Face Forward: Detecting Mind Wandering from Video During Narrative Film Comprehension

Angela Stewart[1](0000-0002-6004-9266), Nigel Bosch[2](0000-0003-2736-2899), Huili Chen[3](0000-0003-3609-1904), Patrick Donnelly[1](0000-0003-1033-5931), & Sidney D'Mello[1](0000-0003-0347-2807)
[1]University of Notre Dame, Notre Dame, IN, USA
[2]University of Illinois at Urbana-Champaign, Urbana, IL, USA
[3]Massachusetts Institute of Technology, Cambridge, MA, USA
{astewa12, sdmello}@nd.edu

Abstract. Attention is key to effective learning, but mind wandering, a phenomenon in which attention shifts from task-related processing to task-unrelated thoughts, is pervasive across learning tasks. Therefore, intelligent learning environments should benefit from mechanisms to detect and respond to attentional lapses, such as mind wandering. As a step in this direction, we report the development and validation of the first student-independent facial feature-based mind wandering detector. We collected training data in a lab study where participants self-reported when they caught themselves mind wandering over the course of completing a 32.5 minute narrative film comprehension task. We used computer vision techniques to extract facial features and bodily movements from videos. Using supervised learning methods, we were able to detect a mind wandering with an F_1 score of .390, which reflected a 31% improvement over a chance model. We discuss how our mind wandering detector can be used to adapt the learning experience, particularly for online learning contexts.

Keywords: Mind Wandering · Attention Aware Interfaces

1 Introduction

Consider a situation where you are enrolled in an online anthropology course. Every week, you are assigned a documentary film to watch and discuss in an online forum. Your forum posts are graded based on your demonstration of film comprehension and your ability to relate the subject matter to current cultural trends. While watching this week's documentary on linguistics in early American society, you are initially engaged in the film. However, your thoughts inevitably begin to drift away from task-related thoughts to unrelated thoughts about your grocery shopping list for tonight's dinner. Using your computer's webcam, the online educational interface has been monitoring your facial expressions and detects that you are not attending to the content even though you appear to be looking at the screen. The interface pauses the video and asks you a question about the film's content, which you answer incorrectly. Based on this, the interface provides an explanation to reinforce certain concepts that you were not attending to, before asking whether you would like to continue viewing the video. This reengages your attention, leading to a deeper understanding of the course content, and consequently a higher score in the course.

Educational interfaces that detect and respond to attentional states, such as the one described above, are on the horizon in the next 5-10 years [1]. Here, we focus on a specific form of inattention, known as mind wandering (MW). MW is a ubiquitous phenomenon where attention unintentionally shifts from task-related to task-unrelated thoughts. The widespread incidence of MW has been documented during a host of real-world activities. In one highly-cited, large-scale study, MW was tracked in 5,000 individuals from 83 countries working in 86 occupations, using an iPhone app that prompted people to report their thoughts at random intervals throughout the day [2]. People reported MW for 46.9% of the prompts, which confirmed numerous lab studies on the pervasiveness of MW (e.g., [3]), which is estimated to occur approximately 20-50% of the time, depending on the person, task, and the environmental context [2, 4].

In addition to being frequent, MW is also detrimental to performance across a number of tasks, such as reading comprehension [5] and retention of lecture content [6]. In fact, a recent meta-analysis of 88 samples indicated a negative correlation between MW and performance across a variety of tasks [7], a correlation which increased in proportion to task complexity. When compounded with its high frequency, MW can have serious consequences on performance and productivity, particularly in learning environments where attention is key to learning and retaining material. Therefore, we believe that next-generation personalized learning technology could benefit from some mechanism to detect and address MW [1]. Of course, an interface must first detect MW before it can respond to it, which is the focus of this work.

As reviewed below, previous work on MW detection, particularly in educational domains, has mainly focused on reading tasks. Here, we focus on MW detection in the novel context narrative film comprehension. Further, for the first, time we consider automated detection of MW from facial features and bodily movements obtained from commercial-off-the-shelf (COTS) webcams.

Related Work. Attention-aware education interfaces are not a new idea. Real-time analysis of eye gaze has been proposed as a way of monitoring and responding to attention [1]. Considerable work has provided offline methodologies to model attention in educational domains; however, real-time attention detection and response systems are still in their infancy [1]. Most work has been limited to eye gaze analysis. We aim to expand work in the field through the use of automatically extracted facial features.

Most of the work on MW detection has been done in the context of reading. These studies use a variety of features, such as eye-gaze [3, 8], reading times [5], and physiological signals [9]. For example, Bixler and D'Mello used eye gaze to detect both probe-caught [3] and self-caught reports [8] of MW during reading. Probe-caught MW reports required users to indicate if they were MW in response to auditory thought probes triggered at pseudo-random intervals during reading. Self-caught reports were obtained whenever users caught themselves MW. The authors achieved above-chance accuracies of 17% to 45% in detecting MW in a user-independent fashion.

