Tracking Sustained Attention with Webcam-Based Eye Gaze and Blink Pattern Tracking

Sven van der Voort
Supervisors: MSc. Yoon Lee and Prof. Dr. Marcus Specht
Delft University of Technology

June 27, 2021

Abstract

Sustained attention is a cognitive state where the learners' attention is completely focused on the learning environment and content-related thoughts for a continuous stretch of time. Sustained attention is vital to perform well on learning tasks, such as reading. Learning analytics platforms that detect changes in sustained attention can prevent ineffective learning by providing direct feedback to the learner. Prior research has found that eye gaze and blink patterns can be good indicators of cognitive state. In this paper we investigate the following main research question: "How can webcam-based eye gaze and blink pattern tracking indicate changes in learners' sustained attention in the remote learning context?". While other studies rely on expensive eye trackers to perform detection, this research explores the use of widely used laptop webcams for detecting changes in sustained attention. We collected webcam data through a small case study involving several different reading tasks. A machine learning classification model was trained on the collected webcam data. The resulting detection model performs well on validation data with a F1-score of 0.889. The model does not perform well on testing data however, showing that it is not usable in practice. We give several possible explanations for this behavior, most of them originating from an overfitted model due to the small size of the user study. Our findings indicate that future work should focus on different experimental settings and larger user studies.

1 Introduction

In the current digital age, education and learning are becoming more computer-mediated every day. Remote learning methods like blended learning and Massive Open Online Courses (MOOCs) are becoming more popular, but require students to spend more time studying from a digital device [Terras and Ramsay, 2015]. Furthermore, the current COVID-19 pandemic has forced many students to study solely from home in a remote learning context.

Although remote learning paradigms like blended learning can contribute to increased learning performance [Baragash and Al-Samarraie, 2018], it introduces several new problems for students like maintaining attention, engagement and self-regulation [Pedrotti and Nistor, 2019]. Students could benefit from attention monitoring and feedback on their concentration levels. Hence we propose a detection system which monitors the learners' sustained attention. This can be used to give learners insights into their own attention and learning

performance. In the future the detection system could also be part of a bigger system that provides direct feedback on the learners' attention.

One of the challenges to construct such a detection system is to find indicators of a change in the learners' sustained attention. The Learning Analytics community finds eye gaze and blink tracking to be a promising candidate for such an indicator [Mills et al., 2016, Bixler and D'Mello, 2016, Huang et al., 2019]. It is a long standing hypothesis that there is a link between external information and eye movements during reading, the so-called eye-mind link [Rayner, 1998]. The eye-mind link could provide indicators of the cognitive state of the learner. This concept has been explored in several recent studies, which show promising results. Mills et al. studied eye gaze and blink patterns while participants watched a narrative film to detect mind wandering [Mills et al., 2016]. The study shows that a mind wandering detection model that uses local features of the content (e.g. salient areas of the screen) performs better than a model that uses global features or a combination global and local features. They used a Tobii TX 300 eye tracker and later processed the collected data using 12 different machine learning algorithms. Bixler and D'Mello provide a similar study, but with the focus on mind wandering detection during reading. Using data from a Tobii T60 eye tracker a detection model was built that provides a detection accuracy of 72% (expected accuracy by chance 60%) [Bixler and D'Mello, 2016]. Contrary to the previous study based on film viewing, this study based on reading finds global eye gaze features to contribute more towards detection accuracy. In another study by Huang et al. the relation between eye vergence (eyes moving towards or from each other) and internal thought was investigated. It was shown that eye vergence can be a strong indicator of internal thought during lecture viewing [Huang et al., 2019]. This study used a Tobii EyeX remote eye tracker to collect data and Random Forest as a classification algorithm. Another part of this study reviewed an attention detection and feedback device during lecture viewing. The setup is similar to the detection and feedback system we proposed earlier. Participants in the study reported their experience with the device as positive, although the detection accuracy needed improvement. Current detection and feedback systems have limited usability due their dependency on expensive eye trackers. There is a need for improving the deployability and usability of these systems.

To close this gap in deployability, this study aimed to contribute to the field of Learning Analytics by answering the following main research question: "How can webcam-based eye gaze and blink pattern tracking indicate changes in learners' sustained attention in the remote learning context?". Where previous research mainly used expensive external eye trackers for tracking eye gaze and blinks, the novelty of this study is that the eye tracking was done using a commodity laptop webcam. Although this posed challenges with for example tracking accuracy, using a commodity webcam allows a future detection and feedback system to be widely deployed to learners in the remote learning context.

Recent literature suggests that Learning Analytics systems can benefit from sensor multimodality to improve detection accuracy [Ochoa and Worsley, 2016, Noroozi et al., 2020, Mitri et al., 2017]. Supplementary to the eye gaze and blink model we collaborated with a research team of four BSc students to find indicators of changes in sustained attention in a multimodal sensor setting. We attempted to answer the following sub-question: "Can multimodal sensor data improve the performance of the machine learning model for detecting changes in the learners' sustained attention in the remote learning context?".

This paper is structured according to three sub-questions and one technical sub-question:

- RQ1. What is sustained attention and how can it be recognized?

- RQ2. Can a machine learning model using webcam-based gaze and blink tracking detect changes in the learners' sustained attention in the remote learning context?
- RQ3. Can multimodal sensor data improve the performance of the machine learning model for detecting changes in the learners' sustained attention in the remote learning context?
- Technical sub-question: How can gaze and blink patterns be tracked with a commodity webcam?

