Automatic Detection of Mind Wandering During Reading Using Gaze and Physiology

Robert Bixler¹, Nathaniel Blanchard¹, Luke Garrison¹, Sidney D’Mello¹,²
¹Department of Computer Science and Engineering
²Department of Psychology
University of Notre Dame, Notre Dame, IN, 46556, USA
{rbixler, nblancha, lgarrison, sdmello}@nd.edu

ABSTRACT
Mind wandering (MW) entails an involuntary shift in attention from task-related thoughts to task-unrelated thoughts, and has been shown to have detrimental effects on performance in a number of contexts. This paper proposes an automated multimodal detector of MW using eye gaze and physiology (skin conductance and skin temperature) and aspects of the context (e.g., time on task, task difficulty). Data in the form of eye gaze and physiological signals were collected as 178 participants read four instructional texts from a computer interface. Participants periodically provided self-reports of MW in response to pseudorandom auditory probes during reading. Supervised machine learning models trained on features extracted from participants’ gaze fixations, physiological signals, and contextual cues were used to detect pages where participants provided positive responses of MW to the auditory probes. Two methods of combining gaze and physiology features were explored. Feature level fusion entailed building a single model by combining feature vectors from individual modalities. Decision level fusion entailed building individual models for each modality and adjudicating amongst individual decisions. Feature level fusion resulted in an 11% improvement in classification accuracy over the best unimodal model, but there was no comparable improvement for decision level fusion. This was reflected by a small improvement in both precision and recall. An analysis of the features indicated that MW was associated with fewer and longer fixations and saccades, and a higher more deterministic skin temperature. Possible applications of the detector are discussed.

Categories and Subject Descriptors
H.5.m [Information Interfaces and Presentation]: Miscellaneous

General Terms
Human Factors

Keywords
Mind wandering; gaze tracking; user modeling; affect detection

1. INTRODUCTION
Mind wandering (MW) is a ubiquitous phenomenon that is characterized by an involuntary shift in attention away from the task at hand toward task-unrelated thoughts. Studies have shown that MW occurs frequently [17, 18, 29, 36] and that it is negatively correlated with performance across a variety of tasks requiring conscious control (see meta-analysis [25]). For example, MW is associated with increased error rates during signal detection tasks [33], lower recall during memory tasks [35], and poor comprehension during reading tasks [10, 31]. Based on this research, it is evident that performance on tasks that require attentional focus can be hindered by MW. This suggests that there is an opportunity to improve task performance by attempting to reduce MW and correct its negative effects. Automatic detection of MW is a fundamental step needed towards creating a system capable of responding to MW. As reviewed below, prior work has mainly focused on unimodal MW detection. This paper explores the potential benefits of multimodal MW detection by fusing eye gaze data and physiology data along with contextual cues during computerized reading.

1.1 Related Work
MW detection is most closely related to the field of attentional state estimation. Attentional state estimation has been explored in a variety of domains and with a variety of end-goals. For example, attention has been used to evaluate adaptive hints in an educational game [22] and to optimize the position of news items on a screen [23]. Attentional state estimators have been developed for several tasks such as identifying object saliency during video viewing [42], and for monitoring driver fatigue and distraction [8]. These studies mainly focus on unimodal detection of attention using eye gaze, though the studies on driver fatigue also use driving performance measures such as steering wheel motion.

Some studies combine multiple modalities for attention detection. For example, Stiefelhagen et. al [37] attempted to detect the focus of attention of individuals in an office space based on gaze and sound. They found that accuracy increased to 75.9% when combining both modalities, compared to an accuracy of 73.9% when using gaze alone. Similarly, Sun et. al [38], focused on detecting attention using keystroke dynamics and facial expression. They logged keystrokes and mouse movements while recording participants’ faces as they completed three tasks related to conducting research, such as searching for and reading academic papers. Combining features from these modalities resulted in an attention detection accuracy of 77.8%, which reflects a negligible improvement compared to the best unimodal model with an accuracy of 76.8%.
This work focuses on MW detection and shares similarities and differences from previous work on attentional state estimation. Although both attentional state estimation and MW detection entail identifying aspects of a user’s attention, MW detection is concerned with detecting more covert forms of involuntary attentional lapses. MW can be considered to be a form of looking without seeing – gaze might be focused on relevant parts of the interface (e.g., an image on a multimedia presentation), but attention is focused on completely unrelated thoughts (e.g., what to have for dinner).

