WEDAR

Webcam-based Attention Analysis via Attention Regulator Behavior Recognition with a Novel E-reading Dataset

Lee, Yoon; Chen, Haoyu; Zhao, Guoying; Specht, Marcus

DOI
10.1145/3536221.3556619

Publication date
2022

Document Version
Final published version

Published in
ICMI 2022 - Proceedings of the 2022 International Conference on Multimodal Interaction

WEDAR: Webcam-based Attention Analysis via Attention Regulator Behavior Recognition with a Novel E-reading Dataset

Yoon Lee
y.lee@tudelft.nl
Delft University of Technology
Leiden-Delft-Erasmus Centre for Education and Learning
Delft, Netherlands

Haoyu Chen
chen.haoyu@oulu.fi
University of Oulu
Oulu, Finland

Guoying Zhao
guoying.zhao@oulu.fi
University of Oulu
Oulu, Finland

Marcus Specht
m.m.specht@tudelft.nl
Delft University of Technology
Leiden-Delft-Erasmus Centre for Education and Learning
Delft, Netherlands

ABSTRACT

Human attention is critical yet challenging cognitive process to measure due to its diverse definitions and non-standardized evaluation. In this work, we focus on the attention self-regulation of learners, which commonly occurs as an effort to regain focus, contrary to attention loss. We focus on easy-to-observe behavioral signs in the real-world setting to grasp learners' attention in e-reading. We collected a novel dataset of 30 learners, which provides clues of learners' attentional states through various metrics, such as learner behaviors, distraction self-reports, and questionnaires for knowledge gain. To achieve automatic attention regulator behavior recognition, we annotated 931,440 frames into six behavior categories every second in the short clip form, using attention self-regulation from the literature study as our labels. The preliminary Pearson correlation coefficient analysis indicates certain correlations between distraction self-reports and unimodal attention regulator behaviors. Baseline model training has been conducted to recognize the attention regulator behaviors by implementing classical neural networks to our WEDAR dataset, with the highest prediction result of 75.18% and 68.15% in subject-dependent and subject-independent settings, respectively. Furthermore, we present the baseline of using attention regulator behaviors to recognize the attentional states, showing a promising performance of 89.41% (leave-five-subject-out). Our work inspires the detection & feedback loop design for attentive e-reading, connecting multimodal interaction, learning analytics, and affective computing.

CCS CONCEPTS

• **Applied computing** → **E-learning**.

KEYWORDS

WEDAR dataset; Attention regulator behaviors; Neural networks

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11.
https://doi.org/10.1145/3536221.3556619

ACM Reference Format:
Yoon Lee, Haoyu Chen, Guoying Zhao, and Marcus Specht. 2022. WEDAR: Webcam-based Attention Analysis via Attention Regulator Behavior Recognition with a Novel E-reading Dataset. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3536221.3556619

1 INTRODUCTION

Keeping a high level of attention is considered a prerequisite for successful learning, being associated with more effective (e.g., comprehension), efficient (e.g., efforts put per time), and appealing (e.g., duration of engagement) learning experiences and outcomes [11, 12, 56]. In this regard, in the fields of learning sciences, multimodal interaction, and affective computing, there have been attempts to measure learners' real-time attention with mind-wandering [37], switches of inner thoughts [69], working memory [19], level of interest [46], and goal-directed thoughts [67]. In this work, we define attention as consciousness towards an ongoing task without an attention redirection. With various sensors and model implementations, attention management through real-time feedback loop design has been endeavored [42, 53]. Significantly, the current transition to hybrid and online learning environments during the pandemic has accelerated the need for attention detection and management in diverse e-learning scenarios.

In e-learning, learners' attention management is different from what they have had in the traditional classroom [3], with limited human educators' involvement and the lack of timely intervention accordingly [61]. Therefore, attention management in e-learning has been highly dependent on learners' self-regulation compared to on-site learning [3]. During e-learning practices, learners experience several iterations of attention fluctuations [29, 70]. In the process, learners recognize their own distractions and try to re-engage in their tasks [49] as a voluntary attentional control [23]. In this work, our focus is on finding learners' self-regulatory behaviors based on learners' own awareness, which leads to self-regulatory efforts to sustain a good level of attention. We define such behaviors as "attention regulator behaviors": Learners' earliest self-awareness of attention loss and following observable behavioral changes as self-regulation. We find those moments important since those are

the moments that learners are willing to and are still able to re-engage in their learning tasks.

In previous studies, diverse multimodal cues have been investigated as observable predictors of subjects' diverse internal states (e.g., cognitive and affective status), such as attention [43, 68], engagement [14, 58], affects [24, 25], and emotion [48]. However, they were often criticized for being difficult to measure or interpret. Iris extension, gaze direction, the position of hands and legs, the style of sitting, walking, standing or lying, body posture, and movement are known to be relevant behaviors for a person's internal states [13, 44, 48]. Diverse parameters of eyes, such as pupil diameter, blinks, and saccades [37], have often been directly used to assess learners' attentional states with dedicated eye trackers. Learners' valence and arousal were often understood primarily through facial expressions [48], with expansion with sensors, such as a photoplethysmograph (PPG), Galvanic Skin Response (GSR), Electroencephalography (EEG), and Electrocardiography (ECG) [38]. Poses and gestures have been interpreted as means to assess engagement [58], affective and cognitive states [24, 25].

However, the current framework has shown that the interpretation of internal states should be understood within the context [5] on macro (e.g., cultural) and micro levels (e.g., situational, personal features) [34]. It indicates that specific cues can be significant indicators of attention in one learning activity, while the same cue does not necessarily represent the same in the other type of learning activity. In this sense, we choose to collect a novel dataset in an e-reading scenario with cognitive and behavioral parameters, which we hypothesize interlinking with attentional changes in e-reading. We chose e-reading since it is the most common and fundamental form of e-learning practice in higher education, which can support other learning activities. This work focuses on which attention regulator behaviors occur following the perceived distraction via the statistical analysis and model implementation. We hope this interdisciplinary study can nurture an understanding of attentive e-reading. Our contributions to the field are listed below.

- To our best knowledge, it is the first attempt to introduce attention regulator behaviors in e-learning for attention analysis and prediction. Compared to conventional subjective measurements of attention, such as self-report, the attention regulator behaviors are easier and intuitive to capture and more objective to evaluate.
- We collected a novel dataset, WEDAR, from 30 subjects with various metrics, including learner status, affects, behaviors, and learning outcomes. Self-report of distraction is also provided as ground truths to verify the effectiveness of the attention regulator behaviors as a predictor.
- Diverse machine learning models are implemented as a baseline to recognize learners' attention regulator behaviors and attentional states. Those baselines can further be applied to diverse e-reading system designs.
- The framework provides a webcam-based attention analysis. It does not require dedicated hardware implementation for obtaining the attention recognition features and can thus be applied to diverse real-world settings.