Based on the main research questions and the sub-questions, two hypotheses were constructed:

- H1. A machine learning model can be trained using webcam-based gaze and blink tracking to detect changes in the learners' sustained attention in the remote learning context.
- H2. The performance of the machine learning model can be improved by training it on features from additional multimodal sensors.

The remainder of this paper is structured as follows. Section 2 Methodology describes some background on sustained attention, the experiments designed to collect data and the post-processing and training on the collected data. Section 3 Experimental Setup and Results covers the details of the experimental environment and the results obtained from the collected data. Section 4 Responsible Research discusses the reproducibility of the research and the ethical concerns related to the study. Section 5 Discussion provides a comprehensive discussion of the results. Finally, section 6 Conclusions and Future Work summarizes our findings and makes recommendations for future research.

2 Methodology

Sustained attention, also called overt attention, is a cognitive state where the learners' attention is focused on the learning environment and consists of content-related thoughts for a continuous stretch of time [D'Mello, 2016]. Any other attentive state, such as mind wandering, internal thought, distraction, etc. are not sustained attention and will be considered as inattentive for this study. Please note that cognitive states like internal thought still allow the learner to think about the learning material, but this is outside the scope of this research.

Sustained attention is important when reading/studying long texts to be able to process and store the information in long term memory [Stern and Shalev, 2013]. Since reading long texts still makes up a majority of learning tasks, especially in higher education, being able to maintain sustained attention for longer periods of time is important. To limit the scope of the research, we focused on reading tasks as a learning task in particular.

Dataset collection

The goal of this research is to build a unimodal model that can detect changes in the learners' sustained attention in the remote learning context, specifically using gaze and blink tracking by a commodity laptop webcam. Additionally, we will attempt to improve detection by using a multimodal approach in collaboration with other members of the research team, who built

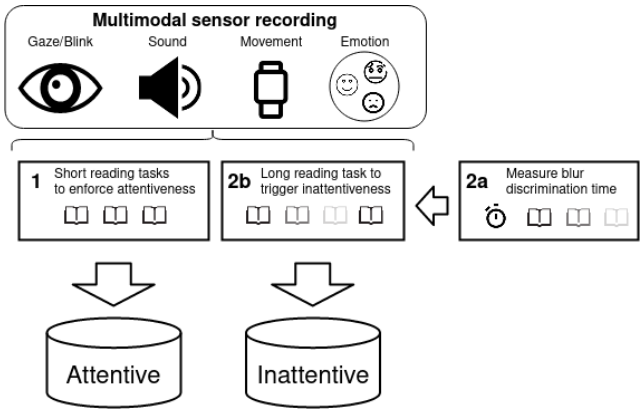


Figure 1: A schematic overview of the three experiments. During experiment 1 we collect "attentive" sensor data using short reading tasks. During experiment 2b we collect similar sensor data during a long reading task, while annotating the participant's attention using the gradual blurring paradigm and self-reports. Experiment 2a allows us to calibrate the blur discrimination time of the gradual blurring annotation paradigm.

their own attention/distraction detection model using facial emotions tracking, smartwatch and smartphone sensors, a temperature sensor and microphones. To train the unimodal and multimodal detection models and to test hypothesis H1 and H2, multimodal training and validation sensor data is needed. For both models supervised classification machine learning was used to build the models. These classification algorithms require data labeled into classes. A small user study was carried out to record and label this data. During the small user study, participants performed three experiments. Below a detailed description can be found of all experiments. Figure 1 provides a schematic overview of the three experiments.

The first experiment (experiment 1) focused on collecting sensor data related to an attentive state of the learner. Participants read a number of short texts from a laptop screen with the explicit instruction to stay attentive while reading. All sensors recorded timestamped data while the participants read the short texts. After finishing all texts, the recording was stopped and participants were asked if they were in fact attentive. If not, the collected data was discarded. This resulted in a collection of sensor data recordings that were all labeled as "attentive".

Experiments 2a and 2b focus on collecting sensor data from an inattentive learner. Research suggests that learners shift their attention every 10 minutes on average [Wankat, 2002]. To achieve multiple such shifts of attention, experiment 2b consisted of a long reading task of about 1 hour. Participants were asked to read the full text just like they would during a normal learning task. During the complete length of the reading task, sensor data was recorded similarly to experiment 1. Before and after reading the long text, participants filled in a questionnaire (see GitHub repository [1]) to measure their knowledge of the contents of the text. The results from this questionnaire were used in another study by a fellow research team member, but they were not used in this study. Also relating to another study, several sounds were played on a speaker during the long reading task, to measure how distracting the sounds would be.

To find out when the attention of the participant shifted during experiment 2b, a ground truth annotation paradigm introduced by Huang et al. was used [Huang et al., 2019].