There have recently been a number of studies that have attempted to create MW detectors using a variety of modalities, including acoustic prosodic information [9], eye gaze [4], physiology [5, 24], reading time [11], and interaction patterns [20]. A subset of these are briefly reviewed below.

Drummond and Litman [9] were the first to build a MW detector using acoustic-prosodic information. Participants were instructed to read and summarize a biology paragraph aloud while indicating at set intervals their degree of “zoning out” on a 7 point Likert scale. Responses of 1-3 were taken to reflect “low zone outs” while responses of 4-7 corresponded to “high zone outs.” Their model was able to discriminate between “low” versus “high” zone outs with an accuracy of 64%, which reflects a 22% improvement over chance. However, it is unclear if their model generalizes to new users as the authors did not use user-independent cross validation.

Eye gaze has shown considerable promise for MW detection due to decades of information showing support for a link between external information and eye movements during reading [16, 26]. For example, compared to normal reading, MW during reading is associated with longer fixation durations [27], a higher frequency of subsequent fixations on the same word [39], a higher frequency of blinks [36], and larger pupil diameters [12, 32]. Capitalizing on these relationships, Bixler et. al [4], used eye gaze features to detect MW while students read instructional texts with a computer interface. Their models attained an above chance classification accuracy of 28% using user-independent cross validation.

Physiology might also be a suitable modality for MW detection due to the relationship between physiological arousal from the sympathetic nervous system and attentional states [1]. In particular, a relationship between MW and skin conductance has previously been documented [33]. Blanchard et. al [5] leveraged this information to detect MW using galvanic skin response and skin temperature collected with the Affectiva Q sensor. The authors were able to achieve an above-chance classification accuracy of 22% in a manner that generalized to new users.

1.2 Current Study: Novelty and Contribution
The current study builds on our previous work on MW detection from eye gaze [4] and physiology [5] by considering a combination of modalities along with the inclusion of contextual cues. To the best of our knowledge, this represents the first multimodal MW detector and one of the few attempts to combine eye gaze, physiology, and contextual factors. This particularly unique combination of modalities poses interesting challenges (discussed below) since the three channels operate at rather different time scales.

The multimodal MW detector was developed in the context of computerized reading, which is a critical component of many real-world tasks. By focusing on a more general activity such as reading, we aim for the detector to be applied more broadly. There is also a particularly strong negative correlation between MW and reading comprehension [10, 29, 34], suggesting the need for automated methods to detect and eventually address MW during reading.

Our approach entails simultaneously collecting eye gaze data via an eye tracker and physiology data via a wearable sensor while participants read instructional texts on a computer screen. MW was tracked using auditory thought probes (discussed further below). Features are extracted from the eye gaze signal, physiology signal, and contextual cues from the reading interface. Supervised machine learning models used these features to detect the presence or absence of MW (from thought probes). We adopted rather simple modality fusion approaches in this early stage of research. Feature level fusion entailed creating one model after fusing features from the individual modalities. Decision level fusion entailed training a separate model for each modality and combining their decisions. We constructed and validate our models in a user-independent fashion that involves no user overlap across training and testing sets.

A challenge when combining gaze and physiology data pertained to how much data should be used to compute features. It is desirable to detect MW using the least amount of data possible so it can be detected sooner. Eye gaze is a modality that quickly reflects stimuli-related changes, while physiological signals like skin temperature and skin conductance take longer to respond [33]. Although smaller amounts of data might work when using eye gaze to detect MW, the same amount of data might not adequately capture changes in the physiological signal. The previous related studies built MW detectors using between 3-30 seconds of data, so it was important to investigate how the amount of data monitored affected detection accuracy. It could also be the case that a different amount of data are optimal for unimodal versus multimodal detection. Hence, we built a variety of single and multimodal models using different window sizes and combinations of window sizes.

2. DATA COLLECTION
2.1 Participants
The entire dataset consisted of 178 undergraduate students that participated for course credit. Of these, 93 participants were from a medium-sized private Midwestern university while 85 were from a large public university in the South. The average age of participants was 20 years (SD = 3.6). Demographics included 62.7% female, 49% Caucasian, 34% African American, 6% Hispanic, and 4% “Other.” There was no physiology data available for the large public university in the South. As a result, we focused on data from the medium-sized private Midwestern university.