2 RELATED WORK

2.1 Attention "regulator" behaviors

Diverse learning theories have been constructed to understand learners' internal states through various tangible predictors. Our work is based on the framework of [27], which focuses on how diverse *stimuli* (e.g., external condition, verbal representation, awareness, intentionality, external feedback, delivered information) can be *interpreted* (e.g., arbitrary, iconic, intrinsic) and connected to *functional* nonverbal behaviors (e.g., emblems, illustrators, regulators, affect displays, adaptors).

According to the behavior categorization of [27], "regulator" behaviors occur as a self-regulatory action with the purpose of successful task performance (e.g., head nods, eye contact, slightly forwarded body, small postural shifts, and eyebrow raises in human-to-human interaction). Those are subconscious and habitual actions triggered by behavior agents' "awareness" of their internal and external states (e.g., attention loss). We hypothesize that such self-regulatory behaviors (i.e., attention regulator behaviors) also occur in e-reading. In this work, we try to define the types and frequencies of attention regulator behaviors in e-reading. The framework of [27] also indicates the expandability of their categorical framework, which supports our attempt.

2.2 Multimodal attention recognition in real-world e-reading settings

Previous research has highlighted the importance of contextual interpretation of multimodal indicators [34]. Instead of finding global features for attention in diverse learning scenarios, we explicitly investigate theoretical and empirical behavioral cues of attention regulation in e-reading. We investigate a data collection method that is non-intrusive and closer to real-world settings, which allows a more widespread application of our framework in diverse e-reading scenarios. In the following, we introduce previous research aimed explicitly at real-world implementation based on webcam and mouse-click.

[43] aimed for the subject-independent model development in e-reading based on eyebrow, lip, head movements, and mouse orientation. Specific behaviors (e.g., leaning forward) have been combined and labeled as more generic categories (e.g., body) to avoid overlapping features in different classes. [68] focused on head orientation, eyelid and mouth height, gaze direction, and emotion (i.e., confusion and happiness) during e-reading. Six hand-labeled attention levels (i,e., sleepiness, drowsiness, fatigue, distraction, attention shift, concentration) has been used as ground truths. However, we assume that each attention class is not exclusive enough to the other, so there is a high chance that the machine can not classify different attention levels with higher performance. According to our best knowledge, very little empirical work has been done for attentive e-reading, which premises real-world settings.

2.3 Multimodal attention regulator behaviors

This section explicitly explores multimodal learning behaviors that function as attention regulators in e-reading. Instead of investigating features that can be found with dedicated sensors and devices, we focus on features recognizable to observers.

Eyebrow. The movements of eyebrow has been associated with the activation of cognition [25, 28], arousal [21] and emotions [21, 28], having most of the framework applied to social communication with rare empirical studies [33]. Though eyebrow movement is often observed in e-reading, only several empirical works indicated that eyebrow movements correlate to attentional changes [43]. As far as we know, theoretical behavior frameworks dedicated to e-reading have not been established yet. [26] understood eyebrow movements as ritualized behavior of attention signals, while [21, 28] interpreted it as sign of "wanting to know more", which is connected to the cognitive arousal. The framework of [28] defined eyebrow movements as a representation of surprise, question, and fear. In this work, we focus on the arousal function of eyebrow movements that are shown with combinations of inner and upper brow raise and lowering movements [21]. With a few solid evidence, we hypothesize eyebrow movements as voluntary self-regulatory behavior to re-engage in the task, aiming at a better attention level.

Blink. Correlations between various eye movements and cognitive actions have been revealed in diverse task performance scenarios, such as reading, scene perception, and visual search [54]. Based on the environmental task demands, humans are known to adapt their blink patterns spontaneously, voluntarily, and reflexively [36]. In e-reading, a reduced blinking rate by 4% has been observed with higher perceived fatigue, compared to paper-based reading, having dry eyes and eye discomfort as major causes [17]. As a result of fatigue, changes in blink patterns in frequency and duration are observed [60]. Blink flurries, which are defined as three or more blinks within a 3-second window, occur [17] and are interpreted as a spontaneous effort to sustain the attention and increase wakefulness [4]. Voluntary prolonged blinks are observed as a behavior to reduce the fatigue levels in eyes [16], showing different ranges in interblink interval variability, degree of completeness, duration of the closure, and the force involved, compared to spontaneous blinking [45].

Mumble. Verbalization during reading is one learning strategy known to help readers' cognitive processing, reading development, and comprehension [64]. Verbalization is also known as read out loud, oral reading, and mumble reading, allowing readers to focus and monitor their real-time comprehension, as opposed to silent reading [51]. We use the term "mumble reading" in this work since our target behavior does not indicate active usage of the verbalization technique. However, it is more inclined to semi-spontaneous mumble behavior as a self-regulatory action to achieve better attention. Mumble reading is more commonly applied to teach young learners. However, it is also known to assist adult learners with decoding difficult passages. By mumbling the text, learners internalize the meaning and information of the sentence as coherent sets [51] with auditory stimulation. Diverse eye movement patterns are known to be correlated with mumble reading behavior [41]: Mumble also works as a stimulus to blink [17], showing internal consistency as an attention regulator behavior.

Hand. Self-touch is known to be an action that re-engages people's attention by soothing themselves during stressful moments, causing self-enjoyment [7]. Aside from the stress-release effect, self-touch during the task performance is known to bring better self-regulation, too [7]. Inhibiting effect from such tactile stimulation helps learners ignore distractions and refocus on the task [39]. Especially when working on a task that demands working memory with the presence of distractors, more spontaneous self-touches on the face tend to appear with the increasing necessity of refocusing [47]. Self-touch should also be interpreted within the context since certain self-touching behaviors lead to relaxation, while others work as arousal (e.g., self-squeezing, rubbing, scratching, stroking) [7]. Therefore, we define calming self-touching behaviors on the body and face as one category of attention regulator behaviors.

Body. While the face delivers more information about types of emotions, the body is known to convey affects and intensity [27] of emotions via diverse amplitude, speed, and fluidity of movements [10]. In previous studies, body postures have shown a direct correlation with attentive [55] and affective states [24]. The direction of the body is known to imply the affective states, such as boredom, confusion, delight, flow, and frustration [24] while the leaning forward pose works as a sign of active cognitive state [25]. Head direction indicates the subject of attention [66]. [9] understood postural shift as an action to move on to the next phase during the task performances.