[1] `https://github.com/MultimodalLearningAnalytics/rp-group-19-common`

The gradual blurring paradigm exploits human blur perception by gradually blurring the screen at random intervals, asking the participant to press a button to deblur the screen when they notice the blurring. Longer deblur times T_{deblur} indicate mind wandering, a common form of inattentiveness. Studies by Huang et al. have shown that this method is effective for obtaining ground truth data on attention. Benefits of this approach are that the attention assessment contains precise timing data and is less prone to subjectivity. This is in contrast to other existing method such as post-hoc self-reports, probe-caught or self-caught paradigms [Smallwood and Schooler, 2006]. A limitation of this approach as mentioned by Huang et al. is that short deblur times are not a direct indication of attentiveness. That is why we collect sensor data from attentive learners during experiment 1. To facilitate studies by fellow research team members, participants were also asked to press a button if they felt distracted during the long reading task. Timestamps of both long deblur times as well as distraction button presses were recorded and time windows before these events were labeled as "inattentive" and "distracted" respectively.

Calibration of the blur discrimination time for experiment 2b was performed using experiment 2a. Although the original study mentions that all deblur times bigger than the blur discrimination time T_d of 1.5s can be labeled as mind wandering [Huang et al., 2019], we needed to verify these findings on our setup. Reason for this is that we can assume that slight differences in learning task, blurring animation and environment affect the blur discrimination time. The design of experiment 2a is similar to Huang et al.'s Study II, where participants watched short lecture videos while the screen started blurring at short random intervals. Participants were asked to stay attentive and deblur the screen by pressing a button as soon as they noticed the blur. The deblur times T_{deblur} were recorded to create an estimate of the time it takes an attentive reader to deblur the screen, the blur discrimination time T_d. We substituted watching short lecture videos by reading short texts, to imitate the setting of experiment 2b.

The participants in the small user study were three members of the research team (all male students, aged between 20 and 24). The ongoing COVID-19 pandamic prevented us from inviting a bigger group of participants to the study. Due to the small and homogeneous group of participants, this study should be considered a case study and will not be generalizable to larger groups. However, case studies can still be relevant, especially in social sciences [Feagin et al., 1991], and single user testing is frequently used for first time data validation.

Feature selection

From the recorded sensor data specific eye gaze and blink features were calculated for time windows labeled as attentive or inattentive. Previous research on attention detection from eye gaze and blink patterns gives an indication of features that could be useful for detection. Features like pupil diameter [Bixler and D'Mello, 2016, Mills et al., 2016], pupillary distance and eye vergence [Huang et al., 2019] are good indicators of changes in attention. Unfortunately it is impossible to detect fluctuations in these features from a low resolution commodity webcam as used in this study. The relatively low accuracy of state-of-the-art webcam gaze tracking [Wood et al., 2015] also means that local features (such as the location of text on the screen) cannot be used. Therefore only global features which do not require high accuracy were selected.

The selected features can be seen in Table 1. The features encode two different types of eye movements: saccades and blinks. Saccades are rapid eye movements to move the focus

Feature	Description	Dimensions
Horizontal saccades [Mills et al., 2016, Bixler and D'Mello, 2016]	Rapid eye movements over x-axis	Count, **distance**, **time between**
Vertical saccades [Mills et al., 2016, Bixler and D'Mello, 2016]	Rapid eye movements over y-axis	Count, **distance**, **time between**
Overall saccades [Mills et al., 2016, Bixler and D'Mello, 2016]	Rapid eye movements (euclidean distance x-axis and y-axis)	Count, **distance**, **time between**
Blinks [Mills et al., 2016, Bixler and D'Mello, 2016]	Facial Action Unit 45 [Baltrušaitis et al., 2015]: blinking	Count, **duration**, **time between**
Blink ratio	Ratio of eyes closed to eyes open	Single dimension
Distance covered by gaze	Total distance covered by gaze in time window (euclidean distance x-axis and y-axis)	Single dimension

Table 1: Eye gaze and blink features calculated from the collected data (54 total). **Bold** dimensions use statistical measures to describe the feature: mean, standard deviation, median, min, max and range.

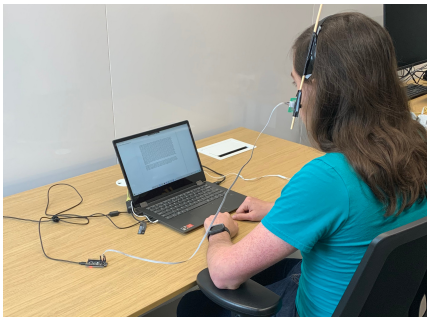
of the eye from one location to the next. Blinks are short periods of time that the eyes are closed. Saccades are split up in three different features: horizontal saccades, vertical saccades and overall saccades, refering to the direction of the saccade. Overall saccades are movements in any direction that cover a large euclidean distance. For all saccades we measured their covered distance, time between saccades and total saccade count. Likewise, for blinks we measured duration, time between blinks and total blink count. We analyzed several statistical measures of these features, such as mean, standard deviation, median, min, max and range. Finally, we measured the ratio between eyes open to eyes closed as the blink ratio and the total euclidean distance covered by the gaze in the window as the distance covered by gaze. This gives a total of 54 features.