2.2 Texts and Experimental Manipulations
Participants read four different texts on research methods topics (experimenter bias, replication, causality, and dependent variables) adapted from a set of texts used in the educational game Operation ARA! [15]. On average the texts contained 1,500 words (SD = 10) and were split across 30-36 pages with approximately 60 words per page. Each page was presented on a computer screen with size 36 Courier New font.

There were two within-subject experimental manipulations: difficulty and value. The difficulty manipulation consisted of presenting either an easy or a difficult version of each text. A text was made more difficult by replacing words and sentences with
more complex alternatives while retaining semantics, length, and content. The value manipulation involved designating each text as either a high-value text or a low-value text. Participants were informed that a post-test followed the reading, and questions from high-value texts would be worth three times as much as questions from low-value texts.

Each participant read four texts on four topics in four experimental conditions (easy-low value; easy-high value; difficult-low value; difficult-high value). The order of the four texts, experimental conditions, and assignment of condition to text was counterbalanced across participants using a Graeco-Latin Square design. The difficulty and value manipulations were part of a larger research study [19] and are only used here as contextual features when building the MW detectors.

2.3 Mind Wandering Probes

Nine pseudorandom pages in each text were identified as probe pages. An auditory thought probe (i.e., a beep) was triggered on probe pages at a randomly chosen time interval 4 to 12 seconds after the page appeared. These probes were considered to be within-page probes. An end-of-page probe was triggered if the page was a probe page and the participant attempted to advance to the next page before the within-page probe was triggered. Participants were instructed to indicate if they were MW or not by pressing keys marked “yes” or “no,” respectively. MW was defined to participants as follows: “At some point during reading, you may realize that you have no idea what you just read. Not only were you not thinking about the text, you were thinking about something else altogether.” Thought probes are a standard and validated method for collecting online MW reports [21, 34]. Although misreports are possible, alternatives for tracking MW such as EEG or fMRI have yet to be sufficiently validated and are not practical in many real world applications due to their cost.

2.4 Procedure

All procedures were approved by the ethics board of both Universities prior to data collection. After signing an informed consent, participants were asked to be seated in front of either a Tobii TX 300 or Tobii T60 eye tracker depending on the university (both were in binocular mode). The Tobii eye trackers are remote eye trackers, so participants could read freely without any restrictions on head movement. In addition, an Affectiva Q sensor was strapped to the inside of the participant’s non-dominant wrist, a standard placement to measure skin conductance.

Participants completed a multiple choice pretest on their knowledge of the research methods topics discussed by a 60-second standard eye tracking calibration procedure. Participants were then instructed how to respond to the MW probes based on the definition discussed above and instructions from previous studies [10]. Next, they then read four texts for an average reading time of 32.4 minutes (SD = 9.09) on a page-by-page basis, using the space bar to navigate forward. Participants then completed a post-test and were fully debriefed.

2.5 Instances of Mind Wandering

There were a total of 6,408 probes with responses from the 178 participants. However, there were two factors that reduced the final number of instances. First, there was only physiology data for participants from one of the two schools, which reduced the number of instances to 3,361. Second, within-page and end-of-page probes capture different types of MW, so separate models had to be built for each type of probe. This resulted in 2,556 instances for within-page probes and 805 instances for end-of-page probes. End-of-page models are not analyzed further because of the relatively low number of instances. Of the remaining 2,556 within-page probes, participants responded “yes” 663 times, resulting in a MW rate of 26%.

3. SUPERVISED CLASSIFICATION

The goal was to build supervised machine learning models from short windows of data prior to each MW report. We began by detecting eye movements from the raw gaze signal and computing features based on the eye movements within each window. We then processed the physiological signal by filtering measurements compromised by abrupt movements based on accelerometer readings from the Affectiva Q [5]. A variety of operations were applied to the datasets, such as resampling the training set to correct class imbalance through downsampling the majority class or oversampling the minority class, and removing outliers. Our models were evaluated with a leave-one-participant-out cross-validation method, where data for each participant was classified using a model built from the data from the other participants.

3.1 Feature Engineering

3.1.1 Gaze Features

The first step was to convert the raw gaze data into eye movements. Gaze fixations (points where gaze was maintained on the same location) and saccades (eye movements between fixations) were estimated from raw gaze data using a fixation filter algorithm from the Open Gaze And Mouse Analyzer (OGAMA), an open source gaze analyzer [40]. Next, the time series of gaze fixations was segmented into windows of varying length (4, 6, and 8 seconds), each ending with a MW probe. The windows ended immediately before the auditory probe was triggered in order to avoid confounds associated with motor activities in preparation for the key press in response to the probe.

