Predicting Mind-Wandering with Facial Videos in Online Lectures

Taeckyung Lee Dain Kim Sooyoung Park Dongwhi Kim Sung-Ju Lee
KAIST
{taeckyung, amy279, sypark0614, dhkim09, profsj}@kaist.ac.kr
https://nmsl.kaist.ac.kr/projects/attention

Abstract

The importance of online education has been brought to the forefront due to COVID. Understanding students' attentional states are crucial for lecturers, but this could be more difficult in online settings than in physical classrooms. Existing methods that gauge online students' attention status typically require specialized sensors such as eye-trackers and thus are not easily deployable to every student in real-world settings. To tackle this problem, we utilize facial video from student webcams for attention state prediction in online lectures. We conduct an experiment in the wild with 37 participants, resulting in a dataset consisting of 15 hours of lecture-taking students' facial recordings with corresponding 1,100 attentional state probings. We present PAFE *(Predicting Attention with Facial Expression), a facial-video-based framework for attentional state prediction that focuses on the vision-based representation of traditional physiological mind-wandering features related to partial drowsiness, emotion, and gaze. Our model only requires a single camera and outperforms gaze-only baselines.*

1. Introduction

Paying attention to the lecture has a significant impact on students' learning performance [38, 49]. Measuring and maintaining the student's attention is crucial and thus many lecturers monitor the students' attentional state during the lecture. Depending on the attentional state, they could adapt contents during the lecture, intervene with the students to refresh their attention, or provide positive reinforcements [3]. Since many offline classes have been shifted to online due to COVID, grabbing the students' attention has become more important. However, students are more easily distracted during online lectures [21, 25] and catching their attentional status is harder for the lecturers [22].

Recent work has shown that attentional states are predictable with sensors such as eye trackers, electroencephalography (EEG) sensors, electrodermal activity (EDA) sensors, and functional magnetic resonance imaging (fMRI) machines. Therefore, we could use the predicted attention to help students improve learning performance. For example, GazeTutor [15] detects students' attentional states with gaze tracking and provides a dialogue to reengage the students and improve the learning gains. Attention-Aware Learning Technology (AALT) [24] provides student interventions (e.g., asking questions, revisiting contents, and calling the name) based on their mind-wandering predictor with eye trackers [23].

Existing work on machine-predicted attentional states has three limitations when applied to video streaming lectures. First, they require each student to be equipped with special hardware (e.g., eye trackers, EEG sensors, EDA sensors, and fMRI machines). Second, they provide low attentional state prediction accuracy. For example, eye-tracking-based AALT [24] could not provide any interventions to half of the total sessions with a low accuracy model of 0.51 F1 score. Third, they are evaluated only in controlled environments and their performance in realistic environments is unknown. For example, camera-based attentional state prediction is explored over the controlled lab environment for narrative film viewing scenario [46, 47]; and engagement-based approaches (e.g., Student Engagement Analytics Technology (SEAT) [3] require education experts to label the facial video [4].

We present PAFE (**P**redicting **A**ttention with **F**acial **E**xpression), an automatic student attention prediction framework for online video lectures. To collect the facial video with corresponding attentional states, we conduct a fully-online user study in the wild with 37 participants. After removing low-quality data and untrustworthy probing responses, we end up with the 15 hours dataset from 15 participants. From this dataset, we build an attention prediction model with diverse mind-wandering-related features of eye-aspect-ratio, emotion, gaze, and head movement.

PAFE does not require any special hardware except webcams from a student computer used in viewing the lecture. The camera is used for capturing the students' facial ex-

pressions. We utilize findings from mind-wandering with specialized sensors (e.g., emotion [29], drowsiness [2, 44], and eye gaze [5, 7, 11, 17, 23, 45]) for feature extraction and discover that our features, including eye-aspect-ratio (that indirectly indicates partial drowsiness), are highly related to attentional states. Our experimental evaluation indicates that our attention prediction model based on facial features achieves AUROC=0.67, outperforming the gaze-based models (AUROC=0.56). Furthermore, PAFE only utilizes the camera recordings, which are easy to collect, thus can be used for various real-world applications such as real-time student interventions.

We make the following contributions:

- We design and present PAFE, a facial-video-based attentional state prediction framework that focuses on vision-based representation of traditional physiological mind-wandering features.
- We collect and release the first public in-the-wild video dataset for attentional state prediction in online education, consisting of 15-hour facial recordings from 15 students with corresponding attentional state probings.

2. Related Work

Mind-wandering is defined as 'thinking about something else rather than the current primary task' [23, 43]. Mind-wandering occurs when students fail to utilize their working memory resources, both intentionally and unintentionally [38]. Shifts towards mind-wandering occur from 5- to 30-seconds intervals [30], where the general mind-wandering ratio in a certain task is roughly known to be between 20% and 50% [23, 27, 29, 32, 38, 49].

With the increasing importance of the attentional state in education and learning, recent works focus on developing attention-aware learning technology (AALT) to detect and intervene mind-wandering. GazeTutor [15] is an intelligent tutoring system (ITS) that reengages students with dialogues when their gaze is invalid for more than five seconds. Hutt et al. [24] detects mind-wandering with eye-tracker-based model [23] and provides student interventions by asking questions, revisiting the learning material, and calling students' names. Evaluation in high schools showed their intervention system reduced the predicted likelihood of mind-wandering and improved the long-term performance of students.

