Automated Physiological-Based Detection of Mind Wandering during Learning

Nathaniel Blanchard[1], Robert Bixler[1], Tera Joyce[1], and Sidney D'Mello[1,2]

[1] Department of Computer Science
[2] Department of Psychology, University of Notre Dame, Notre Dame, IN 46556
{nblancha,rbixler,tjoyce4,sdmello}@nd.edu

Abstract. Unintentional lapses of attention, or mind wandering, are ubiquitous and detrimental during learning. Hence, automated methods that detect and combat mind wandering might be beneficial to learning. As an initial step in this direction, we propose to detect mind wandering by monitoring physiological measures of skin conductance and skin temperature. We conducted a study in which student's physiology signals were measured while they learned topics in research methods from instructional texts. Momentary self-reports of mind wandering were collected with standard probe-based methods. We computed features from the physiological signals in windows leading up to the probes and trained supervised classification models to detect mind wandering. We obtained a kappa, a measurement of accuracy corrected for random guessing, of .22, signaling feasibility of detecting MW in a student-independent manner. Though modest, we consider this result to be an important step towards fully-automated unobtrusive detection of mind wandering during learning.

Keywords: skin conductance, skin temperature, mind wandering, machine learning.

1 Introduction

Almost everyone has had the experience of attempting to concentrate on a learning task and suddenly realizing that their mind has drifted elsewhere. As a result they may have missed key pieces of information and are forced to review the missed material. This phenomenon, called mind wandering (MW), can be described as involuntarily engaging in conscious off-task thoughts without the metacognitive realization that this has occurred [1]. MW has been linked to lower performance on a number of tasks including poor comprehension during reading [2] and low recall during memory encoding [3]. Furthermore, MW is difficult to address immediately because people initially lack conscious awareness of that fact that they are MW. Given the ubiquity and negative consequences of the phenomenon, it might be beneficial for intelligent tutoring systems (ITSs) and other educational technologies to detect when MW occurs and then intervene to restore attention to the task at hand. As an initial step in this direction, this paper reports research aimed at developing a fully-automated system to detect momentary occurrences of MW in a manner that generalizes to new students.

S. Trausan-Matu et al. (Eds.): ITS 2014, LNCS 8474, pp. 55–60, 2014.

Related Work. MW detection is a relatively unexplored field. Drummond and Litman (2010) were one of the first to attempt automatic MW detection. They used prosodic and lexical features of student responses to a spoken ITS. Students were probed at set intervals into if they were MW. Their models were able to discriminate high and low MW with an accuracy of 64%. However, their models were only applicable to ITSs with student speech, and their validation method did not ensure generalization to new students [4].

D'Mello, Cobian, and Hunter (2013) furthered work on MW detection by building supervised classification models that automatically detected MW during reading from eye gaze features obtained with commercial eye trackers. Their model obtained a kappa, a measurement of accuracy corrected for chance, of 0.23 [5]. Though their validation method ensured generalizability to new students, their approach is limited to reading tasks. Furthermore, the use of eye tracking has some scalability concerns.

Current Study. The present study focused on detecting MW by monitoring two physiological signals: skin conductance (SC) and skin temperature (ST). These signals were collected using a wearable sensor at a fraction of the cost of commercial eye trackers. The use of physiology to track MW is motivated by the relationship between sympathetic nervous activity (captured by SC and ST) and attentional states [6].

A previous study found a higher rate of MW was related to overall lower levels of skin conductance (SC) [7]. However, this result was not leveraged to build automated MW detectors. To our knowledge, no attempt has been made to build models capable of detecting MW using SC or ST signals, nor has there been research attempting to link ST and MW. Taking a step in this direction, we collected a large data set where students were periodically probed to report instances of MW during computerized learning from instructional texts. These signals were used to create machine learning models that predicted MW.

2 Methods

Data Collection. Participants were 70 undergraduate students from a medium-sized private mid-western University in the U.S. Students were seated in front of a computer and an Affectiva Q sensor was strapped to the inside of the student's non-dominate wrist, a standard placement to measure SC [8]. The Affectiva Q [9] provides a nonintrusive way to measure SC and ST of the student at sampling rates of 8 Hz.

Students were asked to study four texts, each on key research methods topics: experimenter bias, replication, causality, and dependent variables. On average, each text contained 1500 words (SD = 10 words) with approximately 60 words per page. Students were informed that they would be asked a series of test questions on each text after reading. Before each text, students were made aware of the point value of test questions related to the text – "high-value" text questions were worth three times more than "low-value" text questions. This was the value manipulation. In addition, there were also difficult vs. easy versions of the texts equated in terms of content and length (difficulty manipulation). These manipulations were integral to a larger research study, but are not the focus of this research.