The window sizes of 4, 6, and 8 seconds were selected based on the constraints of the experimental design in that probes occurred between 4 to 12 seconds into the onset of a page. Windows shorter than four seconds were not considered because they likely did not contain sufficient data to compute gaze features. Window sizes above 8 were not considered because there were too few pages where the probe occurred past the 8 second mark to build meaningful models. Furthermore, windows that contained less than five fixations were also eliminated due to insufficient data for the computation of gaze features.

Two sets of gaze features were computed: 46 global gaze features and 23 local gaze features, yielding 69 gaze features overall. The features are listed below but the reader is directed to [3] for detailed descriptions of the gaze features.

Global gaze features were independent of the words being read and consisted of two categories. Eye behavior measurements captured descriptive statistics of five properties of eye gaze: fixation duration, saccade duration, saccade distance, saccade angle, and pupil diameter. For each of these five behavior measurements, we computed the minimum, maximum, mean, median, standard deviation, skew, kurtosis, and range, thereby yielding 40 features. Additional global features included number of saccades, proportion of horizontal saccades, blink count, blink time, fixation dispersion, and fixation duration/saccade duration ratio.
Unlike global features, local gaze features were sensitive to the words being read. The first set of local features captured information pertaining to different fixation types. These included first pass fixations, regression fixations, single fixations, gaze fixations, and non-word fixations [4]. Specific local features included the mean and standard deviation of the durations of each fixation type and the proportion of each fixation type compared to the total number of fixations. This resulted in 15 local features. The number of end-of- clause fixations was also used as a feature, based on the well documented sentence wrap-up effect [41], which suggests that fixations at the end of clauses should result in longer processing times.

Four additional local features captured the extent to which well-known relationships between characteristics of words (in each window) and gaze fixations were observed. These included Pearson correlations between fixation durations and length, hypernym depth, global frequency [2], and synset size of a word. These features exploit known relationships during normal reading, such as a positive correlation between word length and fixation duration, which are expected to break down due to the perceptual decoupling associated with MW [28].

Three final local features captured relationships between the movements of the eye across the screen with respect to the position of the words on the screen. These included the proportion of cross line saccades, words skipped, and the reading time ratio.

3.1.2 Physiology Features

Physiology features were calculated from skin conductance (SC) and skin temperature (ST) signals collected with an Affectiva Q sensor attached to participants’ wrists. In all, 89 physiology features were calculated. Similar to the gaze data, features were calculated from a specific period of time prior to each probe. However, physiology data is not as sensitive to changing screens (when participants advance to the next page), so it was feasible to compute larger windows that extended beyond page boundaries. Window sizes were chosen based on results from a previous study [5], and included both short (i.e., 3 and 6 seconds) and long (i.e., 20, and 30 seconds) windows.

The physiology data were preprocessed in several ways. First, both SC and ST signals were z-score standardized at the participant level to mitigate individual differences in physiological responses. Second, a low pass 0.3 Hz filter was applied to the SC data in order to reduce noise. Finally, abrupt movements were detected using the accelerometer data (also collected by the sensor) and were removed because they can introduce noise in the physiological signals.

The same 43 features were calculated from the SC signal and the ST signal, resulting in 86 features. Signal processing techniques were employed to create different transformations of each signal and the mean, standard deviation, maximum, ratio of peaks to signal length, and ratio of valleys to signal length of the different versions were computed and used as features. The different transformations included the standardized signal, an approximation of the derivation of the signal (D1) obtained by taking the difference between the data points of the original signal, the second derivative (D2) obtained by taking the difference between the data points of D1, the phase component and the magnitude component of the signal computed with a Fourier transform; the spectral density of the original signal obtained using Welch’s method; the autocorrelation of the original signal; and the magnitude squared coherence. Finally, the slope and y-intercept of the slope coefficient of the linear trend line were also calculated. Additional details on the features can be found in [5].

3.1.3 Context Features

Context features were considered to be secondary to gaze and physiological features because they mainly captured situational rather than individualized aspects of the reading task. Context features included session time, text time, session page number, text page number, average page time, previous page time, ratio of the previous page time to the average page time, current difficulty, current value, previous difficulty and previous value. In all, there were 11 context features that are fully described in [4].