Predicting attentional states are crucial for education. Most existing mind-wandering prediction schemes utilize specific bio-markers with specialized and expensive hardware: eye-tracker (e.g., gaze, fixation, saccade, and blink) [5, 7, 11, 17, 23, 45], EEG [2, 5, 6, 10, 16, 26, 53], EDA [8, 11, 42], or fMRI [14, 50]. However, they suffer from a significant performance drop when replaced with affordable devices (e.g., eye-tracking [39, 56]).

Facial video-based approaches have been used to detect mind-wandering in online education. Hutt et al. [23] built a multi-modal classifier with eye-tracking features and facial action units. They employed an action unit to improve the original eye-tracking model but still require eye-trackers. Other camera-based approaches are all limited to controlled in-lab environments, targeting narrow focus of narrative film viewing [46, 47] or showing unreliable performances [57].

3. Dataset

Our goal is to overcome the limitations of utilizing specialized devices or working only in controlled in-lab environments. As a first step, we collect the facial video dataset during online lectures in the wild. We carefully designed the experiment (Section 3.1), proceeded with the experiment (Section 3.2), and preprocessed the data to improve data quality (Section 3.3). Finally, we summarize our dataset content (Section 3.4).

3.1. Experiment Design

To obtain a dataset that resembles the real world, we aimed to collect the data with minimum extra cognitive load invoked by the experiment. Therefore, we designed the experience sampling method and selected a target lecture.

3.1.1 Experience Sampling

We use a probe-caught method that utilizes periodic probing to ask whether the participant is focused or mind-wandered. The probe-caught method is known to capture both intentional and unintentional focus shifts [43] and allows us to collect data in fine-grained intervals. Therefore, this method is widely utilized for collecting data for attentional state analyses and predictions in various fields [5, 7, 8, 17, 23, 26, 42, 50].

Determining the probing interval is important: probing too often might cause participants to be conscious of their mental state [38]; while less probing leads to an insufficient amount of data to build the prediction model. Therefore, we first conducted an in-lab pilot study with seven participants to evaluate the effect of probing intervals. Considering the typical period of attentional state shifting [30], we used the 40 seconds probing interval and five of seven participants felt comfortable with this interval.

When probed, participants could select among (i) *Focused* (thinking of anything related to the lecture), (ii) *Not-Focused* (mind-wandering; thinking or doing something unrelated to the lecture), and (iii) *Skip* (participants could not immediately decide the response).

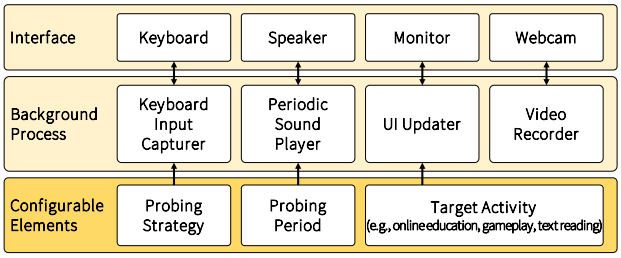

Figure 1. Our experiment program written in Python includes four background processes that asynchronously execute with the main UI process. Our tool is readily applicable to any computer-assisted tasks requiring periodic probing by configuring the probing strategy (e.g., probing options and corresponding keys), probing period, and target activity (e.g., online education, gameplay, and text reading).

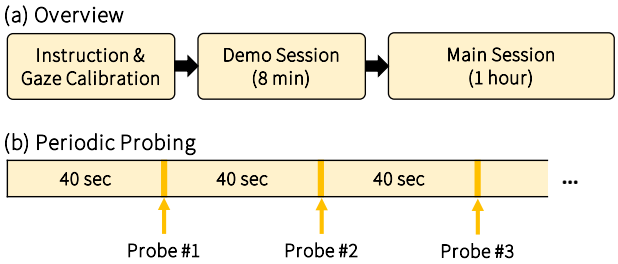

Figure 2. (a) Overview of experiment our procedure. (b) Periodic probing for a demo session and the main session. Note there is a 5 seconds padding when the main lecture starts (omitted in the figure).

3.1.2 Online Lecture

The lecture we use for the experiment should target beginners that do not contain complicated concepts or complex technical terms, which could over-utilize working memory resources and affect the learning performance [52]. Considering our participants are mainly science and engineering majors, we chose "AI For Everyone" by Andrew Ng, a top-rated instructor in Coursera. The lecture consists of a beginner-level introduction to machine learning and deep learning. We merged the first week's lecture videos into one short introductory video for the demo probing session and an 1-hour long video for data collection.

3.2. Participants and Materials

We recruited 51 participants from 7 universities in Korea. The study is approved by the University IRB (ID Number: KH2020-140, KH2021-034) and all participants are over 18 years old. Each participant watched the lecture video at their preferred date in their room alone as a typical online lecture-taking scenario. We encouraged participants to disable any possible distractions such as smartphone alerts, desktop messenger apps, and OS-level notifications, which could cause unexpected focus shifts from the lecture taking to other tasks.

The experiment program was implemented in Python and converted to the executable for Windows and Mac devices. Figure 1 shows the structure of our experiment tool. By running the program, participants are asked to follow the steps in Figure 2: (1) read the experiment instructions, (2) adjust the camera angle and distance, then perform gaze calibration [20], (3) practice the probing process with an 8-min demo lecture, and (4) watch the 1-hour lecture with probing.