Classification models

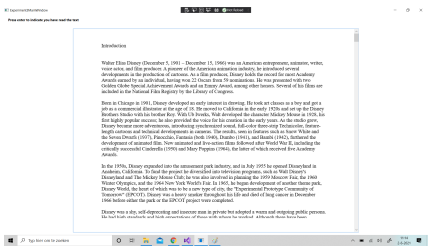
On the collected data, including their classes/labels, a range of different classification models were trained, varying in window size (10s, 20s, 30s), included classes (attentive + only distraction button, only long deblur time or both) and classification algorithm. What window size to use is an ongoing debate in literature, where some argue that shorter windows are better [Huang et al., 2019], while others argue the opposite [Mills et al., 2016]. By varying the window size, we cover all cases. The Weka Machine Learning Software [Hall et al., 2009] provided all implementations of the classification algorithms. Three supervised classification algorithms were selected based on success in previous studies: Naive Bayes (used by [Bixler and D'Mello, 2016]), Decision Table (used by [Mills et al., 2016]) and Random Forest (used by [Huang et al., 2019]). The performance of all trained models was evaluated according to their F1-score after 10-fold cross-validation. The F1-score provides a good measure of the performance on validation data of the model by taking into account false positives and false negatives.

3 Experimental Setup and Results

This section describes the details of the experimental setup of the data collection experiments and an analysis of the obtained results. As described in the section 2 Methodology, a small user study was performed to collect training, validation and testing data. The participants of the small user study were three members of the research team themselves. All participants were male students and aged between 20 and 24. The three experiments all consisted of one or more reading tasks. For experiment 1 12 short technology news articles were selected (number of words mean: 152, SD: 27.2, min: 90, max: 192). For experiment 2a 12 different short technology new articles were selected (number of words mean: 206, SD: 29.3, min: 144, max: 271). For experiment 2b a single long text on the topic of Walt Disney and the Disney theme parks was selected (number of words: 10189). The texts were new to all participants. All texts can be found in the research GitHub repository [2].



(a) Experimental setup participant



(b) Screenshot experiment software, text not blurred

(c) Screenshot experiment software, text blurred

Figure 2: Experimental setup: (a) Participant with all sensors attached. A temperature sensor is mounted to the head, a smartwatch attached to the wrist and a smartphone in the pocket. (b) Screenshot of the special software created to record the sensor data and blur/deblur the screen. the interface was designed to mimic a PDF reader. (c) Same screenshot, but with the text blurred.

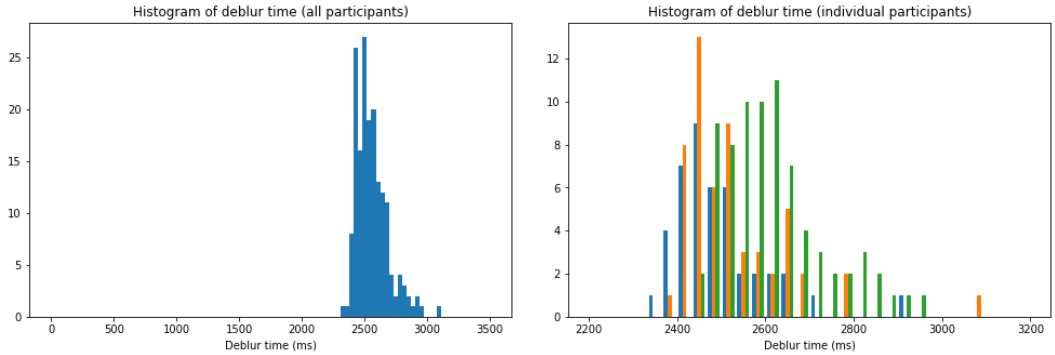
The user study consisted of three experiments. All participants participated in all experiments on the same day. All experiments were conducted in the same type of environment: a quiet room with the participant seated at a desk in front of a Lenovo Yoga 530-14ARR laptop with a 14.0 inch 1920x1080 screen and a HD 720p webcam. Special software was created to facilitate the experiments. The software displayed text files from a selected folder in random order to the participant. The text was shown with font size 18 in the center of the screen with margins similar to a PDF reader. Figure 2b shows a screenshot of this special software without text blur and Figure 2c shows a screenshot with text blur (see GitHub for source code [2]). The software enabled recording timestamped sensor data for later analysis though Microsoft Platform for Situated Intelligence (PSI) [Bohus et al., 2021]. It also allowed blurring and deblurring the displayed text at pseudo random intervals as required by experiment 2a and 2b. Figure 2a shows a picture of the experiment setup with all sensors attached. A temperature sensor was mounted to the head, a smartwatch attached to the wrist and a smartphone in pocket. It must be noted that the experiments were not performed in this exact room, but in similar room to the one in the picture. During experiment

[2] `https://github.com/MultimodalLearningAnalytics/rp-group-19-common`

2b, several different sounds were played on a speaker in the room in order to distract the participant. This was done the context of another study by a fellow research team member.

Determining blur discrimination time

Calibration of the blur discrimination time of the gradual blurring paradigm was performed through experiment 2a. During experiment 2a participants were asked to read 12 short texts. The software started a blurring animation on the text at random intervals of 2-5s. Participants were instructed to press the space bar as soon as they noticed the screen blurring. After reading one of the short texts, participants pressed enter to take a short break and pressed enter again to display the next text. Participants were explicitly instructed to stay attentive while reading, which was possible due to the short nature of the texts.