Distraction self-reports. Distraction self-reports are commonly used as the ground truth to reflect people's internal states [30]. The model of [29] introduce two types of distraction: 1) Task-related distraction and 2) task-unrelated distraction. [29] explains that task-related thoughts are correlated with the objective performances, while task-unrelated mind wandering functions as an impairment to the ongoing task performances. In this regard, we collect two types of distraction self-reports in e-reading.

Based on the execution, distraction reports are two types. The first method is to collect distraction reports real-time at the choice of participants during the task performance, putting more importance of more timely aspect of distraction reports. The second method uses a specific time or event to trigger the question regarding the current distraction levels [30]. The first method is criticized as participants might not be aware of their attention loss or forget about reporting. The second method is faulted for bothering the primary task performance.

We implemented the first method since our objective of the distraction self-reports collection is to find behaviors at the moment of learners' perceived distraction, which is used as the ground truth of the model training in our work. To minimize the possible intrusiveness in the self-report process, we carefully designed a simple and intuitive self-report interface, introduced in the following "Distraction self-reports" section. In this way, we obtained the ground truth of the attention levels of every subject through their frame-level distraction self-reports.

3 WEDAR-DATASET

3.1 Participants

30 learners (gender: 15 males, 15 females; age: M=27.89, SD=3.39) in higher education, who use the English language for their daily education, have been invited for an e-reading task. Participants voluntarily joined the experiment via an advertisement on campus.

3.2 Materials

The text "how to make the most of your day at Disneyland Resort Paris" has been implemented on a screen-based e-reader, which

we developed in a pdf-reader format. An informative but entertaining text was adopted to capture learners' attentional shifts during knowledge acquisition. The text has 2685 words, distributed over ten pages, with one subtopic on each page (e.g., how to book tickets online the same day). The e-reader has been implemented on a 13-inch laptop monitor with resolutions of 960×720, having the text with 11 pt. A built-in webcam on Mac Pro and a mouse have been used for the data collection, aiming for real-world implementation only with essential computational devices. A height-adjustable laptop stand has been used to compensate for participants' different eye levels.

3.3 Measurements

We collected various cues that reflect learners' moment-to-moment and page-to-page cognitive states to understand the learners' attention in e-reading. Fig 1 shows an overview of measurements used in the WEDAR dataset collection.

Video recording. Video recordings were made at 30 fps. The recording has initiated with "start session" button and ended with the "end session" buttons on the screen interface, pressed by learners. The collected video recording has an average duration of 16.2 minutes (SD= 5.2 minutes). Video processing will be described in the following "Data analysis and results" chapter.

Distraction self-reports. Learners were asked to report their distractions on two levels during the reading: 1) In-text distraction (e.g., still reading the text with low attentiveness) or 2) out-of-text distraction (e.g., thinking of something else while not reading the text anymore). We implemented two noticeably-designed buttons (33×22) on the right-hand side of the screen interface to minimize the possible distraction coming from the reporting task.

Blur stimuli. We implemented blur stimuli on the text in the random range of 20 seconds after the trigger of a new page. It ensures that the blur stimuli occur at least once on each page. This is based on the finding that average learners read 230-250 words per minute [6]. Participants were asked to click the de-blur button on the text area of the screen to proceed with the reading. The button has been implemented to the whole text area, with 400×480 resolutions, so participants can minimize the effort to find and click the button. Reaction time for de-blur has been measured, too, to grasp the arousal of learners during the reading.

Pre-test and post-test. We asked participants to answer pre-test and post-test questionnaires related to the reading material. Participants were given ten multiple-choice questions before the session, while the same set of questions was given after the reading session (i.e., formative questions) with added subtopic summarization questions (i.e., summative questions). It can provide insights into the quantitative and qualitative knowledge gained through the session and different learning outcomes based on individual differences.

3.4 Procedure

30 learners in higher education have been invited for a screen-based e-reading task (M=16.2, SD=5.2 minutes). A pre-test questionnaire with ten multiple-choice questions was given before the reading to check their prior knowledge level about the topic. There was no specific time limit to finish the questionnaire. Afterward, instructions on secondary tasks were given: 1) Deactivating the blur stimuli on the screen by clicking the text area and 2) reporting distractions (i.e., in-text distraction, out-of-text distraction). Learners were left alone in a room to perform a screen-based reading task. Once participants finished the reading, they were given a post-test questionnaire with the same question set as the pre-test. However, in the post-test questionnaire, there were added questions for summarizing ten subtopics by filling in the sentences starting with " How to…".

3.5 Dataset: WEDAR

The final outcome of the WEDAR dataset is presented in Table 1, including the objectives of data collection, modalities, features, evaluation, and interpretation. In this work, indicators in ***bold*** are used for the attention regulator recognition and attention prediction. Note that the WEDAR is built not only for attention regulator behavior recognition but, more importantly, for exploring the learners' attentional states during the e-reading events. Thus, we collected various metrics as cues of the learners' attentional states, such as reactions to stimuli, distraction self-reports, and knowledge gain. All those metrics were obtained by learners' self-reports. The annotation is frame-level (one value for one reading case) for the metrics of reactions to stimuli and distraction self-reports. The annotation is instance-level for the metric of knowledge gain, which has been measured before and after the reading.

4 DATA ANALYSIS AND RESULTS

This section presents preliminary experimental results conducted on the WEDAR dataset. We first report a relevant statistical analysis of the WEDAR dataset using Pearson's correlation coefficient. Several classical models are implemented as the benchmark for recognizing different attention regulator behaviors. Lastly, high-level attention analysis is conducted using the attention regulator behaviors and attention span.

4.1 Annotation and baseline analysis

Annotation of attention regulator behaviors. The video dataset of 931,440 frames has been annotated with the attention regulator behaviors using an annotation tool that plays the long sequence clip by clip, which contained 30 frames. Two annotators (doctoral students) have done two stages of labeling. In the first stage, the annotators were trained on the labeling criteria and annotated the attention regulator behaviors separately based on their judgments. In the second round, the labels were summarized and cross-checked to address the inconsistent cases. We used six categories that we found to be relevant to attention regulation based on the literature study: Behaviors shown from eyebrow (26,535 frames), blink (17,761 frames), mumble (22,214 frames), hand (101,700 frames), and body (155,880 frames), contrary to the neutral (607,350 frames) state (Figure 2). Since the importance of our work is not merely on the recognition of behaviors itself but on connection with hidden behavioral functions (e.g., attention regulator) [44], we combined multiple specific behaviors (e.g., squint) into a general category (e.g., eyebrow). It also helps avoid redundancy among features [43] which could negatively affect the model's performance. The labeled data has been used as two input formats: images segmented by each frame and videos segmented every second (30 frames).

