12	-0.61
d_{bo}	-0.49	-0.66	-0.57	-0.55	-0.17	-0.64
n_b	0.23	0.51	0.48	0.39	0.16	0.41
t_b	0.43	0.04	0.01	0.14	0.29	0.59
$\bar{r}_o$	-0.76	-0.78	-0.62	-0.72	-0.58	-0.23
a_o	-0.71	-0.74	-0.48	-0.57	-0.72	-0.51

3.3 Verification of Features

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. We employ Polyserial correlation to determine whether there is a relation between the features defined in Section and the coded values of level of engagement.

As for the numerical variable, we consider the proposed features described in Section and for the ordinal variable we use the labels assigned by the expert coders. For estimating the correlation values, we opt for using a maximum likelihood approach, which maximizes the bivariate-normal likelihood with respect to thresholds (i.e. the ordinal variable). For optimization, a general-purpose method based on Nelder-Mead, quasi-Newton and conjugate-gradient algorithms is used [0]. In implementation, we used rpy2 package which is a back-end for R programming language of statistical computing to Python [0].

This shows that duration of blinks t_b and number of blinks n_b have positive correlation with the engagement. Namely, when level of engagement decreases (i.e. assigned label increases), the average duration of blinks get longer and the number of blinks increase as well. This finding is inline with the suggestion of [0], which states that blinking helps humans disengage from the outside stimuli, which in our case is the e-learning material, in favor of the other cognitive processing. It would be reasonable to assume that when the duration of blinks is longer, the user is likely to get bored and sleepy.

On the other hand, inter-ocular distance, d_{io}, bi-ocular distance d_{bo}, eye aspect ratio r_o, normalized eye size a_o have positive correlation with the level of engagement (i.e. a negative correlation with the assigned label). This indicates that when the user is concentrated on the task, he looks at the screen more often and constantly; and his face is close to the screen.

3.4 Deriving Probability Distributions of Features

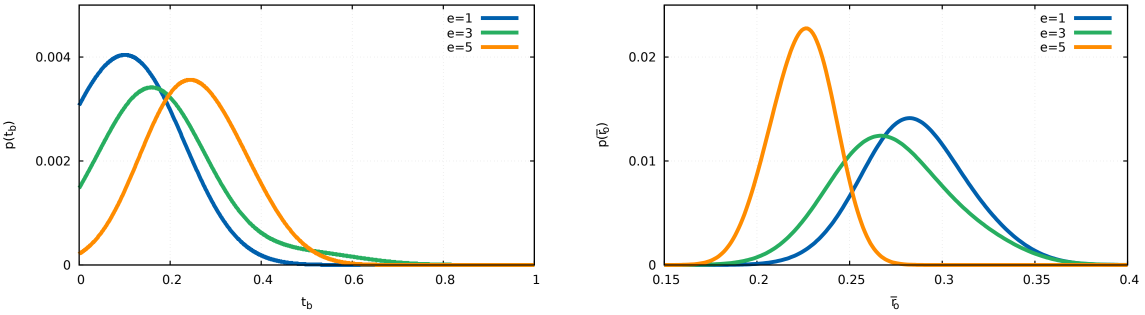
In order to derive probability density function (pdf) of the features described in Section from empirical observations, we utilize Kernel density estimation (KDE). In statistics, KDE is a non-parametric way to estimate the probability density function of a random variable. Since, it learns the shape of the density from the data automatically, it offers a number of advantages, and thus is one of the common methods to estimate the underlying probability density function of a data set.

Let $(x_1, x_2, \ldots, x_n)$ be a univariate random and identically distributed sample drawn from some distribution with an unknown density f. KDE can be expressed as follows,

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

where K is the kernel (i.e. a non-negative function) and $h > 0$ is a smoothing parameter called the bandwidth, and thus is a hyper-parameter of KDE. In most cases, using a Normal distribution as a kernel is considered to give satisfactory results. Although choosing the scaling parameter h as small as the data allows is preferable, there is always a trade-off between the bias of the estimator and its variance. Therefore, bandwidth selection is an inherent problem of kernel density estimation and limits the application of the cross-validation estimate. Since the bandwidth estimate selected by the least squares cross-validation is known to be subject to large sample variation, as a remedy we used grid search over a given interval at evenly spaced points.