3.2 Model Building

Supervised classifiers were built using the aforementioned features to discriminate positive instances of MW (responding “yes” to a MW probe) from negative instances (responding “no” to a MW probe) of MW. Thirteen supervised machine learning algorithms from Weka [13] were used, including Bayesian models, decision tables, lazy learners, logistic regression, and support vector machines. Models were built from datasets with a number of varying parameters in order to identify the most accurate models as well as to explore how different factors affect classification accuracy.

The first parameter varied was the window size used. For gaze data, windows were 4, 6, or 8 seconds before each MW probe. A wider range of window sizes was available for physiology data, so window sizes of 3, 6, 20, and 30 were used. The same window sizes were used for both ST and SC data.

Second, feature selection was applied to the training set only in order to remove collinear features (e.g., number of fixations and number of saccades) and to identify the most diagnostic features. Features that were strongly correlated with other features but weakly correlated with MW reports were ranked using correlation-based feature selection (CFS) [14]. The top 20%, 30%, 40%, 50%, 60%, 70%, or 80% of features ranked by CFS were included; the specific percentage was another parameter that was varied.

Third, outlier treatment was performed in order to remove the destabilizing effect of outliers on our models. Outlier treatment consisted of replacing values greater/lower than 3 standard deviations above/below the mean with the corresponding value +3 or -3 standard deviations above/below the mean. Datasets were constructed either with outliers treated or with outliers retained.

Fourth, the training data was resampled to maintain an even class distribution (“no” MW responses accounted for 74% of all responses), as an uneven class distribution can have adverse effects on classification accuracy. Dowsampling consisted of removing instances from the most common MW response (i.e., “no” responses) at random until there were an equal number of “yes” and “no” responses in the training set. Oversampling consisted of using the SMOTE (Synthetic Minority Oversampling Technique) algorithm [6] as implemented in Weka to oversample the minority class by generating synthetic surrogates. Importantly, the sampling methods were only applied to the training data, but not the testing data.

Finally, we varied the features used and the method of fusing modalities. Unimodal models included gaze and context features or physiology and context features. Context features were combined with gaze and physiology features rather than included
alone because previous research indicated that are mainly complementary features that improve accuracy rather than stand on their own [4, 5]. As the inclusion of the context features was consistent across models, any multimodal improvement could be attributed to the fusion of the physiology and gaze features. We also tested two standard methods of fusing the modalities as discussed below.

### 3.3 Multimodal Fusion

Two approaches were used for multimodal fusion. For feature level fusion, gaze and physiology features were calculated separately. The first step was to combine the feature vectors corresponding to the same probe. If either gaze features or physiology features were missing for a given probe the entire instance was removed. For feature level fusion, it was deemed important that each modality be equally represented in the set of features. Therefore, an equal number of features was selected from each modality. For example, suppose there were 32 features after collinear features were removed. If feature selection was set to select 50% of the features, then eight features from one modality and eight features from the other would be selected, resulting in a total of 16 features.

Decision level fusion consisted of building two unimodal models and combining the results to reach a final conclusion. The unimodal models provided a confidence (ranging from 0 to 1) for each instance being classified as MW. The confidence values for each model were averaged. If the average confidence was above 0.5, the instance was classified as a positive instance of MW, otherwise it was classified as a negative instance of MW. In the case where there was confidence from only one modality due to missing data, the final decision was based on the modality with available data only.

### 3.4 Validation

A leave-one-participant-out validation method was used to ensure that data from each participant was exclusive to either the training or testing set. Using this method, data from one participant were held aside for the testing set while data from the remaining participants were used to train the model. Furthermore, a nested cross validation was performed on the training set in order to optimize feature selection. Specifically, data from a random 66% of the participants in the training set were used to perform feature selection. Feature selection was repeated five times in order to remove the variance caused by a random selection of participants. The feature rankings were averaged over these five iterations, and a certain percentage (ranging from 20%-80% - see above) of the highest ranked features were included in the model. Sampling was performed after feature selection and was repeated five times.

Model performance was evaluated using the kappa value [7]. Which corrects for random guessing in the presence of uneven class distribution, as is the case for our data. The kappa value is calculated using the formula \( K = (\text{Observed Accuracy} - \text{Expected Accuracy}) / (1 - \text{Expected Accuracy}) \), where Observed Accuracy is equivalent to recognition rate and Expected Accuracy is computed from the marginal class distributions. Kappa values of 0, 1, > 0, and < 0 indicate chance, perfect, above chance, and below chance agreement, respectively.