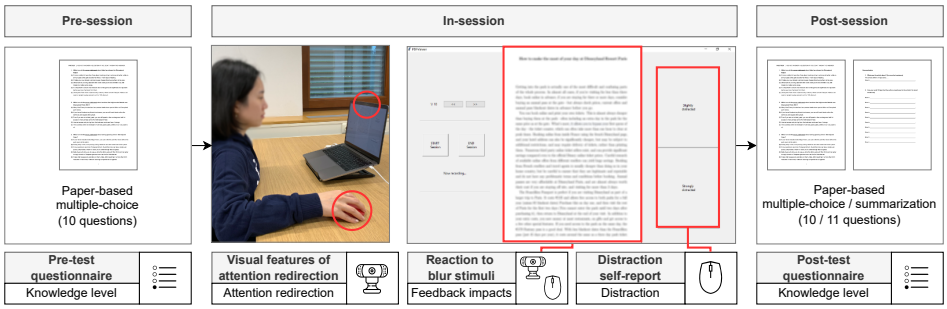

Figure 1: The experiment settings show an overview of our WEDAR dataset collection.

Table 1: Our WEDAR dataset contains diverse dimensions of attention: Objectives, modalities, features, evaluation, and interpretation of attention indicators.

Objectives	Modalities	Features	Evaluation	Interpretation
Learner behaviors	Video (avi.)	-Affective states of learners -Behavioral states of learners	Objective/ Subjective	Short-term attention
Reactions to stimuli	Timestamp (txt.)	-Blur triggered -Blur deactivated -Reaction time	Objective	Short-term attention
Formative & summative assessment	Text (txt.)	-Pre-test (multiple-choice) -Post-test (multiple-choice, summarization) -Knowledge gain	Objective	Long-term/holistic attention
Distraction self-reports	Timestamp (txt.)	**-Distraction in the context of reading** **-Distraction outside the context of reading**	Subjective	Short-term attention

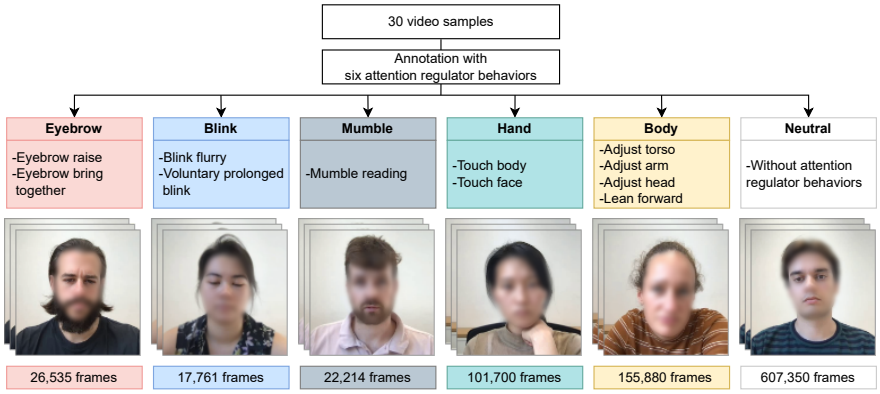

Figure 2: Annotation of attention regulator behaviors: Eyebrow, blink, mumble, hand, body, and neutral.[1]

[1] Images were blurred for identity protection purposes. All images were consented to be used for publication.

4.2 Preliminary analysis: Pearson's correlation

We conducted a preliminary second-to-second analysis using Pearson's correlation among the overall, in-text, out-of-text self-reported distractions and attention regulator behaviors. We aimed at comprehensive insights into how each behavior category can be correlated to perceived distractions. As can be seen from Table 2, the total number of self-reported distractions and in-text distractions showed a significant correlation with eyebrow and body behavior categories. Out-of-text distraction has correlated with the most behavior categories: Eyebrow, blink, hand, and body. Though mumbling did not directly correlate with any types of distractions, it has been correlated with other behavior categories, such as eyebrow, hand, and body. Various behavior categories have shown correlations among each other. The unimodal correlation analysis based on Pearson's correlation coefficient has presented: 1) The internal consistency among the attention regulator behaviors and 2) the potential of attention prediction model training based on multimodal behavioral cues related to attention self-regulation. Note that Pearson's correlation coefficient is a preliminary examination that only shows the linear correlation of two variables, revealing their potential association in the temporal domain. However, when it comes to attention regulator behavior-based distraction recognition, the performance might vary greatly because the relationship between attention regulator behaviors and distraction level is complex and non-linear, which cannot simply be described by Pearson's factor.

Table 2: Preliminary Pearson's correlation[2] analysis between distraction self-reports and attention regulator behaviors.

		Total distraction[3]	In-text distraction	Out-of-text distraction	Eyebrow	Blink	Mumble	Hand	Body
Total distraction	Pearson's r	-							
	(p-value)								
In-text distraction	Pearson's r	**0.938** ***	-						
	(p-value)	**(<.001)**							
Out-of-text distraction	Pearson's r	**0.342****	-0.004	-					
	(p-value)	**(<.001)**	(0.469)						
Eyebrow	Pearson's r	**0.030****	**0.021****	**0.028****	-				
	(p-value)	**(<.001)**	**(<.001)**	**(<.001)**					
Blink	Pearson's r	0.019	0.011	**0.025****	**0.025****	-			
	(p-value)	(0.001)	(0.055)	**(<.001)**	**(<.001)**				
Mumble	Pearson's r	0.004	0.006	-0.006	**0.041****	-0.005	-		
	(p-value)	(0.518)	(0.274)	(0.270)	**(<.001)**	(0.440)			
Hand	Pearson's r	0.006	0.002	**0.012***	**0.053****	**0.028****	**0.045****	-	
	(p-value)	(0.325)	(0.783)	**(0.036)**	**(<.001)**	**(<.001)**	**(<.001)**		
Body	Pearson's r	**0.045****	**0.035****	**0.035****	**0.095****	**0.044****	**0.037****	**0.375****	-
	(p-value)	**(<.001)**	**(<.001)**	**(<.001)**	**(<.001)**	**(<.001)**	**(<.001)**	**(<.001)**	

[2]Note. * p < .05, ** p < .01, *** p < .001, [3]Total distraction=In-text distraction+Out-of-text distraction

Table 3: Attention regulator behavior recognition performance on the test set of the WEDAR.