For the sake of brevity, in Figure we provide examples on the kernel density estimation (of pdf) for two features, namely t_b and $\bar{r}_o$. Besides, for both features, we provide a comparison for 3 values of engagement labels, namely $e = 1$ (fully engaged), $e = 3$ (medium engaged), and $e = 5$ (completely disengaged). From this figure, it is clear that as level of engagement increases (i.e. e decreases), there is a tendency of observing longer blinks. In other words, the peak of the t_b distribution shifts towards larger values for growing values of e. On the



(a)

(b)

Figure 3.4 KDEs for (a) t_b and (b) $\bar{r}_o$ for three levels of engagement.

other hand, regarding $\bar{r}_o$, as the level of engagement decreases (i.e. e increases), the expected eye aspect ratio decreases, indicating an apparent lack of engagement. In addition, these observations are in line with the findings presented in Table .

3.5 Resolving Level of Engagement Probabilistically

In estimating the level of engagement, we adopt a probabilistic approach. Without loss of generality, let us consider a single feature. For instance, for the duration of blinks t_b, we obtain a set of N_f observations for each video clip, where N_f is the number of frames. By evaluating these observations in the kernel density estimate of t_b relating $e = 1$, we compute the likelihood that this set comes from a fully engaged user,

$$L_e(t_b) = \prod_{\forall t} p\left(t_b(t)|e = 1\right).$$

Assuming independence of features, we compute the likelihood that the user is fully engaged, $L_e = L_e(d_{io})L_e(d_{bo})L_e(n_b)L_e(t_b)L_e(\bar{r}_o)L_e(a_o)$. Similarly, by evaluating this observation set in the kernel density estimates relating $e = 5$, we compute the likelihood L_d that these observations come from a completely disengaged user. Considering that the two values of e, $e = 1$ and $e = 5$, define the two extremes, we can derive in an empirical way, the probability of being engaged p_e and the probability of being disengaged p_d, where

$$p_e = \frac{L_e}{L_e + L_d},$$

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

Section 4

Results

We can apply this approach to each individual feature to evaluate the effectiveness of those features in estimation of level of engagement. In addition, we can apply it to the group of all features in order to assess the potential of integration of the information coming from various features.

In Figure , we demonstrate the probability of being engaged for each of the six features described in Section as well as the set of all features. From this figure, it is clear that there is a tendency of decrease in probability of being engaged for clips with increasing values of e (i.e. decreasing level of apparent engagement). In the case of interocular distance d_{io} and biocular distance d_{bo}, this probability is monotonically decreasing, but this property is not satisfied for some other features such as a_o or t_b. Nevertheless, the overall tendency still presents supporting evidence for the efficacy of the proposed features. In addition, by integrating the information from all the features, we obtain a clear improvement in estimation of engagement. This integration not only produces a monotonic decrease but also it yields a clear separation between values of e from 1 to 3 (i.e. fully engaged to medium engagement) and values of e from 4 to 5 (i.e. low engagement and complete disengagement). In particular, we see that when the user is engaged ($e \in [1, 3]$) the probability of being engaged p_e is over 0.90, whereas it decreases sharply to approximately 0.60 as e drops to 4 and further to 0.20 when $e = 5$. These findings suggest that by estimating the probability of being engaged with the proposed method and setting s threshold for p_e for some value around 0.80, we can detect the cases with low engagement with a satisfactory accuracy.