As students progressed through the texts they were instructed to report if they were MW by responding to auditory probes. Auditory probes occurred at a random point 4 to 12 seconds from the beginning of pseudo randomly chosen probe pages. These probes are classified as "within-page" probes. If students attempted to advance to the next page before the probe appeared, they were probed with an "end of page" probe. Once an auditory probe occurred, students used a keyboard to indicate MW with a "yes" or normal reading with a "no" by selecting appropriate keys on the keyboard.

Students reported MW to end of page probes 16.9% of the time (N = 108), and they reported MW to within page probes 26.1% of the time (N = 526).

Model Building. Supervised classification was conducted to detect instances of MW from physiological signals and contextual features (discussed below). Models were built using WEKA [10] and were validated at the student-level - data was randomly split on students, with 67% for training and 33% in the testing set and repeated for 25 iterations. SC and ST signals were z-score standardized at the student level and a low pass filter was applied to the SC data at 0.3 Hz to reduce noise in the signal.

To account for physiological measurements compromised by abrupt movements, the average difference between consecutive x, y, and z accelerometer readings for each student was calculated from an accelerometer in the Affectiva Q. A threshold of five times the average was used to eliminate compromised data, as has been used in previous studies [11]. In instances where this threshold was reached, data 5/8ths of a second before the movement through 5/8ths of a second after the movement was discarded.

Features were extracted from windows of signal data between the triggering of the auditory probe and a variable number of seconds before the probe. Separate datasets were constructed for window lengths of 3, 6, 12, 20, and 30 seconds.

Physiological features were extracted from the SC and ST signals included the mean, standard deviation, maximum, the ratio of maxima, and ratio of minima [12]. These statistical features were calculated for: the *standardized signal*; an approximation of the *derivation of the signal* (D1) obtained by taking the difference from one data point to the next; an approximation of D1, or the *second derivate* (D2) [13]; the *frequency*, and *magnitude* obtained from the Fast Fourier transformation [11]; the spectral density of the signal with *Welch's* method; the *autocorrelation* of the signal at lag 10, and, in models where both ST and SC of the same window were used, the *magnitude squared coherence* between the signals. Other physiological features included slope and y-intercept of the slope *coefficient* of the linear trend line [13].

In all, 43 features from the SC signal and the same 43 from ST were extracted. A separate dataset was created for each combination of window sizes of SC and ST data in order to address different temporal combinations of these signals (e.g. SC data was extracted for a window size of 3 seconds while ST was extracted for a window size of 30 seconds). Coherence statistics were used if the window sizes matched.

Context features captured the context of the learning task and included features for text, timing, and difficulty and value. Difficulty and value features included the *current difficulty* and *current value* of the text and the *previous difficulty* and *previous value* of the previous text. Timing features include *total time elapsed* since the student started the reading portion of the experiment, the *time since starting the current* text, the *average page time*, the *previous page time*, and the ratio of *previous page time to average page time*. Text features were the *total number of pages* that the student had

read since starting the reading portion of the session and the *page number* of the current text. In all, there were 11 context features.

Data treatments were applied in various combinations to determine which combination of data treatments resulted in the most accurate model. First, tolerance analysis was used to eliminate features that exhibited multicollinerity. Second, three feature selection algorithms (Gain-Ratio, Info-Gain, or ReliefF) were used (on training data only) to rank the contribution of each feature, and either 25%, 50%, 75%, or 90% of the top features were selected. Third, the data was winsorized by setting outliers greater than 3 standard deviations from the mean to the corresponding value 3 standard deviations from the mean. Fourth, downsampling was applied to the training data to obtain an equal distribution of responses by randomly removing instances of the more frequent class until the classes were balanced. Fifth, SMOTE (oversampling) was applied to the training data by adding random synthetic samples of the less frequent class until the classes were balanced. Sixth, when context features were not used, probes were eliminated if the student spent less than 4 seconds on a probe page, as the student likely either was not reading or accidently advanced prematurely.

3 Results

Table 1 presents the kappa, a measurement of accuracy which corrects for random guessing, of the best models (highest kappa). The best models were standardized and outliers were winsorized. Neither of the best models used tolerance, downsampling, or oversampling. Within page MW responses were easier to detect (kappa = .22, *SD* across iterations = .11) than end of page probes (kappa = .14, *SD* = .11). As seen from the confusion matrices in Table 2, although the best models have a high true negative rate (accurately detecting when not MW), the hit rate (correctly detecting MW) was low.

Table 1. Models with kappas

Probe Type	Features	Window (SC, ST)	No. Feat	Classifier	Kappa
Best WP	SC+ST+CF	(3, 12)	36	Filtered Classifier	0.22
Best EoP	ST	20	34	LADTree	0.14
Alt. WP	SC+CF	30	7	LADTree	0.15
Alt. EoP	ST+CF	6	23	AdaBoost M1	0.10

Note. WP – within page; EoP = end of page; Alt = Alternative;

To address the low hit rates, we considered alternate models as shown in Table 2. These models have a lower kappa for within page (kappa = .15, *SD* = .11) and end of page probes (kappa = .10, *SD* = .09), but have higher MW hit rates. Both alternative models were standardized within subjects, winsorized, used context features, and were trained with upsampling. Neither model used tolerance analysis. The use of upsampling in both models may indicate that with more positive MW reports, higher rates of MW can be detected.