Method	Framework	Accuracy (%)	
		Subject-dependent	Subject-independent
ResNet-18 + fine-tuning	Frame-level	39.76	25.90
ResNet-50 + fine-tuning		30.92	23.84
ResNet-101 + fine-tuning		31.26	16.39
ResNet-18 + kNN		69.98	18.43
ResNet-50 + kNN		69.95	18.23
ResNet-101 + kNN		69.73	15.76
CNN-RNN-imbalanced	Video-level	**75.18**	**68.15**
CNN-RNN-balanced		**75.70**	**68.43**

4.3 Low-level attention regulator behavior recognition

We propose the benchmark of classical models with two types of frameworks (i.e., frame-level and video-level recognition) on the WEDAR dataset to first recognize attention regulator behaviors.

Here, we followed the classical 70%-30% protocol from other large-scale action/activity datasets, such as ActivityNet [31] and Kinetics-400 [40]. Given 30 video samples with frame-level annotations (931,340 frames), we aimed to recognize the six attention regulator behavior categories accurately. Besides, we conducted an evaluation with both subject-dependent and subject-independent protocols. In subject-dependent protocol, we randomly shuffled all the samples and split the training and testing set with a ratio of 70% and 30%. In subject-independent protocol, we split the subjects with a ratio of 70% and 30%. Thus, all the samples from 21 subjects were used for training, and samples from the remaining nine subjects were used for testing. Note that we used the same protocol and evaluation settings for all the evaluation methods to make a fair comparison. Table 3 shows the overall accuracy of the testing set.

Frame-level attention regulator behavior recognition. In this section, we conduct the attention regulator behaviors recognition using frame-by-frame image inputs. We implemented ResNet architecture as the backbone with its three variants (ResNet-18, ResNet-50, ResNet-101) [35], which are pre-trained on ImageNet [22], and fine-tuningd by fixing the layers $1000d\ fc$ and above. ResNet architecture became one of the most popular architectures in various computer vision tasks. Its shortcut connections architecture yields compelling results. First, each frame has been converted to 224×224 grid RGBs as image inputs. The higher-level features have been extracted with the layer going deeper, combining primitive features from images on earlier layers. To avoid the imbalanced data issue brought by a large number of neutral behaviors, we evenly sampled each category based on the class with the minimum category number (17,761). The number of training features from ResNet-18, ResNet-50, and ResNet-101 were 1000×74778, and testing features were 1000×31788, respectively. All the models have been trained with 32 batch sizes. In the process, fast Stochastic Gradient Descent (SGD) [52] with standard momentum parameters were applied.

Furthermore, we implemented a simple multiclass kNN (k-Nearest Neighbour) classifier stacked to the output features from the layers $1000d\ fc$ of ResNet-18, ResNet-50, and ResNet-101 to achieve the attention regulator behavior recognition. Our rationale lies in the observation that the target dataset (WEDAR) is relatively small and different from the source dataset (ImageNet). The images in the WEDAR are also with high homogeneity. Thus, the fine-tuning of the WEDAR dataset will highly likely make it overfit. Therefore, we implemented the multiclass kNN classifier to verify it. Various k variables have been applied to ResNets for the comparative performance analysis.

Video-level attention regulator behavior recognition. Since attention regulator behaviors are the aggregation of instant actions over the temporal domain, frame-level recognition tends to lose rich, dynamic information. In that sense, we adopted a video-level framework compatible with the video inputs, having a "temporal" feature in its learning process. Comparative analyses have been conducted between frame-based and video-based models to achieve a better recognition result of attention regulator behaviors. Specifically, we implemented a hybrid architecture that consists of convolutions (for spatial information) and recurrent layers (for temporal information). We used a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) consisting of GRU layers [18], popularly known as a CNN-RNN [2, 50]. We chose the InceptionV3 architecture [63] as the CNN backbone, which has

been pre-trained on ImageNet [22], benefiting from its lightweight structure, which is suitable for the temporal modeling. The output features of each frame have been fed into GRU with three layers (with GRU units as 16) and stacked by a fully connected layer as output. Besides, we noticed that imbalanced-data issues brought by a large number of neutral behaviors might affect the model's performance. Thus, we present two types of data sampling strategies: 1) Evenly sampling each category based on the class with the minimal number (balanced) and 2) using all the samples from each category (imbalanced).

Experimental results of the attention regulator behavior recognition. A comparative performance analysis has been conducted among models aimed at recognizing attention regulator behaviors (Table 3). Note that all models followed the same evaluation protocol mentioned above for fair comparisons. 1) Video-level models (CNN-RNN) have shown better performances than frame-level models (75.70% vs. 69.73% in subject-dependent settings and 68.43% vs. 25.90% in subject-independent settings) by large margins, with the more temporal information involved. It means capturing temporal dynamics (temporal reasoning) is important for behavior recognition. 2) The performances of the models vary significantly based on the evaluating protocol. ResNet-kNN architecture performed better than ResNet-finetuning architecture on the subject-dependent protocol. ResNet-kNN (-18, -50, -101) has achieved 69.98%, 69.95%, and 69.73% accuracy, respectively. However, when it comes to subject-independent protocol, ResNet-kNN models have a significant performance drop of more than 50% accuracy, while the performances of Resnet-finetuning models are relatively steady. This result indicates that the high performance of the ResNet-kNN model is benefited from the subject-dependent setting via overfitting to our WEDAR dataset. 3) The comparison between different ResNet variants has shown the best result in ResNet-18 with slight performance differences compared to other models with higher learnable parameters. Because WEDAR is a relatively small dataset, learning could have been converged early with smaller learnable layers. Our result emphasizes the importance of the compatibility of the sizes of the model and datasets [1, 15].

4.4 High-level attention analysis with attention regulator behaviors

This section introduces our attention analysis based on attention regulator behaviors. The task is recognizing the attentional states (i.e., attention or distraction) based on the attention regulator behaviors within a given small video instance.