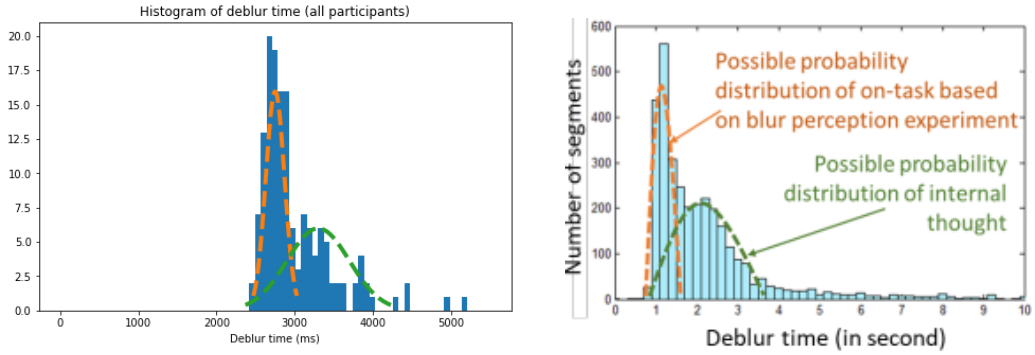
(a) All participants

(b) Individual participants

Figure 3: Experiment 2a deblur times for (a) all participants and (b) individual participants indicated by different colors. The clear peak around the median of 2530 ms shows that we can confidently base blur discrimination time T_d on these measurements.

In total 174 deblur times were collected with a median of 2530 ms, standard deviation of 126 ms and a 98th-percentile of 2902 ms. Figure 3b shows a histogram of the individual deblur times of the particpants. Although the distributions do not fully overlap, we considered them to be sufficiently similar for the purpose of the experiment, namely determining the blur discrimination time for attentive readers. Figure 3a shows a histogram of the deblur times of all participants. The histogram shows a clear peak around the median of 2530 ms, which indicated that we could confidently base our discrimination time T_d on this data. We rounded the 98th-percentile value of 2902 ms to 2900 ms to be used as the blur discrimination time T_d. As expected, this result was different from what was found in the paper that introduced the gradual blurring paradigm. They found a blur discrimination time T_d of 1.5s (including 0.3s human reaction time, analogous to our result) [Huang et al., 2019]. A number of factors that could explain this difference are the use of a different blurring animation/speed and text format, a different screen and a different learning task (reading instead of lecture watching). The difference shows that it is useful to do a calibration of the blur discrimination time on every different setup, also in future work.

During the long reading task of experiment 2b a total of 150 deblur times were collected (median: 2857 ms, SD: 470.2 ms). Figure 4a shows a histogram of all deblur times during experiment 2b. The left-most orange peak around 2700 ms corresponds to the distribution



(a) Deblur times during reading of experiment 2b

(b) Deblur times during lecture viewing as reported by Huang et al. [Huang et al., 2019]

Figure 4: Comparison of deblur time distributions overlaid with possible probability distribution for attentive (orange) and inattentive (green) (a) deblur times from experiment 2b (b) deblur times as reported by Huang et al.

that we found during experiment 2a for attentive readers (see Figure 3a). A smaller green peak around 3300 ms likely corresponds to deblur times of an inattentive reader. This confirms a upper bound on the blur discrimination time of around 2900 ms as was found with experiment 2a. The results agree with the findings of Huang et al. shown in Figure 4b. These results validate and provide confidence in our implementation of the gradual blurring annotation paradigm.

Extracting training and validation data

From the recorded sensor data of the short reading tasks (experiment 1) and the long reading task (experiment 2b) labeled training and validation data was extracted to feed to the classification algorithm. The raw webcam data required post-processing to extract gaze and blink data. The webcam recorded with a resolution of 1280x720 pixels at 14.7 frames per second (fps) on average. The post-processing was done using the OpenFace facial behavior toolkit [Baltrušaitis et al., 2018]. A custom C++ OpenFace bridge in combination with a custom PSI script enabled fast post-processing of the webcam video data into gaze and blink data, stored in a PSI store.

To extract gaze and blink features relating to attentiveness or inattentiveness, time windows of sensor data of three different lengths (10s, 20s and 30s) were extracted from the gaze and blink data stream. The data in these windows were aggregated into the features that are mentioned in section 2 Methodology (see Table 1). Saccade detection was done by comparing the gaze angle value with a gaze angle value 500ms later. If the difference was bigger than 0.05 for x-axis (gaze angle speed $>$ 0.10 per second) or 0.10 for y-axis (gaze angle speed $>$ 0.20 per second), a saccade was registered. Figure 5a shows a saccade in a graph of the gaze angle over time. The start of a blink was detected when Facial Action Unit 45 "Blink" (AU45) [Baltrušaitis et al., 2015] was active for more than 300ms. The blink ended when AU45 was not active for more than 300ms. Figure 5b shows a blink in the graph of AU45 over time. For more information on the post-processing see the source code and extracted datasets on GitHub [3].

[3] https://github.com/MultimodalLearningAnalytics/eye-gaze-blink-tracking

(a) Gaze angle x-axis over time showing a saccade

(b) AU45 indication over time, showing a blink

Figure 5: Graphs showing features: (a) a saccade and (b) a blink

From manual inspection of the webcam video data and the extracted blink data it could be seen that AU45 was not only active when a person blinked. AU45 was also active when the eyes were slightly closed or the participant faced the webcam at a certain angle. However, we still considered it worthwhile to keep the blink data, since slightly closed eyes or a certain head pose could just as well be indicators of changes in sustained attention.