During the main lecture session, participants periodically reported the attentional states (i.e., *Focused*, *Not-Focused*, or *Skip*) at the time just before the ding sound by pressing the keyboard button. The probing sounds were played with 40-second intervals (starting from 45 seconds after the beginning of the lecture), resulting in 88 probes for the main session. During the experiment, our program automatically recorded the facial video of each participant in 640p 30fps. In addition, we collected the participants' hardware specifications (e.g., monitor size and resolution) for further gaze-based analysis.

In the end, 37 participants successfully finished the experiment and were rewarded with approximately 25 USD. However, 14 participants either failed to run the experiment program, failed to achieve 640p 30fps, could not fit to experiment schedule, or stopped the experiment at their own will. Note that we clarified that the monetary reward would be provided regardless of their attentional states during the lecture, so they would not manipulate their state to earn the reward.

3.3. Data Preprocessing

While our fully online in-the-wild data collection resembles real world scenario, it has a limitation on data quality. For example, we could not observe face boundaries due to low luminance, or eyes were not distinguishable from the face. Therefore, we removed the video if the scene is too bright or dark, or camera is unstable, which affects the face/gaze detection performance. We then applied the few-shot gaze calibration [33] to remove participants with validation error $\geq 5.0°$. Gaze validation ensures our gaze-tracking error is less than 0.1 cm in a typical lecture-viewing scenario of under one-meter distance between the camera and the eye.

We also resolved the probing quality issue. For example, we observed that few participants were asleep while pressing the *Focused* button. We removed two participants with *Focused* response ratio $< 40\%$, which represents the participants were not paying attention to the lecture, possibly sleeping. Two authors independently labeled the participants' drowsiness state as *Awake* or *Drowsed*, following the

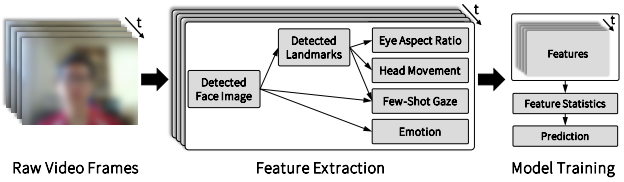

Figure 3. Overview of attentional state prediction pipeline.

mechanism of drowsiness detection [51]. The drowsiness labels agreed with Krippendorff's alpha = 0.76, where Krippendorff's alpha represents the inter-coder reliability [31]. For the mismatched labels, the authors discussed converging to the same label.

To evaluate the participant's probing responses, we built a new metric: attentional state probing unreliability. As participants were fully informed of the labeling strategy, they should not respond with *Focused* when asleep or drowsing. Therefore, we calculated the attentional state probing unreliability U_p for participant p as:

$$U_p = \frac{1}{N}\sum_{t=1}^{N}(Focused_{p,t} \wedge Drowsed_{p,t}) \quad (1)$$

where $Focused_{p,t}$ is a binary value representing participant's *Focused* response and $Drowsed_{p,t}$ is a binary value from drowsiness coding. We removed the entire participants' data with $U_p > 0.1$, considering the human error of probing and drowsiness coding. Otherwise, we only removed *Drowsed* labels while preserving the remaining data.

3.4. Dataset Content

Our finalized dataset consists of 640p 30fps RGB facial video recordings with timestamps, gaze calibration recordings with corresponding screen coordinates, attentional state probings (*Focused / Not-Focused / Skip*), and drowsiness coding. Our dataset includes 15 participants and a total of 1,100 *Focused / Not-Focused* probes. For the 15 participants, seven are male and eight female, with age (Min=19, Max=28, Mean=22.6). Six participants wore glasses.

We open-source our experiment tool built in Python. The program is easily configurable of probing methods, probing interval, and target activity.

4. Predicting the Attentional State

We present PAFE, a facial-video-based attentional state prediction framework. We first select the feature extraction window and extract vision-based features (Section 4.1). Based on statistical analysis on features, we then design the prediction models (Section 4.2). We plot the overview of our system in Figure 3.

Table 1. Vision-based features extracted from the raw facial video. We have 49 features: 6 EAR features, 14 emotion features, 13 gaze features, and 16 head movement features.

Feature	Sub-Feature	Statistic
EAR	-	Mean, SD, 1q, 2q, 3q, MAD
Emotion	Neutral, Angry, Disgust, Fear, Happy, Sad, Surprise	Mean, SD
Gaze	Speed, Dispersion	Mean, SD, 1q, 2q, 3q, MAD
	Horizontal Movement Ratio	-
Head Movement	Translation (x, y, z, RMS), Rotation (x, y, z, RMS)	Mean, SD

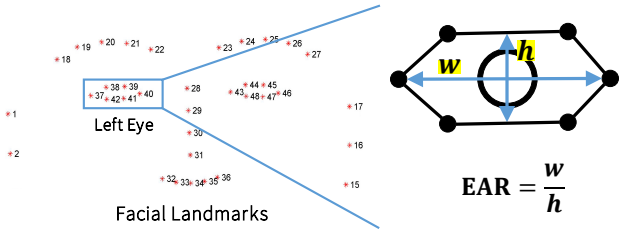

Figure 4. The eye aspect ratio (EAR) represents the ratio of eye width over height, which reflects both short-term effects (e.g., blinks) and long-term consequences (e.g., drowsiness). Landmark annotations are from the 300-W dataset [40].