### 4. RESULTS

#### 4.1 Feature Level Fusion

We built a feature level fusion model for each variation of the parameters as described above and selected the best performing feature level fusion model based on two criteria. We first identified models with a precision greater than .5. From those models, we chose the one with the highest kappa value. We then compared these models to two unimodal models (gaze-context, physiology-context) chosen in the same manner. The unimodal models were built using only the instances that contained data from both modalities so that the same instances were used to build both the unimodal models and the feature level fusion models. The best feature level fusion models and the corresponding unimodal models for each combination of window sizes are shown in Table 1.

<table>
<thead>
<tr>
<th>Window Size</th>
<th>N</th>
<th>Kappa</th>
<th>% Better</th>
</tr>
</thead>
<tbody>
<tr>
<td>G+C</td>
<td>6</td>
<td>.07</td>
<td>.11</td>
</tr>
<tr>
<td>P+C</td>
<td>6</td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td>G+C</td>
<td>10</td>
<td>.10</td>
<td>.11</td>
</tr>
<tr>
<td>P+C</td>
<td>10</td>
<td>.12</td>
<td>.11</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td>G+C</td>
<td>30</td>
<td>.11</td>
<td>.09</td>
</tr>
<tr>
<td>P+C</td>
<td>30</td>
<td>.11</td>
<td>.09</td>
</tr>
<tr>
<td>G+C</td>
<td>6</td>
<td>.09</td>
<td>.11</td>
</tr>
<tr>
<td>P+C</td>
<td>6</td>
<td>.10</td>
<td>.11</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
<td>.09</td>
<td>.11</td>
</tr>
</tbody>
</table>

Note: G = gaze; P = physiology; C = context; F = feature level fusion

For each window size combination, we calculated the percent improvement of the feature level fusion model over the best unimodal model. In three cases the gaze-context model performed best; this occurred for window sizes of 4 and 20 and 8 and 20 for gaze and physiology respectively. However, in every other case the feature level fusion model performed either better than or equivalent to either unimodal models alone. On average, there was an 11% improvement in multimodal classification accuracy compared to accuracy of the best unimodal model. The overall best model was a feature level fusion model with a kappa of .19, reflecting a 27% improvement over the best unimodal model (kappa of 0.15).

A previous study [4] found that gaze window size had an effect on classification accuracy. In order to investigate this effect for our feature level fusion models, we averaged the best models of each gaze and physiology window size (Table 2). We note that as the gaze window size increases, so too does the accuracy of the unimodal models and the feature level fusion models. This is likely due to more available information for the larger window sizes. However, the improvement due to feature level fusion remained consistent among all window sizes. The physiology window size did not have a noticeable effect on the overall accuracy of the models, though the greatest improvement was seen with a physiology window size of 3.

We then analyzed the effect of feature level fusion by comparing the precision and recall between the best feature level fusion model and the two unimodal models with the same window size. The gaze model had a precision of .607 and a recall of .327 while the physiology model had a precision of .503 and a recall of .336. The feature level fusion model improved both the number of
instances of MW that were correctly classified, with a precision of .613, as well as increased the number of instances classified as MW, with a recall of .351. Therefore, feature level fusion resulted in a subtle but definitive improvement.

Table 2. Results for feature level fusion models aggregated across window sizes

<table>
<thead>
<tr>
<th>Window Size</th>
<th>Avg N</th>
<th>Kappa</th>
<th>% Better</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>G+C</td>
<td>P+C</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>.10</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>.12</td>
<td>.13</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>.14</td>
<td>.16</td>
</tr>
<tr>
<td>P</td>
<td>3</td>
<td>.11</td>
<td>.14</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>.13</td>
<td>.14</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>.13</td>
<td>.14</td>
</tr>
</tbody>
</table>

Note: G = gaze; P = physiology; C = context; F = feature level fusion

4.2 Decision Level Fusion

The results for decision level fusion are shown in Table 3. The decision level models and unimodal models were chosen using the same selection criteria as was used for selecting the feature level fusion models. Of note is that the unimodal models in this comparison contained a larger number of instances than the unimodal models in the feature level fusion comparison. This is because any instances without data from both modalities were removed when building the feature level fusion models. However, each unimodal model within the decision level model fusion used all the available instances regardless of the other modality.

In contrast to the feature level fusion models, only three decision level models resulted in a positive improvement (window sizes of 4 and 6, 8 and 3, and 8 and 30 for gaze and physiology respectively). Every other decision level model resulted in a lower kappa value than the best associated unimodal model. On average, decision level fusion had a negative effect on classification accuracy, and will not be analyzed further.