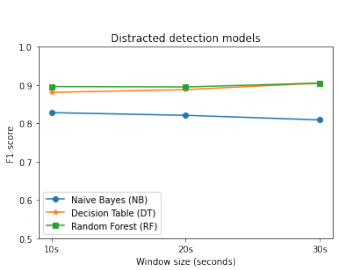
Model training and validation

Multiple classification models were created through training on the labeled extracted and aggregated eye gaze and blink features. From experiment 1 consecutive "attentive" time windows of 10s, 20s and 30s were extracted: in total 163 10s windows, 72 20s windows and 42 30s windows of sensor data. (The first recording of experiment 1 of participant 1 was discarded because the recording was corrupt, leaving 11 recordings of participant 1 during experiment 1.) From experiment 2b in total 130 not attentive data points were extracted from time windows either before participants pressed the distraction button (labeled as "distracted", 67 data points) or before participants had a longer deblur time (labeled as "inattentive", 63 data points). For each combination of window size (10s, 20s, 30s) and type of inattentiveness (only "distracted", only "inattentive" or both) three different classification algorithms were run rendering a total of 27 model training configurations. To address the class imbalance present in some datasets, random downsampling was applied to the majority class until the classes were equally balanced. All models were trained using the default hyperparameter settings of Weka. Cross-validation with 10 folds was used to do model validation. The F1-scores of all models can be seen in Table 2. Graphs of the F1-scores for different time windows can be seen in Figure 6.

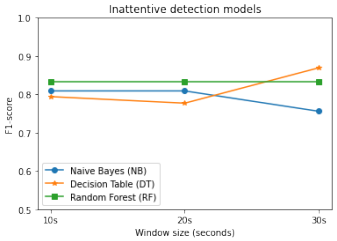
Window (seconds)	10s			20s			30s		
Classification algorithm	NB	DT	RF	NB	DT	RF	NB	DT	RF
Distracted (self-report)	0.828	0.881	0.896	0.821	0.888	0.895	0.809	0.905	0.905
Inattentive (long deblur time)	0.809	0.794	0.833	0.809	0.777	0.833	0.756	0.869	0.833
Distracted or inattentive	0.815	0.865	0.881	0.833	0.833	0.889	0.737	0.821	0.854

Table 2: F1-scores of all trained models. Every row represents which non-attentive labels were used for training: only distracted (self-report), only inattentive (long deblur time) or both. For every row 9 models were trained with windows of 10s, 20s and 30s and with three different classification algorithms: Naive Bayes (NB), Decision Table (DT) and Random Forest (RF). Best performing models per row are colored green.

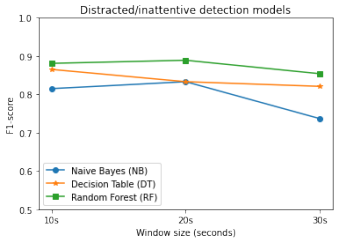
From the results in Table 2 we can see that the best performing model for detecting self-reported distraction is either Decision Table or Random Forest with 30s windows. The best performing model for detecting inattentiveness through long deblur times is Decision Table with 30s windows. The best performing model to detect not paying attention in general



(a) F1-scores distracted models



(b) F1-scores inattentive models



(c) F1-scores distracted or inattentive models

Figure 6: F1-scores for models of different window sizes

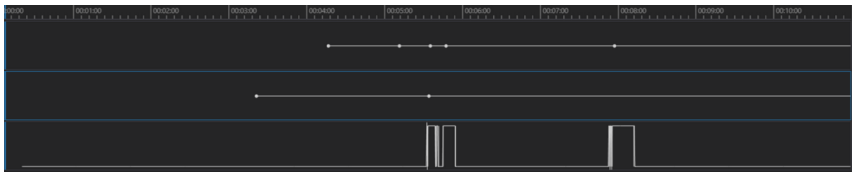
(either distraction or inattentiveness) is Random Forest with 20s windows. Although 20s and 30 windows tend to perform better than 10s windows, the differences are not significant. The same goes for the Random Forest classifier: it performs best is nearly all cases, but the differences are too small to be significant.

Model testing

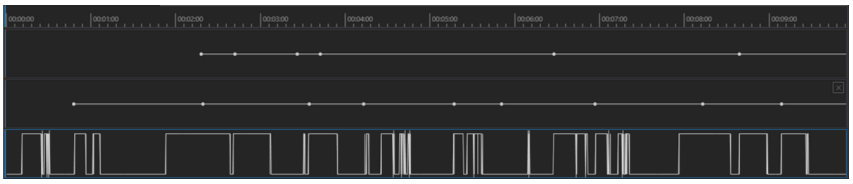
The full recording of the long reading task (experiment 2b) was used as means to test the best model to check whether the model would work in practice. For each frame of the recording the preceding 20 seconds were aggregated into features and labeled by the trained detection model either "attentive" or "inattentive". The results were unanticipated: the model would label almost all frames as "inattentive" (47121 out of 47229). Although the individual frames are not all labeled by a ground truth, common sense tells us that this is incorrect: at least for a large portion of the time, the participants had to be attentive to be able to read the text. Repeating the same actions on recordings from short reading tasks (experiment 1) resulted in 100% of the frames being labeled as "attentive" (which is correct, considering the participants were asked to stay attentive). This is a strong indication that the trained model does not perform well in practice.