4.1. Feature Extraction

We extracted physiological features instead of utilizing raw video frames, to provide an interpretable system. Previous mind-wandering studies report that attentional states are related to emotion [29], partial drowsiness [2, 44], and eye gaze [5, 7, 11, 17, 23, 45].

Therefore, we extracted features that indicate emotion, partial drowsiness, or gaze. Precisely, from the raw frames, we extracted facial emotions for predicting emotions, eye aspect ratio, and head movement for partial movement, and appearance-based gaze for eye gaze (see Table 1).

Eye Aspect Ratio: For each facial video frame, we extracted facial landmarks with HRNet [48]. From facial landmark positions, we extracted eye aspect ratio (EAR) [12] representing the eye width with respect to height (Figure 4). Even if the scale of the eye changes, the ratio remains scale-invariant. Moreover, the face detection and landmark detection system is translation-invariant, while the landmark detection and Perspective-n-Point (PnP) algorithm provide rotation-invariance. Therefore, EAR is robust to any possible variances unless the eye is detectable.

Existing works utilized short-term EAR for blink-based task difficulty assessment [13] and blink-based sleeping detection [19]. Instead, we aim to extract partial drowsiness from long-term EAR variations. We applied Kalman filtering to smooth the face bounding box and facial landmark positions to reduce the short-term effects of blinking. Furthermore, to remove the per-participant dependency, we scaled the entire EAR values with median and median-absolute-deviation (MAD) of the first 20 seconds of the data collection, with an assumption that people are fully engaged at the beginning of the experiment. This scaling strategy is applicable to real-world application as we can simultaneously apply the scaling method for any session. Missing values due to face detection failures are linearly interpolated. We extracted mean and standard deviation (SD) to understand EAR containing outliers such as (partially) closing eyes. We also extracted 1q, 2q, 3q, and MAD as outlier-robust statistics to observe long-term effects.

Emotion: We extracted emotional features from the facial analysis framework LightFace [41]. Emotion is classified as a vector of length seven: neutral, angry, disgust, fear, happy, sad, and surprise. Although few sub-features (e.g., disgust and fear) are less likely to occur in lecture-taking scenarios, such features might occur in mind-wandering episodes. Utilizing the full emotion vector aligns with diverse emotions at a workspace [28]. Every emotion vector is scaled to be the sum of 1.0. We extracted the mean and SD from each emotion sub-features.

Gaze: We extracted eye gaze features with a few-shot vision-based gaze-estimation model FAZE [33]. We applied few-shot learning and validation with the gaze calibration data to adapt to an individual's device and eye appearance. As described in Section 3.3, all participants in attentional state prediction have the gaze validation error under $5.0°$. Kalman filter is used for gaze extraction to reduce gaze estimation errors. Moreover, each participant's gaze grid is normalized with one's monitor resolution to be comparable in the exact resolution. Considering the low framerate of webcam recordings compared with eye-trackers, we extracted gaze speed and gaze dispersion (distance from the center of the bounding circle) with statistical values (e.g., mean, SD, 1q, 2q, 3q, MAD). We also extracted the horizontal percentage of gaze movement, which is widely used in gaze-based mind-wandering prediction [7, 23].

Head Movement: With the PnP algorithm, we calculated head movement vector rotation and translation based on facial landmark positions of the nose (landmarks #28 $\sim$ #36) and eye corners (landmarks #37, #40, #43, and #46). Refer to Figure 4 for the landmarks. Note that all landmark positions are Kalman filtered to reduce random noise. We extract mean and SD from rotation and translation for all axes (x, y, z) and root-mean-square (RMS) along all axes. Calculating head movements require the camera intrinsic parameter information. However, to simplify the data collection procedure, we did not ask participants to calibrate their cameras. We assumed that modern digital cameras have square pixel sizes and the principal point is close to the image center to overcome the lack of intrinsic parameters [34].

To reduce the effect of probing in participants' attentional state, we utilized only the last 20 seconds of 40 second intervals for feature extraction. This approach aligns with the previous finding that thought shifts occur from 5- to 30-second intervals [30]. On the other hand, we could remove probing-related behavioral patterns with a windowing strategy. For example, some participants rotated their heads to find the keyboard button, which is unrelated to the lecture.

4.2. Prediction Model

We report important vision-based features by statistical analysis to provide insights into mind-wandering prediction. We compared the value of the facial features (eye aspect ratio, emotion, gaze, and head movement) of window 20 between the *Focused* and the *Not-Focused* groups. The t-test revealed that five emotional features, one gaze feature, one head movement feature, and six EAR features showed a significant difference ($p < 0.05$) between the two groups. Surprisingly, every EAR feature (mean, SD, 1q, 2q, 3q, and MAD) showed strong significance with $p < 0.01$. Table 2 details statistically significant features.

The importance of EAR is interesting; since we removed all drowsed data, the significance of EAR belongs to partial drowsiness, not sleeping. The result aligns with previous findings that partial drowsiness is connected to mind-wandering [2, 44]. Furthermore, we interpret the emotional features confirm mind-wandering is related to negative emotions [29]. In our experiment, participants showed more sadness and less happiness during mind-wandering. Therefore, we conclude that our facial features are strong indicators of mind-wandering.