Evaluation protocols. For the attentional state recognition, as the task is highly subject-dependent, we chose to use a leave-subjects-out protocol to verify the generalizability of the method. We obtained the ground truths of attention and distraction instances from participants' distraction self-reports. We took 8-second duration as an average attention span of human beings based on a literature study [8, 62]. Therefore, we set the last 8 seconds to the moment of distraction self-report as "distraction" while following 8 seconds from the moment of distraction self-report as "attention" state. We also took 16-second, 4-second duration and 2-second duration as comparisons. 383 distraction self-reports have been observed in the dataset, resulting in two sets of 383 × instances of "attention"

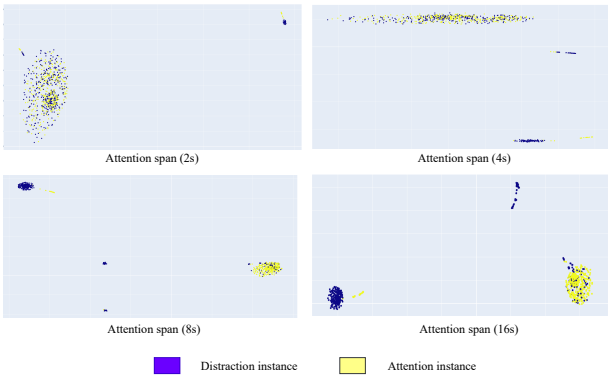

Figure 3: The t-SNE visualization of the features for attentional states. The feature embeddings are obtained based on the attention regulator behaviors happened during the given attention span. Each dot stands for attentional states.

and "distraction" states. We split the 30 subjects into six folds; each fold contains five subjects. To conduct the leave-subjects-out evaluation, we used all the attention instances from 25 subjects for the training and all the instances from the remaining five subjects for testing at each fold evaluation. Each instance belonged to a specific state (attention or distraction). We reported the average and standard deviations of the recognition accuracy in percentage. Note that we only focused on the recognition task of "recalled" and "reported" distractions. Thus, although "false-negative" errors of the self-reports (e.g., participants forgot to or ignored reporting the distraction) exist, they will not be included in this analysis.

Attentional state recognition. We provide six machine learning-based methods for attentional state recognition, using attention regulator behaviors as cues. We first encoded the distribution of attention regulator behaviors that happened within a given attention span as feature vectors with dimensions of $1 \times N$. N is the number of attention regulator behaviors, as six in practice. Since we used 30 fps for the annotation, which is redundant to count the attention regulator behaviors, we downsample the frame rate from 30 to 8. The resulting feature vectors were fed into the classifiers to predict the final binary attentional states (i.e., attention or distraction). We experimented with different classical machine learning classifiers combined with feature embedding: Bayesian network [13], Multi-layer Perceptron with Relu non-linearity (MLP) [57], k-nearest neighbors (kNN) [32], and Adaptive Boosting (AdaBoost) [59]. As can be seen from Table 4, the MLP classifier has achieved the best performance (69.55%, 89.41%, and 87.57%) in the attention span settings of 4s, 8s, and 16s while the SVM classifier has shown the best performance over the attention span of 2s. We can also observe that a shorter attention span of an instance has brought a significant performance drop (87.57% to 57.84% from 16s to 2s) in the recognition. We assume it is because the shorter attention span implementation does not provide enough information on the attention regulator behaviors to build up the probabilistic distribution model for further inferences.

Visualization of the features for attentional state recognition. In this section, we visualized the feature embeddings constructed from the attention regulator behaviors, using the t-SNE technique

Table 4: Attention regulator behavior-based attention recognition results from various classifiers. The attention span is the instance duration before and after the distraction self-reports. We show the average and standard deviations over six leave-five-subject-out runs.

Methods	Attentional state recognition (%)			
	Attention span (2s)	Attention span (4s)	Attention span (8s)	Attention span (16s)
Random guess	0.50	0.50	0.50	0.50
kNN [32]	51.69 ± 5.62	61.86 ± 11.10	88.91 ± 7.98	80.02 ± 15.67
SVM [20]	**58.09 ± 4.95**	68.83 ± 7.67	89.31 ± 6.92	86.98 ± 7.43
AdaBoost [59]	57.84 ± 5.48	69.14 ± 7.51	88.12 ± 6.92	85.642 ± 6.83
MLP [57]	57.84 ± 5.48	**69.55 ± 7.83**	**89.41 ± 6.91**	**87.57 ± 7.46**

[65]. As shown in Figure 3, features from a short attention span are not discriminative enough, while features from a longer attention span show much larger margins.

5 DISCUSSION AND LIMITATIONS

5.1 Discussion

Distraction self-reports vs. attention regulator behaviors. Self-reported distractions during the e-reading practices can be regarded as ground truths of attentional states of participants to some extent, as they are the direct reflection of internal activities provided by participants. However, there are three major limitations of the distraction self-reports. 1) Self-reports are based on a dual-task condition. The participants might be distracted when keeping the reporting task in their minds, which could affect their attention level. 2) self-reported metrics are not always reliable as participants often forget to record their distractions. Thus, false-negative errors are evident in some case, which could be a severe issue when evaluating the performances of online detection algorithms. 3) Lastly, the self-report is subjective, making it difficult for machines to learn the patterns. In contrast, attention regulators-based attentional state analysis has advantages as follows. 1) Those patterns are concrete and rather easy to observe in the images so that machines can easily learn. 2) Significant correlations found between distraction reports and attention regulator behaviors indicate that observable attention regulator behaviors as a good predictor of attention.

Further implementation in e-learning. Since our work aimed for a real-life application based on a webcam, we believe that the work can be extended to other reading-based e-learning scenarios with an investigation of attention regulator behaviors in the specific learning activity. By combining various feedback types with diverse instructional designs, platforms, and modalities from different feedback agents, more timely feedback provision can be achieved for learners and instructors.

Defining attention span. We defined the attention span by taking the duration before and after the distraction reports (e.g., 2s, 4s, 8s, and 16s). We found that the definition of the attention span can affect the performance of attention recognition by a large amount, as a longer period will contain more behavioral patterns for the recognition. Existing methods [43, 58, 68] mainly worked short-term or even frame-level attention recognition, while our findings can inspire the upcoming research to work on the direction of attention span by showing potential for holistic attention recognition in instances with a longer attention span.

Rich cues for attention analysis. In this work, we only presented some preliminary baselines using attention regulator behaviors and self-reports as cues and ground truths. However, rich cues provided in WEDAR, such as knowledge gains and reaction time, can offer more opportunities for a more holistic and long-term attention analysis.

5.2 Limitations

Differentiating spontaneous behavior vs. voluntary behavior. In this work, we focused on finding regulatory behavior that helps learners sustain their attention. We primarily focused on voluntary or semi-voluntary behaviors from learners with consciousness. However, it was often challenging to differentiate voluntary behaviors from spontaneous behaviors through human observation, which might have affected our labeling and prediction results.