Alternative methodology

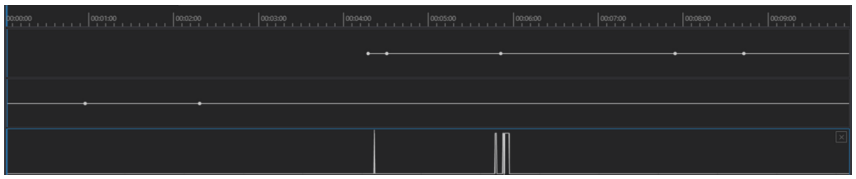
To explore alternatives to our methodology, another detection model was trained on a dataset constructed through a completely different methodology. To train this model a dataset was constructed with only time windows from the long reading task (experiment 2b). Time windows where no distraction button was pushed and where no long deblur times were detected, were labeled as "attentive". Time windows preceding a distraction button push or a long deblur time, were labeled "inattentive", similar to the original methodology. All data was split up in three datasets: the first 10 minutes of the recording belonged to the test dataset, the second 10 minutes to the validation dataset and the remaining 30-34 minutes belonged to the training dataset. A classification model was trained using the Random Forest classifier, 20s windows and no downsampling on 186 "attentive" and 74 "inattentive" data points from the training dataset. The model performed well on the validation dataset with a F1-score of 0.723. Analysis of predictions on the validation dataset shows that the model performs relatively well, at least better than the original model. It detected 9 out of 16 distracted points and 9 out of 12 long deblur events correctly. Inspection of the full 10 minute testing dataset in Figure 7 shows that the model's false positive rate highly differs per participant, indicating the limitations of the usability of this model.



(a) Testing dataset participant 1



(b) Testing dataset participant 2



(c) Testing dataset participant 3

Figure 7: Predictions by alternative model on its testing dataset. The two upper lines represent distraction button presses and long deblur times with dots respectively. The bottom line represents the model prediction of inattentiveness: high is "inattentive", low is "attentive". We can see that the false positive rate highly differs between participants.

4 Responsible Research

In the spirit of Open Science, all source code and derived datasets used to conduct this research are published under an open source license on GitHub [4]. The open source repository contains scripts to run the described experiments and to collect the necessary data. The PSI stores and datasets containing gaze and blink data are also available. The original video data from the experiments is not published, because it is sensitive data of the experiment participants (members of the research team). Upon reasonable request with the researcher this data can be shared.

Due to the small number and low diversity of the participants in the experiment (n=3, all male, all students, aged 20-24), it is likely that the published models and corresponding results are not generalizable to a different group of people or other experiment contexts. To create a more generalizable and reproducible result, future studies should aim for a higher number and a more diverse set of participants to avoid model overfitting.

It must be mentioned that although this research was conducted with the best of intentions, the described technology can also be used for in an unethical manner. Care must be taken that the privacy sensitive data collected for detection of changes in sustained attention is sufficiently protected. This can be achieved by designing future detection systems according to "Privacy by Design", which is required by Art. 25 of the GDPR [Council of European Union, 2016]. In case of the detection system collected data can immediately be discarded after detection is completed. Additionally, making the detection system mandatory and/or

[4] `https://github.com/MultimodalLearningAnalytics`

assessing learners based on their attention can prevent a safe and trusted learning environment that is required for effective learning [Drachsler and Greller, 2016].

5 Discussion

The main goal of this research was to build a model that detects changes in sustained attention through webcam-based eye gaze and blink tracking. The methodology was carefully designed and based on previous research in detecting attention based on eye gaze and blink tracking, consisting of two different experiments to collect data for the classification model to train on. Although the model performs well in the synthetic context of the training and validation datasets, predictions on testing data show that this does not translate into a usable model in practice. These results are likely related to the small size of the dataset, causing the model to overfit on conditions of the experimental context of the short and long reading tasks, rather than detecting attention. Some of those properties might be: lighting conditions, head pose or webcam orientation. Although these are not direct input features of the model, the model features might be proxies for these conditions. Another possibility is that the model actually detects eye gaze and blink patterns that relate to a short or long reading task. Collecting more data from more participants could help alleviate the problem of overfitting. Additionally, a higher resolution webcam or future (more accurate) eye gaze and blink tracking software could provide solutions.

We explored an alternative methodology to create the dataset. This dataset contained only data from the long reading task, assuming the reader to be attentive whenever the distraction button was not pressed and no long deblur times were detected. A model trained on this data performed better than the original model, although it's false positive rate on testing data is still high. The assumption that a reader is attentive whenever the inattentive ground truth is not active, is also disputed in previous research. Huang et al. explicitly state that one cannot assume attention with short deblur times [Huang et al., 2019].

6 Conclusions and Future Work

Being able to maintain sustained attention is an important factor in achieving effective learning, especially in the remote learning context. A robust framework for detecting changes in sustained attention could open up possibilities for direct feedback systems for learners.

The results of this research produced a number of important findings. First we defined sustained attention as a cognitive state focused on the learning environment and consisting of content-related thoughts for a continuous stretch of time (RQ1). Changes in sustained attention can come from transitions to other cognitive states such as mind wandering and distraction. Second we have successfully applied the gradual blurring annotation paradigm that was introduced by Huang et al. It has proven to be a reliable and unobtrusive method to gather ground truth data on inattentiveness during learning analytics experiments. Third we have demonstrated a novel methodology to build a classification model to detect changes in sustained attention using webcam-based eye gaze and blink tracking. The constructed model performed well on validation data with a F1-score of 0.889, but does not perform well in practice due to overfitting on the experimental context (RQ2). Fourth we explored an alternative methodology using only data from a long reading task showing more promising results. However, it uses the ground truth annotation in a debatable manner by assuming

attentiveness in the absence of inattentiveness. With these results we could not confirm nor reject hypothesis H1.