Table 3. Results for best decision level models

<table>
<thead>
<tr>
<th>Window Size</th>
<th>N</th>
<th>Kappa</th>
<th>% Better</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>G+C</td>
<td>P+C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G+C</td>
<td>P+C</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2020</td>
<td>1995</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>2020</td>
<td>2006</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>2020</td>
<td>2011</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>2020</td>
<td>2013</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1464</td>
<td>1995</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1464</td>
<td>2006</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>1464</td>
<td>2011</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>1464</td>
<td>2013</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>906</td>
<td>1995</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>906</td>
<td>2006</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>906</td>
<td>2011</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>906</td>
<td>2013</td>
</tr>
</tbody>
</table>

Note: G = gaze; P = physiology; C = context; D = decision level fusion

4.3 Feature Analysis

We also investigated how the features varied between positive (“yes” responses to a MW probe) and negative (“no” responses to a MW probe) instances of MW in order to obtain a better understanding of how eye gaze and physiology differ during MW compared to normal reading. The mean values of each feature for “yes” and “no” MW responses were analyzed using a paired samples t-test. We chose to analyze the dataset associated with the best performing feature level fusion model shown in Table 1 above. It was necessary that there to be at least one “yes” response and one “no” response from each participant in order to perform a paired samples t-test. As a result, only 39 participants with both types of responses in this dataset were included in the analysis. The features that were significantly different (two tailed p < .05) between “yes” and “no” MW responses are included in Table 4.

There are several conclusions that can be drawn from Table 4. Saccades were greater in duration and fewer in number (implying fewer fixations as well) during MW compared to normal reading. In addition, there were fewer horizontal saccades during MW. Fixations were also greater in duration, evidenced by a greater mean, median, standard deviation, and maximum overall fixation duration. This is also supported by a greater mean duration for single fixations, first pass fixations, and gaze fixations, and a greater standard deviation for first pass fixations. A greater fixation duration during MW is corroborated by previous research [27]. Thus, with respect to eye gaze, MW was associated with fewer, but longer fixations along with longer and more irregular saccades (since saccades during reading should largely be horizontal). There were also differences in ST, such as a greater number of peaks in the autocorrelation of the ST signal, fewer peaks and valleys in the frequency of the ST signal, and a greater mean ST during MW. These results suggest that the ST was higher and more consistent during MW.

Table 4. Mean value of features for Yes and No responses to MW probes for the best performing feature level fusion model

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean Yes</th>
<th>Mean No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Saccades</td>
<td>51.79</td>
<td>58.89</td>
</tr>
<tr>
<td>Fixation Duration Mean</td>
<td>242.89</td>
<td>226.64</td>
</tr>
<tr>
<td>Fixation Duration Median</td>
<td>216.98</td>
<td>205.64</td>
</tr>
<tr>
<td>Fixation Duration SD</td>
<td>110.58</td>
<td>89.93</td>
</tr>
<tr>
<td>Fixation Duration Max</td>
<td>574.66</td>
<td>504.59</td>
</tr>
<tr>
<td>Single Fixation Duration Mean</td>
<td>249.78</td>
<td>222.34</td>
</tr>
<tr>
<td>First Pass Fixation Duration Mean</td>
<td>239.48</td>
<td>225.94</td>
</tr>
<tr>
<td>First Pass Fixation Duration SD</td>
<td>96.62</td>
<td>81.90</td>
</tr>
<tr>
<td>Gaze Fixation Duration Mean</td>
<td>240.55</td>
<td>222.90</td>
</tr>
<tr>
<td>Saccade Duration Median</td>
<td>26.65</td>
<td>20.33</td>
</tr>
<tr>
<td>Horizontal Saccade Proportion</td>
<td>.96</td>
<td>.98</td>
</tr>
<tr>
<td>Temp Autocorrelation Peak Ratio</td>
<td>.21</td>
<td>-.10</td>
</tr>
<tr>
<td>Temp Frequency Valley Ratio</td>
<td>-.22</td>
<td>.10</td>
</tr>
<tr>
<td>Temp Frequency Peak Ratio</td>
<td>-.21</td>
<td>.10</td>
</tr>
<tr>
<td>Temp Standardized Signal Mean</td>
<td>.10</td>
<td>-.17</td>
</tr>
</tbody>
</table>

Note: SD = standard deviation; degrees of freedom = 38; all differences are significant at p < .05