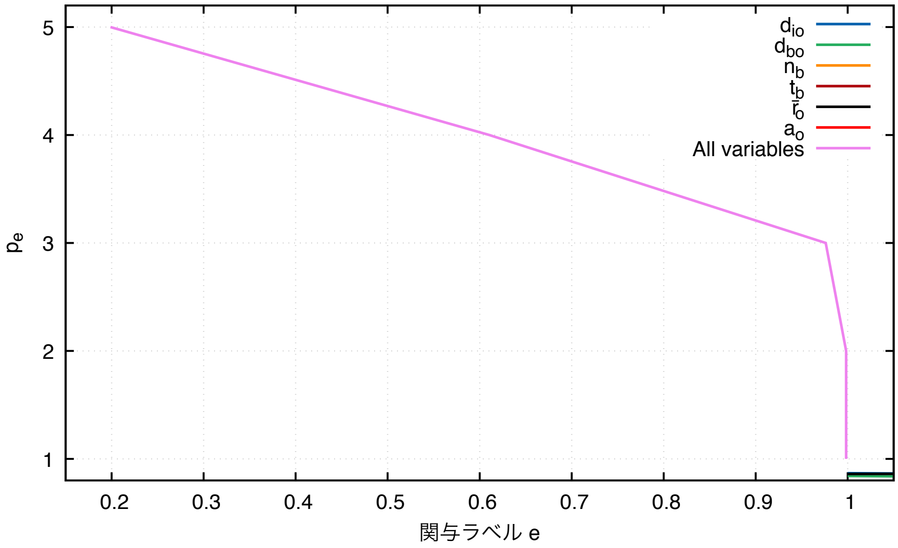


Figure 4.1: The probability of being engaged computed using each of the six features as well as integration of all features.

Section 5

Conclusion

This study proposes using visual feedback from e-learning users to infer about their state of engagement. To that end, we focus on the face area in the video footage of the and compute their facial landmarks. Several features are derived, particularly from the landmarks around the eyes, concerning eye blink patterns and eye size, which are all shown to be correlated with the ground truth levels of engagement. By building a probabilistic method based on the empirical observations of such feature distributions, we assess the level of engagement probabilistically and shown that a significant accuracy is achieved.

The proposed approach can potentially be integrated into the e-learning system so as to provide on-the-fly assessment of engagement, which potentially enables stimulation of the user immediately upon detection of a decline in level of engagement. Additionally, with the proposed method, it is possible to build person-specific estimators by a simple calibration of the fundamental models, which bears the potential to adjust to interpersonal variations in behavior (specifically, blink patterns).

Acknowledgment

First, I would like to thank my advisor, Professor Akito Monden, for his helpful comments. I am deeply grateful to Assistant Professor Zeynep Yücel for her continuing support, extensive discussion and encouragement. I would also like to thank the past and present members of Monden Laboratory for their kind help in my research and life. At last, I would like to thank my family, who have gave me daily support during my years of work.

References

[1] A. Beinicke and T. Bipp, "Evaluating training outcomes in corporate e-learning and classroom training," *Vocations and Learning*, pp. 1–28, 2018.

[2] G. M. Piskurich, "Online learning: E-learning. Fast, cheap, and good," *Performance Improvement*, vol. 45, no. 1, pp. 18–24, 2006.

[3] E. O'Donnell, S. Lawless, M. Sharp, and V. P. Wade, "A review of personalised e-learning: Towards supporting learner diversity," *Int. Journal of Distance Education Technologies*, vol. 13, no. 1, pp. 22–47, 2015.

[4] V. Arkorful and N. Abaidoo, "The role of e-learning, advantages and disadvantages of its adoption in higher education," *Int. Journal of Instructional Technology and Distance Learning*, vol. 12, no. 1, pp. 29–42, 2015.

[5] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of Educational Research*, vol. 74, no. 1, pp. 59–109, 2004.

[6] J. Keller and K. Suzuki, "Learner motivation and e-learning design: A multinationally validated process," *Journal of Educational Media*, vol. 29, no. 3, pp. 229–239, 2004.

[7] G. Ben-Zadok, M. Leiba, and R. Nachmias, "Drills, Games or Tests? Evaluating Students' Motivation in Different Online Learning Activities, Using Log File Analysis," *Journal of E-Learning and Learning Objects*, vol. 7, no. 1, pp. 235–248, 2011.