Lack of categorical frameworks for attention regulator behaviors in e-learning. We strived to classify learner behaviors based on existing theoretical and empirical works. Though our work is a categorical expansion of [27], we still miss the dedicated framework that could be applied in the exploration of attention regulator behaviors in e-reading.

6 CONCLUSION AND FUTURE WORK

In this work, we applied the categorical framework of [27] to an e-reading scenario and identified attention regulator behaviors, which was the first attempt. We collected a novel dataset from 30 higher education learners containing various cognitive, emotional, and behavioral cues. We annotated 931,340 frames of video data second-to-second into six categories. We used various classical models to recognize attention regulator behaviors as a baseline with the highest accuracy of 75.70% (subject-dependent) and 68.43% (subject-independent) with CNN-RNN. Attentional state recognition has been further conducted by leveraging the attention regulator behaviors with a promising performance of 89.41% accuracy with a leave-five-subject-out protocol. Our webcam-based dataset and framework for the attention analysis make it feasible to comply with primary computing devices without sophisticated sensor implementation, allowing real-world implementation. We hope our work contributes to the field by providing insights into attention regulator behaviors in e-reading. The future research includes the system extension with the feedback implementation, which will function as an interactive feedback loop for attentive e-reading.

ACKNOWLEDGMENTS

This research was supported by Leiden-Delft-Erasmus Center for Education and Learning, Academy of Finland for Academy Professor project EmotionAI (grants 336116, 345122), and project MiGA (grant 316765).

REFERENCES

[1] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. 2021. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Applied Sciences* 11, 2 (2021), 796.

[2] Damla Arifoglu and Abdelhamid Bouchachia. 2017. Activity recognition and abnormal behaviour detection with recurrent neural networks. *Procedia Computer Science* 110 (2017), 86–93.

[3] Hasnan Baber. 2021. Modelling the acceptance of e-learning during the pandemic of COVID-19-A study of South Korea. *The International Journal of Management Education* 19, 2 (2021), 100503.

[4] Giuseppe Barbato, Vittoria De Padova, Antonella Raffaella Paolillo, Laura Arpaia, Eleonora Russo, and Gianluca Ficca. 2007. Increased spontaneous eye blink rate following prolonged wakefulness. *Physiology & behavior* 90, 1 (2007), 151–154.

[5] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. Context in emotion perception. *Current Directions in Psychological Science* 20, 5 (2011), 286–290.

[6] Timothy Bell. 2001. Extensive reading: Speed and comprehension. *The reading matrix* 1, 1 (2001).

[7] Rebecca Boehme and Håkan Olausson. 2022. Differentiating self-touch from social touch. *Current Opinion in Behavioral Sciences* 43 (2022), 27–33.

[8] Donald Eric Broadbent. 1957. A mechanical model for human attention and immediate memory. *Psychological review* 64, 3 (1957), 205.

[9] Justine Cassell, Y Nakano, T Bickmore, C Sidner, and Charles Rich. 2001. Annotating and generating posture from discourse structure in embodied conversational agents. In *Workshop on representing, annotating, and evaluating non-verbal and verbal communicative acts to achieve contextual embodied agents*.

[10] Ginevra Castellano, Santiago D Villalba, and Antonio Camurri. 2007. Recognising human emotions from body movement and gesture dynamics. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 71–82.

[11] Chih-Ming Chen. 2009. Personalized E-learning system with self-regulated learning assisted mechanisms for promoting learning performance. *Expert Systems with Applications* 36, 5 (2009), 8816–8829.

[12] Chih-Ming Chen and Sheng-Hui Huang. 2014. Web-based reading annotation system with an attention-based self-regulated learning mechanism for promoting reading performance. *British Journal of Educational Technology* 45, 5 (2014), 959–980.

[13] Haoyu Chen, Xin Liu, Xiaobai Li, Henglin Shi, and Guoying Zhao. 2019. Analyze Spontaneous Gestures for Emotional Stress State Recognition: A Micro-gesture Dataset and Analysis with Deep Learning. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition*. 1–8.

[14] Haoyu Chen, Esther Tan, Yoon Lee, Sambit Praharaj, Marcus Specht, and Guoying Zhao. 2020. Developing AI into explanatory supporting models: An explanation-visualized deep learning prototype. In *The International Conference of Learning Science (ICLS)*.

[15] Haoyu Chen, Zitong Yu, Xin Liu, Wei Peng, Yoon Lee, and Guoying Zhao. 2020. 2nd place scheme on action recognition track of eccv 2020 vipriors challenges: an efficient optical flow stream guided framework. *arXiv preprint arXiv:2008.03996* (2020).

[16] Youngjun Cho. 2021. 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*. 1–12.

[17] Christina A Chu, Mark Rosenfield, and Joan K Portello. 2014. Blink patterns: reading from a computer screen versus hard copy. *Optometry and Vision Science* 91, 3 (2014), 297–302.

[18] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[19] Gregory JH Colflesh and Andrew RA Conway. 2007. Individual differences in working memory capacity and divided attention in dichotic listening. *Psychonomic bulletin & review* 14, 4 (2007), 699–703.

[20] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[21] Connie De Vos, Els Van der Kooij, and Onno Crasborn. 2009. Mixed signals: Combining linguistic and affective functions of eyebrows in questions in Sign Language of the Netherlands. *Language and speech* 52, 2-3 (2009), 315–339.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[23] Douglas Derryberry. 2002. Attention and voluntary self-control. *Self and identity* 1, 2 (2002), 105–111.

[24] Sidney D'Mello and Art Graesser. 2010. Mining bodily patterns of affective experience during learning. In *Educational data mining 2010*. Citeseer.

[25] Sidney D'Mello, Tanner Jackson, Scotty Craig, Brent Morgan, P Chipman, Holly White, Natalie Person, Barry Kort, R El Kaliouby, Rosalind Picard, et al. 2008. AutoTutor detects and responds to learners affective and cognitive states. In *Workshop on emotional and cognitive issues at the international conference on intelligent tutoring systems*. 306–308.

[26] Mayada R Eesa. 2010. Facial Expressions A study of Eyebrow Movement During Conversation. *Ahl Al-Bait Jurnal* 1, 10 (2010).

[27] Paul Ekman and Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica* 1, 1 (1969), 49–98.

[28] Paul Ekman and Wallace V Friesen. 2003. *Unmasking the face: A guide to recognizing emotions from facial clues*. Vol. 10. Ishk.

[29] Michael Esterman and David Rothlein. 2019. Models of sustained attention. *Current opinion in psychology* 29 (2019), 174–180.