Table 2. Confusion matricies for models

Model	Best Models			Alternative Models		
		Predicted			Predicted	
	Actual	*Yes*	*No*	**Actual**	*Yes*	*No*
Within page	*Yes* (.26)	.30	.70	*Yes* (.26)	.57	.43
	No (.74)	.11	.89	*No* (.74)	.38	.62
End of Page	*Yes* (.16)	.14	.86	*Yes* (.17)	.41	.59
	No (.84)	.04	.96	*No* (.83)	.28	.72

Note. Prior probabilities (base rates) are in parantheses

4 General Discussion

We investigated the possibility of detecting MW, a frequent and harmful phenomenon, from two physiological markers and aspects of the interaction context. MW detection is in its infancy; hence our immediate goal was to demonstrate the feasibility of MW detection. The major finding of this work is that SC and ST both contain information that can be used to detect MW. We acknowledge that our detection rates are modest, but consider them to be promising as an initial investigation into the possibility of unobtrusive detection of momentary instances of MW, an elusive state that is difficult to study since it is a highly internal unconscious phenomenon. Our detection is complicated by the relatively low rates of MW (23.9% of probes), which complicates supervised classification. Furthermore, we attempted to detect MW in a student-independent fashion, which is important for generalizability to new students, but more challenging due to individual differences in physiological responding [6].

MW detection has a number of possible applications. Interventions could be initiated during moments of MW in learning sessions to increase engagement. For example, an ITS that has detected MW could reevaluate the difficulty of the task the student is undertaking or could attempt to reengage the student's attention.

There are a few limitations that need to be addressed in future studies. One limitation is the relatively small data set used to train the models, so replicating the study with a larger sample is warranted. The study was conducted in a lab since we were interested in a highly controlled environment for this initial investigation. However, replication in more authentic contexts is warranted. The use of physiological sensors are also somewhat limited in terms of scalability. All participants were undergraduate students, and a large proportion (69%) identified as Caucasian – it would be advisable to retrain the models with a more diverse data set to study generalizability to diverse student populations.

In summary, although the results detailed are promising as a first start, there are multiple directions in which this research can be extended. We are working towards expanding our models to include multimodal data such as eye gaze or facial features. It is possible that by including additional modalities we will be able to achieve improved detection rates than by using any single modality. This is, of course, an empirical question that awaits further investigation.

Acknowledgement. This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

References

1. Schooler, J.W., Reichle, E.D., Halpern, D.V.: Zoning out while reading: Evidence for dissociations between experience and metaconsciousness. Think. Seeing Vis. Metacognition Adults Child, 203–226 (2004)
2. Smallwood, J., McSpadden, M., Schooler, J.W.: When attention matters: The curious incident of the wandering mind. Mem. Cognit. 36, 1144–1150 (2008)
3. Smallwood, J., Schooler, J.W.: The restless mind. Psychol. Bull. 132, 946 (2006)
4. Drummond, J., Litman, D.: In the zone: Towards detecting student zoning out using supervised machine learning. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part II. LNCS, vol. 6095, pp. 306–308. Springer, Heidelberg (2010)
5. D'Mello, S., Cobian, J., Hunter, M.: Automatic Gaze-Based Detection of Mind Wandering during Reading
6. Andreassi, J.L.: Psychophysiology: Human behavior and physiological response. Routledge (2000)
7. Smallwood, J., Davies, J.B., Heim, D., Finnigan, F., Sudberry, M., O'Connor, R., Obonsawin, M.: Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. Conscious. Cogn. 13, 657–690 (2004)
8. Feidakis, M., Daradoumis, T., Caballé, S.: Emotion measurement in intelligent tutoring systems: what, when and how to measure. In: 2011 Third International Conference on Intelligent Networking and Collaborative Systems (INCoS), pp. 807–812 (2011)
9. Picard, R.W.: Measuring affect in the wild. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, p. 3. Springer, Heidelberg (2011)
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explor Newsl. 11, 10–18 (2009)
11. Guo, R., Li, S., He, L., Gao, W., Qi, H., Owens, G.: Pervasive and unobtrusive emotion sensing for human mental health. In: 2013 7th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), pp. 436–439 (2013)
12. Wagner, J., Kim, J., André, E.: From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In: IEEE International Conference on Multimedia and Expo, ICME 2005, pp. 940–943 (2005)
13. Setz, C., Arnrich, B., Schumm, J., La Marca, R., Troster, G., Ehlert, U.: Discriminating stress from cognitive load using a wearable EDA device. IEEE Trans. on Inf. Technol. Biomed. 14, 410–417 (2010)