[8] S. C. Kong, "An evaluation study of the use of a cognitive tool in a one-to-one classroom for promoting classroom-based dialogic interaction," *Computers & Education*, vol. 57, no. 3, pp. 1851–1864, 2011.

[9] C. R. Henrie, L. R. Halverson, and C. R. Graham, "Measuring student engagement in technology-mediated learning," *Computers & Education*, vol. 90, pp. 36–53, 2015.

[10] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," in *Proc. IEEE Winter Conf. Applications of Computer Vision*, pp. 1–9, 2016.

[11] N. Altrabsheh, M. Cocea, and S. Fallahkhair, "Predicting students' emotions using machine learning techniques," in *Proc. Int. Conf. Artificial Intelligence in Education*, pp. 537–540, Springer, 2015.

[12] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester, "Automatically recognizing facial expression: Predicting engagement and frustration," in *Educational Data Mining 2013*, 2013.

[13] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *Digital Image Computing*, pp. 1–8, IEEE, 2018.

[14] S. Chickerur and K. Joshi, "3D face model dataset: Automatic detection of facial expressions and emotions for educational environments," *British Journal of Educational Technology*, vol. 46, no. 5, pp. 1028–1037, 2015.

[15] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015.

[16] R. K. Heaton, G. J. Chelune, J. L. Talley, G. G. Kay, and G. Curtiss, *WCST: Wisconsin card sorting test.* Psychological Assessment Resources, 1993.

[17] O. Monchi, M. Petrides, V. Petre, K. Worsley, and A. Dagher, "Wisconsin card sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging," *Journal of Neuroscience*, vol. 21, no. 19, pp. 7733–7741, 2001.

[18] C. Thomas and D. B. Jayagopi, "Predicting student engagement in classrooms using facial behavioral cues," in *Proc. Workshop on Multimodal Interaction for Education*, pp. 33–40, ACM, 2017.

[19] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1867–1874, 2014.

[20] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *Int. Journal of Computer Vision*, pp. 1–28, 2017.

[21] D. King, "Dlib C++ Library." `http://dlib.net/`, 2018. [Accessed 2018-12-31].

[22] P. Wolkoff, J. K. Nøjgaard, P. Troiano, and B. Piccoli, "Eye complaints in the office environment: precorneal tear film integrity influenced by eye blinking efficiency," *Occupational and Environmental Medicine*, vol. 62, no. 1, pp. 4–12, 2005.

[23] D. Smilek, J. S. Carriere, and J. A. Cheyne, "Out of mind, out of sight: eye blinking as indicator and embodiment of mind wandering," *Psychological Science*, vol. 21, no. 6, pp. 786–789, 2010.

[24] J. W. Schooler, J. Smallwood, K. Christoff, T. C. Handy, E. D. Reichle, and M. A. Sayette, "Meta-awareness, perceptual decoupling and the wandering mind," *Trends in Cognitive Sciences*, vol. 15, no. 7, pp. 319–326, 2011.

[25] M. Rosenfield, "Computer vision syndrome: a review of ocular causes and potential treatments," *Ophthalmic and Physiological Optics*, vol. 31, no. 5, pp. 502–515, 2011.

[26] J. M. Sullivan, "Visual fatigue and the driver," Tech. Rep. UMTRI-2008-50, University of Michigan, Ann Arbor, Transportation Research Institute, 2008.

[27] G. Matthews and P. A. Desmond, "Personality and multiple dimensions of task-induced fatigue," *Personality and Individual Differences*, vol. 25, pp. 443–458, 1998.

[28] T. Soukupová and J. Cech, "Real-time eye blink detection using facial landmarks," in *Proc. Computer Vision Winter Workshop*, 2016.

[29] S. Asteriadis, K. Karpouzis, and S. Kollias, "The importance of eye gaze and head pose to estimating levels of attention," in *Proc. Int. Conf. Games and Virtual Worlds for Serious Applications*, pp. 186–191, IEEE, 2011.

[30] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[31] L. Gautier, “rpy2: A simple and efficient access to r from python.” http://rpy.sourceforge.net/rpy2.html, 2008. [Accessed 2019-01-26].