For attentional state prediction, we utilized 13 statistically significant features of 5, 10, and 20 seconds window sizes. First, we applied traditional machine learning techniques: Support Vector Machine (SVM) with RBF kernel and XGBoost. We also built a simple DNN model with two layers: first DNN layer with 12 nodes and ReLU activation; second DNN layer with a single node and Sigmoid activation. We apply binary crossentropy loss with Adam optimizer.

For the comparison, we generated three baseline models: (1) always predicting the most frequent label, (2) stratified random prediction (predicting with the probability of label distribution of training dataset), and (3) random prediction (predicting with the same probability of 0.5). Moreover, we

Table 2. Statistically significant features ($p < 0.05$) in window 20. A total of 13 features showed significance, while all six features in EAR showed strong significance. EAR is median-scaled for each participant. Emotion and gaze horizontal percentage represent the ratio between 0 and 1. Head translation is represented in 3D real-world coordinates. Statistical results show that EAR is the key indicator of mind-wandering.

Feature	Sub-Feature	Statistic	Focused		Not-Focused		p-value	
			Mean	SD	Mean	SD		
EAR	-	Mean	2.13	2.37	3.34	2.74	<0.001	***
		SD	6.45	4.82	7.55	4.67	0.005	**
		1q	-0.56	0.93	-0.03	1.13	<0.001	***
		2q	0.56	1.31	1.23	1.61	<0.001	***
		3q	2.17	2.24	3.41	2.96	<0.001	***
		MAD	1.26	0.70	1.49	0.79	<0.001	***
Emotion	Angry	SD	0.12	0.10	0.10	0.10	0.033	*
	Happy	Mean	0.07	0.12	0.05	0.12	0.012	*
		SD	0.13	0.12	0.08	0.10	<0.001	***
	Neutral	SD	0.26	0.10	0.23	0.10	0.005	**
	Sad	Mean	0.12	0.10	0.14	0.12	0.006	**
Gaze	Horizontal Percentage	-	0.54	0.08	0.52	0.08	0.002	**
Head	Translation (y)	SD	4.04	4.27	3.17	3.77	0.011	*

generated the gaze-only baseline models as an alternative to existing gaze-based approaches. We utilize XGBoost with 5, 10, and 20 seconds windows for gaze-only models.

We compare the stratified 5-fold classification results in Table 3. Here, folds are divided between participants (leave-several-users-out) to examine if the model could be applied to unseen users. In addition, we used the random under-sampling for training data (except the baseline) to reduce the data imbalance since probing responses consist of $\sim 5\times$ labels of *Focused* than *Not-Focused*.

As a result, XGBoost models showed the highest AUROC, representing the separation capability between two labels. Our model achieves AUROC=0.67, which outperforms the random baseline (AUROC=0.50) and gaze-only baseline (AUROC=0.56). Note that as our model only utilizes a facial video from readily available webcams, our model could be widely deployable than the eye-tracker-based models that require expensive devices.

5. Discussion

5.1. Dataset

Size and Extension: Our dataset contains 15 hours of video recordings of 15 participants, with corresponding 1,100 attentional state probes. Our dataset size is large enough for machine-learning and deep-learning-based methods as we showed comparable accuracy between XGBoost and DNN models. Furthermore, our dataset could be easily extended to other non-real-time lectures by simply loading the video URL (e.g., YouTube links, local video files) in our experiment tool. Our dataset has limited demographic diversity as the study was conducted in Korea and participants were aged from 19 to 28. Therefore, our model and dataset might under-represent groups with unexplored minority, ethnicity, and age.

Primary Task: Primary task for mind-wandering detection must be fixed to the point of interest. Therefore, we restricted participants from intentionally or unintentionally moving the primary task from lecture taking to other tasks (e.g., using smartphones or desktop messenger apps). We asked participants to focus only on lecture taking and required participants to turn off distractions (e.g., smartphone alerts, desktop notifications) and perform the experiment alone in a quiet place.

However, in reality, students could be distracted by events unrelated to the lecture. Some behavioral aspects of distractions are detectable within face detection, head position/rotation, and keyboard/mouse interaction [1, 39].

Effect of Probing: Although self-report probing is widely used in mind-wandering label collection, the effect of probing on attentional states and physiological responses is unknown. In our experiment, 14 of 15 participants (93%) reported that practicing the probing before the main experiment helped them understand the probing strategy (Likert scale $(1 \sim 5) > 3$). In terms of the effect of probing on their attentional state, three participants reported that probing affected their attentional state: *"I got distracted by beeping sounds too often (P1)"*, *"I could actually focus more on the lecture as I do not know when the beep would occur (P2)"*, *"They truly disturbed me, but they refreshed my focus when I was not focusing on the lecture (P8)"*. Still, our probing method did not overwhelm the overall learning process; participants' post-experiment quiz scores were improved by 32% on average, compared with the pre-experiment quiz

Table 3. Each model's mean (± SD) performance for 5, 10, and 20 seconds window data with 5-fold cross-validation.