5. DISCUSSION

The paper focused on building the first multimodal MW detector using eye-gaze, physiology, and context features. In the remainder of this section, we discuss our main findings, consider applications of the MW detector, and discuss limitations and avenues for future work.
5.1 Main Findings

Our results highlight a number of important findings for building MW detectors. Most importantly, we developed the first MW detector built using both gaze and physiological data. Our results show that there is an average improvement in kappa value of 11% when using feature level fusion compared to unimodal detectors alone. In contrast, decision level fusion resulted in a small 5% drop in accuracy when compared to unimodal detection. Feature level fusion also resulted in a small improvement in both precision and recall compared to unimodal classification. We also found that larger gaze window sizes resulted in higher kappa values overall, which is consistent with previous research [3]. This is likely because smaller gaze windows provide less evidence to distinguish MW from normal reading, resulting in lower accuracy. It is also possible that the difference in window sizes is partially due to the methodology of the study. The instances with a gaze window of 8 seconds are a subset of those with a gaze window of 4 seconds. It could be the case that reports that occur before 8 seconds are more difficult to classify than those that occur later. In contrast to gaze window sizes, there was not a clear difference in accuracy across different physiology window sizes, suggesting that the choice of physiology window size matters less. Further research with larger gaze window sizes and with different tasks would be useful towards determining an ideal gaze window size, and if it is consistent across tasks. Finally, we analyzed how features differed between positive and negative instances of MW. Our findings were largely similar to previous findings, in that MW was associated with fewer fixations of greater duration [27] and a deviation from normal reading patterns exhibited by greater saccade durations and fewer horizontal saccades [3]. We also discovered that skin temperature followed more deterministic patterns and was higher during MW compared to normal reading.

5.2 Applications

The primary application of a MW detector is inclusion in an interface with the intent of improving productivity. MW has been shown to negatively affect text comprehension [25], so any interface that includes text comprehension could be improved by dynamically responding to MW. One example of an intervention is to recommend to re-read or self-explain a passage when MW is detected. Interventions should not disrupt the participant if MW is detected incorrectly and should be used sparingly so participants are not overwhelmed. In addition to reading, it is possible that MW detection and intervention could also be attempted across a wider array of tasks and contexts. For example, outside of learning contexts, a MW detector could be employed in situations requiring vigilance such as driving or operation of a control room. More research is needed to understand the need for and potential of automatic MW detection systems in different domains.

5.3 Limitations and Future Work

There are some limitations to the current study. The data was collected in a lab environment and participants were limited to undergraduates located in the United States. This limits our claims of generalizability to individuals from different populations. In addition, physiological data was available for only half of the participants. This ended up resulting in a smaller amount of data available to build models. This is one potential cause of the lower unimodal kappa values in relation to previous MW detectors [4, 5]. With this in mind, a second study with data from both modalities and from both schools is warranted. Another limitation is that an expensive, high quality eye tracker was used for data collection, which limits the scalability of using eye gaze as a modality for MW detection. However, this could be addressed by the decreasing cost of consumer-grade eye tracking technology, such as Eye Tribe and Tobii EyeX, or with promising alternatives that use webcams for gaze tracking [30]. It is also possible that participants did not provide accurate or honest self-caught reports. This is a clear limitation although it should be noted that both the probe-caught and self-caught methods have been validated in a number of studies [33, 34] and there is no clear alternative for tracking a highly internal state like MW. Alternatives such as EEG and fMRI can be costly, and their efficacy is equivocal. Another limitation is our moderate classification results, with the best user-independent kappa of .19. There are several reasons for the moderate accuracy. MW is a subtle and highly internal state and little is known about the onset and duration of MW. Further, we did not optimize our model for individual users or optimize the classification algorithm parameters, which would ostensibly improve performance but also run the risk of overfitting.

5.4 Concluding Remarks

In summary, this study demonstrated the possibility of a MW detector that uses both eye movements and physiology data along with contextual cues. We combined these modalities through feature level fusion and decision level fusion. Feature level fusion resulted in an average improvement of 11% while there was no improvement for decision level fusion. Importantly, our approach used an unobtrusive remote eye tracker and wrist-attached physiological sensor, thereby allowing for unrestricted head and body movement. The next step involves integrating the detector into an adaptive system in order to trigger interventions that attempt to reorient attentional focus when MW is detected.

Acknowledgment. This research was supported by the National Science Foundation (DRL 1235958). Any opinions, findings and conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of NSF.

References