[30] Myrthe Faber, Robert Bixler, and Sidney K D'Mello. 2018. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods* 50, 1 (2018), 134–150.

[31] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[32] Evelyn Fix and Joseph Lawson Hodges. 1989. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique* 57, 3 (1989), 238–247.

[33] Maria L Flecha-Garcia. 2006. Eyebrow raising, discourse structure, and utterance function in face-to-face dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 28.

[34] Katharine H Greenaway, Elise K Kalokerinos, and Lisa A Williams. 2018. Context is everything (in emotion research). *Social and Personality Psychology Compass* 12, 6 (2018), e12393.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[36] David Hoppe, Stefan Helfmann, and Constantin A Rothkopf. 2018. Humans quickly learn to blink strategically in response to environmental task demands. *Proceedings of the National Academy of Sciences* 115, 9 (2018), 2246–2251.

[37] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R Brockmole, and Sidney K D'Mello. 2019. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction* 29, 4 (2019), 821–867.

[38] Sharifah Noor Masidayu Sayed Is, Nor Azlina Ab Aziz, and Siti Zainab Ibrahim. 2022. A comparison of Emotion Recognition System using Electrocardiogram (ECG) and Photoplethysmogram (PPG). *Journal of King Saud University-Computer and Information Sciences* (2022).

[39] Shimpei Ishiyama, Lena V Kaufmann, and Michael Brecht. 2019. Behavioral and cortical correlates of self-suppression, anticipation, and ambivalence in rat tickling. *Current Biology* 29, 19 (2019), 3153–3164.

[40] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[41] Young-Suk Grace Kim, Yaacov Petscher, and Christian Vorstius. 2019. Unpacking eye movements during oral and silent reading and their relations to reading proficiency in beginning readers. *Contemporary Educational Psychology* 58 (2019), 102–120.

[42] Yoon Lee. 2020. FLOWer: Feedback Loop for Group Work Supporter. *The International Learning Analytics and Knowledge Conference (LAK demo session)* (2020).

[43] Jiajia Li, Grace Ngai, Hong Va Leong, and Stephen CF Chan. 2016. Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics. *ACM SIGAPP Applied Computing Review* 16, 3 (2016), 37–49.

[44] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. iMiGUE: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10631–10642.

[45] Charles W McMonnies. 2020. The clinical and experimental significance of blinking behavior. *Journal of Optometry* 13, 2 (2020), 74–80.

[46] Selene Mota and Rosalind W Picard. 2003. Automated posture analysis for detecting learner's interest level. In *2003 Conference on Computer Vision and Pattern Recognition Workshop*, Vol. 5. IEEE, 49–49.

[47] Stephanie Margarete Mueller, Sven Martin, and Martin Grunwald. 2019. Self-touch: contact durations and point of touch of spontaneous facial self-touches differ depending on cognitive and emotional load. *PloS one* 14, 3 (2019), e0213677.

[48] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. 2018. Survey on emotional body gesture recognition. *IEEE transactions on affective computing* 12, 2 (2018), 505–523.

[49] Pablo Oyarzo, David D Preiss, and Diego Cosmelli. 2022. Attentional and metacognitive processes underlying mind wandering episodes during continuous naturalistic reading are associated with specific changes in eye behavior. *Psychophysiology* (2022), e13994.

[50] Preksha Pareek and Ankit Thakkar. 2021. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review* 54, 3 (2021), 2259–2322.

[51] Suzanne M Prior, Kimberley D Fenwick, Katie S Saunders, Rachel Ouellette, Chantell O'Quinn, and Shannon Harvey. 2011. Comprehension after oral and silent reading: Does grade level matter? *Literacy Research and Instruction* 50, 3 (2011), 183–194.

[52] Ning Qian. 1999. On the momentum term in gradient descent learning algorithms. *Neural Networks* 12, 1 (1999), 145–151.

[53] RK Rahul, S Shanthakumar, P Vykunth, and K Sairamnath. 2020. Real-time Attention Span Tracking in Online Education. In *2020 IEEE MIT Undergraduate Research Technology Conference (URTC)*. IEEE, 1–4.

[54] Keith Rayner. 2009. The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology* 62, 8 (2009), 1457–1506.

[55] Abhishek Revadekar, Shreya Oak, Aumkar Gadekar, and Pramod Bide. 2020. Gauging attention of students in an e-learning environment. In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*. IEEE, 1–6.

[56] Ian Roffe. 2002. E-learning: engagement, enhancement and execution. *Quality assurance in education* (2002).

[57] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.

[58] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction*. 305–312.

[59] Robert E Schapire. 2013. Explaining adaboost. In *Empirical inference*. Springer, 37–52.

[60] Robert Schleicher, Niels Galley, Susanne Briest, and Lars Galley. 2008. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics* 51, 7 (2008), 982–1010.

[61] Maryam Shafiei Sarvestani, Mehdi Mohammadi, Jalil Afshin, and Laleh Raeisy. 2019. Students' experiences of e-learning challenges; a phenomenological study. *Interdisciplinary Journal of Virtual Learning in Medical Sciences* 10, 3 (2019), 1–10.

[62] Kalpathy Ramaiyer Subramanian. 2018. Myth and mystery of shrinking attention span. *International Journal of Trend in Research and Development* 5, 1 (2018).

[63] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[64] Daniel Todorovic. 2020. Choosing what to read out loud while studying: The role of agency in production. (2020).

[65] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[66] Andrea Veronese, Mattia Racca, Roel Stephan Pieters, and Ville Kyrki. 2017. Probabilistic Mapping of human Visual attention from head Pose estimation. *Frontiers in Robotics and AI* 4 (2017), 53.

[67] Sonja Walcher, Christof Körner, and Mathias Benedek. 2017. Looking for ideas: Eye behavior during goal-directed internally focused cognition. *Consciousness and cognition* 53 (2017), 165–175.

[68] Liying Wang. 2018. Attention decrease detection based on video analysis in e-learning. In *Transactions on Edutainment XIV*. Springer, 166–179.

[69] Michae Xuelin Huang, Jiajia Li, Grace Ngai, Hong Va Leong, and Andreas Bulling. 2019. Moment-to-moment detection of internal thought from eye vergence behaviour. *arXiv e-prints* (2019), arXiv–1901.

[70] Shan Zhang, Zihan Yan, Shardul Sapkota, Shengdong Zhao, and Wei Tsang Ooi. 2021. Moment-to-Moment Continuous Attention Fluctuation Monitoring through Consumer-Grade EEG Device. *Sensors* 21, 10 (2021), 3419.