Method / Window (s)		Focused		Not-Focused		Weighted F1	AUROC
		Accuracy	F1	Accuracy	F1		
Random Baseline	Most Frequent	1.00±0.00	0.90±0.06	0.00±0.00	0.00±0.00	0.75±0.08	0.50±0.00
	Random	0.48±0.02	0.62±0.03	0.47±0.01	0.25±0.07	0.56±0.02	0.50±0.05
	Stratified	0.80±0.01	0.81±0.03	0.22±0.05	0.19±0.02	0.70±0.06	0.49±0.05
Gaze-Only Baseline	XGBoost / 5	0.49±0.04	0.60±0.06	0.56±0.04	0.28±0.07	0.55±0.05	0.54±0.05
	XGBoost / 10	0.48±0.02	0.65±0.04	0.57±0.03	0.27±0.04	0.54±0.04	0.54±0.05
	XGBoost / 20	0.54±0.04	0.66±0.10	0.57±0.03	0.27±0.06	0.60±0.08	0.56±0.03
Proposed Feature-Based Method	SVM / 5	0.53±0.11	0.62±0.25	0.51±0.21	0.25±0.11	0.57±0.19	0.53±0.08
	SVM / 10	0.61±0.09	0.66±0.19	0.50±0.12	0.28±0.07	0.61±0.15	0.56±0.04
	SVM / 20	0.50±0.05	0.65±0.14	0.57±0.08	0.31±0.08	0.60±0.13	0.59±0.12
	XGBoost / 5	0.60±0.05	0.68±0.09	0.54±0.09	0.31±0.10	0.62±0.08	0.59±0.06
	XGBoost / 10	0.58±0.06	0.59±0.16	0.65±0.05	0.36±0.06	0.64±0.12	0.63±0.06
	XGBoost / 20	0.59±0.04	0.71±0.10	0.59±0.02	0.33±0.08	0.65±0.10	**0.67±0.09**
	DNN / 5	0.61±0.09	0.68±0.14	0.52±0.14	0.27±0.10	0.62±0.12	0.56±0.07
	DNN / 10	0.60±0.04	0.71±0.10	0.53±0.03	0.31±0.08	0.64±0.07	0.61±0.03
	DNN / 20	0.60±0.04	0.73±0.09	0.61±0.01	0.35±0.06	0.68±0.10	0.65±0.08

scores. Since such unpredictable effects of probing are inevitable with probe-caught methods, non-invasive methods such as automatic detection are crucial for mind-wandering assessment.

5.2. Prediction Model

Interpretation: Our proposed feature-based model outperforms both random baseline and gaze-only baseline. As appearance-based gaze tracking carries heavy models, we suggest utilizing various landmark-based features rather than gazes for mind-wandering prediction in online education.

Real-World Interference: Since our system is fully facial-video-based, it shares the limitation of face detection, landmark detection, emotion detection, and few-shot gaze adaptation. For example, students should avoid direct light and brighten the scene to detect their faces and eyes. On the other hand, students' movements might affect the model performance. EAR is rotation-/translation/scale-invariant, and so is emotion. Head movement features are designed to detect rotation and translation variances. Since the few-shot gaze-adaptation model [33] is fine-tuned with few shots of a short period, head movements might affect the gaze prediction accuracy with long-lasting lectures.

Utilizing More Vision Techniques: Action Units (AU) are the movement on the face defined by the Facial Action Coding System [18]. A total of 44 AUs represent the different movements of muscles related to face or eyes (30 face-related; 14 eye- or orientation-related). Recent research has investigated the relationship between AU and student engagement during online lectures [54]. Also, there have been approaches to utilize the AU to detect the mind-wandering in the classroom [9]. We expect action units to contribute in vision-based mind-wandering detection.

Heart rate is also known to be an indicator of mind-wandering. AttentiveLearner [35, 36] utilizes heart rate to detect mind-wandering in mobile MOOC learning, where the heart rate is reconstructed with fingers on the smartphone camera. Remote photoplethysmography (rPPG) is a vision-based technique to reconstruct one's heart rate from facial video recordings [37]. HREmo [55] provides student rPPG to lecturers as an index of concentration. We expect rPPG to be further utilized for mind-wandering detection in online education.

6. Conclusion

We propose PAFE, a facial video-based attentional state prediction framework. To build our model, we collected one-of-a-kind vision dataset in real-world-like online lecture scenarios. The dataset contains 15 hours of video recordings of 15 students with corresponding 1,100 attentional state probings. Our physiological feature-based prediction model achieves the accuracy of AUROC=0.67, which outperforms the gaze-only models. The result provides a new opportunity to deploy mind-wandering detection in real-world scenarios. We open-source our dataset and corresponding data collection tool to foster follow-up research.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.NRF-2020R1A2C1004062) and the Institute of Information & communications Technology Planning & evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00900).