No attempt was made to include the resulting eye gaze and blink tracking model in a multimodal analysis or multimodal detection model, since the model is known not the perform well in practice (RQ3). Therefore we could also not confirm nor reject hypothesis H2.

Although this research does not directly answer the main research question on how to detect changes in sustained attention through webcam-based eye gaze and blink tracking in the remote learning context, a lot can be learned from our findings. We show the importance of a large, diverse dataset and participants group, which is crucial to avoid overfitting on biases and characteristics of individual participants and experimental environments. Additionally, the proposed methodology where data is collected during separate experiments involving short and long reading tasks has some drawbacks, like the risk of model overfitting on experimental context. The alternative methodology where data comes from a single experiment with a long reading task provides some promising early results. We advice future research to focus on this methodology rather than our earlier proposed methodology. Unobtrusively collecting proper attentiveness ground truth is still an open question that needs to be answered by future research. Finally, we encourage future studies to publish their multimodal datasets of learning analytics experiments to advance and accelerate the field of multimodal learning analytics.

Acknowledgments

I would like to thank MSc. Yoon Lee and Prof. Dr. Marcus Specht for their supervision and useful feedback. I would also like to thank my fellow research team members Giuseppe Deininger, Jeffrey Pronk and Jurriaan Den Toonder for their inexhaustible work and attention to design and build the experiments together with me.

References

[Baltrušaitis et al., 2015] Baltrušaitis, T., Mahmoud, M., and Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–6.

[Baltrušaitis et al., 2018] Baltrušaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66.

[Baragash and Al-Samarraie, 2018] Baragash, R. S. and Al-Samarraie, H. (2018). Blended learning: Investigating the influence of engagement in multiple learning delivery modes on students' performance. *Telematics and Informatics*, 35:2082–2098.

[Bixler and D'Mello, 2016] Bixler, R. and D'Mello, S. (2016). Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*, 26:33–68.

[Bohus et al., 2021] Bohus, D., Andrist, S., Feniello, A., Saw, N., Jalobeanu, M., Sweeney, P., Thompson, A. L., and Horvitz, E. (2021). Platform for situated intelligence. Technical Report MSR-TR-2021-2, Microsoft.

[Council of European Union, 2016] Council of European Union (2016). Council regulation (EU) no 679/2016. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504&qid=1622555601886.

[D'Mello, 2016] D'Mello, S. K. (2016). Giving eyesight to the blind: Towards attention-aware aied. *International Journal of Artificial Intelligence in Education*, 26:645–659.

[Drachsler and Greller, 2016] Drachsler, H. and Greller, W. (2016). Privacy and analytics - it's a delicate issue a checklist for trusted learning analytics. volume 25-29-April-2016, pages 89–98. Association for Computing Machinery.

[Feagin et al., 1991] Feagin, J. R., Orum, A. M., and Sjoberg, G. (1991). *A Case for the Case Study*. The University of North Carolina Press.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

[Huang et al., 2019] Huang, M. X., Li, J., Ngai, G., Leong, H. V., and Bulling, A. (2019). Moment-to-moment detection of internal thought from eye vergence behaviour.

[Mills et al., 2016] Mills, C., Bixler, R., Wang, X., and D'mello, S. K. (2016). Automatic gaze-based detection of mind wandering during narrative film comprehension.

[Mitri et al., 2017] Mitri, D. D., Börner, D., Scheffel, M., Ternier, S., Drachsler, H., and Specht, M. (2017). Learning pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data. pages 188–197. Association for Computing Machinery.

[Noroozi et al., 2020] Noroozi, O., Pijeira-Díaz, H. J., Sobocinski, M., Dindar, M., Järvelä, S., and Kirschner, P. A. (2020). Multimodal data indicators for capturing cognitive, motivational, and emotional learning processes: A systematic literature review. *Education and Information Technologies*, 25:5499–5547.

[Ochoa and Worsley, 2016] Ochoa, X. and Worsley, M. (2016). Editorial: Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, 3:213–219.

[Pedrotti and Nistor, 2019] Pedrotti, M. and Nistor, N. (2019). How students fail to self-regulate their online learning experience. In *European Conference on Technology Enhanced Learning*, pages 377–385, Cham. Springer.

[Rayner, 1998] Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124.

[Smallwood and Schooler, 2006] Smallwood, J. and Schooler, J. W. (2006). The restless mind. *Psychological Bulletin*, 132.

[Stern and Shalev, 2013] Stern, P. and Shalev, L. (2013). The role of sustained attention and display medium in reading comprehension among adolescents with adhd and without it. *Research in Developmental Disabilities*, 34:431–439.

[Terras and Ramsay, 2015] Terras, M. M. and Ramsay, J. (2015). Massive open online courses (moocs): Insights and challenges from a psychological perspective. *British Journal of Educational Technology*, 46(3):472–487.

[Wankat, 2002] Wankat, P. (2002). *The Effective, Efficient Professor: Teaching, Scholarship, and Service*.

[Wood et al., 2015] Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., and Bulling, A. (2015). Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.