References

[1] Nese Alyuz, Eda Okur, Utku Genc, Sinem Aslan, Cagri Tanriover, and Asli Arslan Esme. An unobtrusive and multimodal approach for behavioral engagement detection of students. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, MIE 2017, page 26–32, New York, NY, USA, 2017. Association for Computing Machinery. 6

[2] Thomas Andrillon, Angus Burns, Teigane Mackay, Jennifer Windt, and Naotsugu Tsuchiya. Predicting lapses of attention with sleep-like slow waves. *Nature Communications*, 12(1):1–12, 2021. 2, 4, 5

[3] Sinem Aslan, Nese Alyuz, Cagri Tanriover, Sinem E. Mete, Eda Okur, Sidney K. D'Mello, and Asli Arslan Esme. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. 1

[4] Sinem Aslan, Sinem Emine Mete, Eda Okur, Ece Oktay, Nese Alyuz, Utku Ergin Genc, David Stanhill, and Asli Arslan Esme. Human expert labeling process (help): Towards a reliable higher-order user state labeling process and tool to assess student engagement. *Educational Technology*, 57:53–59, 2017. 1

[5] Carryl L. Baldwin, Daniel M. Roberts, Daniela Barragan, John D. Lee, Neil Lerner, and James S. Higgins. Detecting and quantifying mind wandering during simulated driving. *Frontiers in Human Neuroscience*, 11, 2017. 2, 4

[6] Evelyn Barron, Leigh M. Riby, Joanna Greer, and Jonathan Smallwood. Absorbed in thought: The effect of mind wandering on the processing of relevant and irrelevant events. *Psychological Science*, 22(5):596–601, 2011. PMID: 21460338. 2

[7] Robert Bixler and Sidney D'Mello. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*, 26(1):33–68, 2016. 2, 4, 5

[8] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. Automated physiological-based detection of mind wandering during learning. In Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia, editors, *Intelligent Tutoring Systems*, pages 55–60, Cham, 2014. Springer International Publishing. 2

[9] Nigel Bosch and Sidney K. D'Mello. Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing*, 12(4):974–988, 2021. 7

[10] Claire Braboszcz and Arnaud Delorme. Lost in thoughts: Neural markers of low alertness during mind wandering. *NeuroImage*, 54(4):3040–3047, 2011. 2

[11] Iuliia Brishtel, Anam Ahmad Khan, Thomas Schmidt, Tilman Dingler, Shoya Ishimaru, and Andreas Dengel. Mind wandering in a multimodal reading setting: Behavior analysis & automatic detection using eye-tracking and an eda sensor. *Sensors*, 20(9), 2020. 2, 4

[12] Jan Cech and Tereza Soukupova. Real-time eye blink detection using facial landmarks. *Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague*, pages 1–8, 2016. 4

[13] Youngjun Cho. Rethinking eye-blink: Assessing task difficulty through physiological representation of spontaneous blinking. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. 5

[14] Kalina Christoff, Alan M. Gordon, Jonathan Smallwood, Rachelle Smith, and Jonathan W. Schooler. Experience sampling during fmri reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences*, 106(21):8719–8724, 2009. 2

[15] Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, 70(5):377–398, 2012. 1, 2

[16] Henry W Dong, Caitlin Mills, Robert T Knight, and Julia WY Kam. Detection of mind wandering using eeg: Within and across individuals. *Plos one*, 16(5):e0251490, 2021. 2

[17] Robert E. Bixler and Sidney K. D'Mello. Crossed eyes: Domain adaptation for gaze-based mind wandering models. In *ACM Symposium on Eye Tracking Research and Applications*, New York, NY, USA, 2021. Association for Computing Machinery. 2, 4

[18] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976. 7

[19] Reza Ghoddoosian, Marnim Galib, and Vassilis Athitsos. A realistic dataset and baseline temporal model for early drowsiness detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 5

[20] Pupil Labs GmbH. Pupil labs documentation. https://docs.pupil-labs.com/core/, 2022. [Accessed: 10 March 2022]. 3

[21] Philip J. Guo, Juho Kim, and Rob Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, page 41–50, New York, NY, USA, 2014. Association for Computing Machinery. 1

[22] Khe Foon Hew and Wing Sum Cheung. Students' and instructors' use of massive open online courses (moocs): Motivations and challenges. *Educational Research Review*, 12:45–58, 2014. 1

[23] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R Brockmole, and Sidney K D'Mello.

Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, 29(4):821–867, 2019. 1, 2, 4, 5

[24] Stephen Hutt, Kristina Krasich, James R. Brockmole, and Sidney K. D'Mello. Breaking out of the lab: Mitigating mind wandering with gaze-based attention-aware technology in classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. 1, 2

[25] Scott A. Jensen. In-class versus online video lectures: Similar learning outcomes, but a preference for in-class. *Teaching of Psychology*, 38(4):298–302, 2011. 1

[26] Julia W. Y. Kam, Elizabeth Dao, James Farley, Kevin Fitzpatrick, Jonathan Smallwood, Jonathan W. Schooler, and Todd C. Handy. Slow Fluctuations in Attentional Control of Sensory Cortex. *Journal of Cognitive Neuroscience*, 23(2):460–470, 02 2011. 2

[27] Michael J. Kane, Leslie H. Brown, Jennifer C. McVay, Paul J. Silvia, Inez Myin-Germeys, and Thomas R. Kwapil. For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. *Psychological Science*, 18(7):614–621, 2007. PMID: 17614870. 2

[28] Harmanpreet Kaur, Daniel Mcduff, Microsoft Redmond, Alex C Williams, Jaime Teevan, and Shamsi T Iqbal. "I didn't know I looked angry": Characterizing observed emotion and reported affect at work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. 5

[29] Matthew A. Killingsworth and Daniel T. Gilbert. A wandering mind is an unhappy mind. *Science*, 330(6006):932–932, 2010. 2, 4, 5

[30] Eric Klinger. Modes of normal conscious flow. In *The stream of consciousness*, pages 225–258. Springer, 1978. 2, 5

[31] Klaus Krippendorff. Computing krippendorff's alpha-reliability, 2011. 4

[32] Caitlin Mills, Julie Gregg, Robert Bixler, and Sidney K. D'Mello. Eye-mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human–Computer Interaction*, 36(4):306–332, 2021. 2

[33] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3, 5, 7

[34] Adrian Penate-Sanchez, Juan Andrade-Cetto, and Francesc Moreno-Noguer. Exhaustive linearization for robust camera pose and focal length estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2387–2400, 2013. 5

[35] Phuong Pham and Jingtao Wang. Attentivelearner: Improving mobile mooc learning via implicit heart rate tracking. In Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo, editors, *Artificial Intelligence in Education*, pages 367–376, Cham, 2015. Springer International Publishing. 7

[36] Phuong Pham and Jingtao Wang. Attentivelearner2: A multimodal approach for improving mooc learning on mobile devices. In Elisabeth André, Ryan Baker, Xiangen Hu, Ma. Mercedes T. Rodrigo, and Benedict du Boulay, editors, *Artificial Intelligence in Education*, pages 561–564, Cham, 2017. Springer International Publishing. 7

[37] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10):10762–10774, May 2010. 7

[38] Evan F. Risko, Nicola Anderson, Amara Sarwal, Megan Engelhardt, and Alan Kingstone. Everyday attention: Variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology*, 26(2):234–242, 2012. 1, 2

[39] Tarmo Robal, Yue Zhao, Christoph Lofi, and Claudia Hauff. Webcam-based attention tracking in online learning: A feasibility study. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 189–197, New York, NY, USA, 2018. Association for Computing Machinery. 2, 6

[40] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, 47:3–18, 2016. 4

[41] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5, 2020. 5

[42] Jonathan Smallwood, John B. Davies, Derek Heim, Frances Finnigan, Megan Sudberry, Rory O'Connor, and Marc Obonsawin. Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. *Consciousness and Cognition*, 13(4):657–690, 2004. 2

[43] Jonathan Smallwood and Jonathan W. Schooler. The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology*, 66(1):487–518, 2015. PMID: 25293689. 2

[44] David Stawarczyk, Clémentine François, Jérôme Wertz, and Arnaud D'Argembeau. Drowsiness or mind-wandering? fluctuations in ocular parameters during attentional lapses. *Biological Psychology*, 156:107950, 2020. 2, 4, 5

[45] Lena Steindorf and Jan Rummel. Do your eyes give you away? a validation study of eye-movement measures used as indicators for mindless reading. *Behavior research methods*, 52(1):162–176, 2020. 2, 4

[46] Angela Stewart, Nigel Bosch, Huili Chen, Patrick Donnelly, and Sidney D'Mello. Face forward: Detecting mind wandering from video during narrative film comprehension. In Elisabeth André, Ryan Baker, Xiangen Hu, Ma. Mercedes T. Rodrigo, and Benedict du Boulay, editors, *Artificial Intelligence in Education*, pages 359–370, Cham, 2017. Springer International Publishing. 1, 2

[47] Angela Stewart, Nigel Bosch, and Sidney K D'Mello. Generalizability of face-based mind wandering detection across task contexts. *International Educational Data Mining Society*, 2017. 1, 2

[48] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[49] Karl K. Szpunar, Novall Y. Khan, and Daniel L. Schacter. Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, 110(16):6313–6317, 2013. 1, 2

[50] Liila Taruffi, Corinna Pehrs, Stavros Skouras, and Stefan Koelsch. Effects of sad and happy music on mind-wandering and the default mode network. *Scientific reports*, 7(1):1–10, 2017. 2

[51] Shogo Terai, Shizuka Shirai, Mehrasa Alizadeh, Ryosuke Kawamura, Noriko Takemura, Yuki Uranishi, Haruo Takemura, and Hajime Nagahara. Detecting learner drowsiness based on facial expressions and head movements in online courses. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, IUI '20, page 124–125, New York, NY, USA, 2020. Association for Computing Machinery. 4

[52] Tim van der Zee, Wilfried Admiraal, Fred Paas, Nadira Saab, and Bas Giesbers. Effects of subtitles, complexity, and language proficiency on learning from online education videos. *Journal of Media Psychology*, 29(1):18–30, 2017. 3

[53] Dana van Son, Frances M. De Blasio, Jack S. Fogarty, Angelos Angelidis, Robert J. Barry, and Peter Putman. Frontal eeg theta/beta ratio during mind wandering episodes. *Biological Psychology*, 140:19–27, 2019. 2

[54] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014. 7

[55] Keyu Zhai and Xueming Li. HREmo: A measurement system of students' studying state in online group class based on rPPG technology. *Journal of Physics: Conference Series*, 1976(1):012068, July 2021. 7

[56] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Evaluation of appearance-based methods and implications for gaze-based applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery. 2

[57] Yue Zhao, Christoph Lofi, and Claudia Hauff. Scalable mind-wandering detection for moocs: A webcam-based approach. In Élise Lavoué, Hendrik Drachsler, Katrien Verbert, Julien Broisin, and Mar Pérez-Sanagustín, editors, *Data Driven Approaches in Digital Education*, pages 330–344, Cham, 2017. Springer International Publishing. 2