Despite their success, these studies have relied on specialized equipment to collect eye-gaze (Tobii TX300). The prohibitive cost or lack of accessibility of these sensors potentially limits wide-spread adoption outside the context of laboratory settings. To address this, some researchers have considered sensor-free MW detection. Reading times have been particular beneficial in this regard. In one study, reading times for

individual words were tracked using a word-by-word self-paced reading paradigm [5]. Readers were considered to be MW if they spent too little or too much time on difficult sections of the text, as determined by predetermined thresholds on word length, syllables, and word familiarity. Despite success, an obvious limitation with the use of reading time for MW detection is that such a detector is only applicable while reading.

There has been limited work investigating detection of MW during video watching. Pham and Wang use heart rate to detect MW during videos for massively open online courses (MOOCs) with a 22% above-chance accuracy [10]. They detected heart rate by monitoring fingertip transparency using the back camera of an iPhone. While this method makes use of widely-owned equipment (an iPhone in this case), whether this method can be used on non-mobile devices is an open question.

Mills et al. took a different approach to MW detection in narrative film viewing by using eye-gaze features [11]. They used global and local (context-dependent) features, as well as a combination of the two, to build models to detect MW. Their best models yielded a 29% improvement over chance when using only local features. This work demonstrates the feasibility of detecting MW during film viewing tasks. However, the prohibitive cost of the eye-gaze sensors potentially limit widespread adoption of their method for detecting MW.

Contributions and Novelty[1]**.** This study reports the development and validation of the first student-independent facial feature-based MW detector during narrative film comprehension. Our work is novel in two respects. First, while previous work has focused on MW in the context of reading, we consider MW detection during narrative film comprehension. This is a challenging domain, because, compared to reading, where there are detectable patterns that might indicate attentional lapses, such unexpected reading times or failing to advance to the next screen, naturalistic film viewing is less interactive, which provides less context information for detecting MW.

Nevertheless, we chose to study this domain because video-based courses, such as MOOCs, are very popular for a variety of students [6]. Although one previous study [10] focused on MW detection while students viewed MOOC-style videos, our present focus is on commercially-produced narrative films, such as historic documentaries, nature films, and fantasy-drama films that might be assigned in history, sociology, and film appreciation courses, amongst others. We focused on these types of films because professional filmmakers employ a host of cinematic devices to direct and engage viewer attention [4]. Furthermore, films contain both audio and visual content, which would presumably keep attention focused [4]. Despite these efforts to engage the viewer, MW still occurs quite frequently while students watch such films [4] (as well as with typical MOOC-style videos [10]), suggesting that tracking and responding to moments of MW during film viewing could improve online learning from these materials.

Second, previous work has relied on specialized sensors for MW detection, thereby limiting scalability. This work represents the first attempt at a fully automated student-

[1] A preliminary two-page version of this paper was presented as an Extended Abstract Poster at the 24th Conference on User Modeling, Adaptation and Personalization. The present paper describes the methods in more detail, updated results, and expanded analyses not included in the preliminary paper.

independent detection of MW using face videos recorded from COTS webcams. This also raises some challenges because unlike emotional states, where facial correlates have been investigated for decades and video-based automated affect detection is common [12], the facial correlates of MW have yet to be mapped out. It is also an open question if such correlates exist. For example, as Fig. 1 illustrates, facial expressions corresponding to MW reports (left) appear to be highly similar to when MW was not reported (right). Despite these challenges, if successful, our MW detector should be scalable (because it uses webcams) and more broadly applicable to additional contexts (because it does not rely on any features specific to a particular interaction context, like reading times or click-stream analyses).

Our approach to MW detection entailed collecting videos and self-reports of MW while users watched a short film on a computer screen. We used a self-caught method to detect MW in order to avoid the disruptive effects of thought probes. We extracted facial features and bodily movements from the videos and used supervised classification techniques to build models that identified when users were MW across short time windows. The models were constructed and validated in a student-independent fashion so that they would generalize to new students.

2 Data Collection

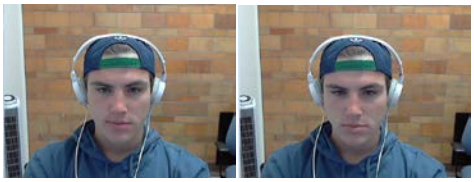

Fig. 1. Video frame of participant corresponding to the presence (left) and absence (right) of MW reports.

Participants were 65 undergraduate students from a medium-sized private Midwestern university and 43 undergraduate students from a large public university in the Southern United States. Of the 108 participants, 66% were female and their average age was 20.1 years. Participants were compensated with course credit. Data from one participant was discarded due to equipment failure.

Participants viewed the narrative film “The Red Balloon” (1956), a 32.5-minute French-language film (with English subtitles). The film has a musical score but only sparse dialogue. This short fantasy film depicts the story of a young Parisian boy who finds a red helium balloon and quickly discovers it has a mind of its own as it follows him wherever he goes. This film was selected because of the low likelihood that participants had previously seen it, and because it has been used in other film comprehension studies [4]. Participants’ faces were recorded while they watched the film with a low-cost ($30) consumer-grade webcam (Logitech C270).

Participants were instructed to report MW throughout the film by pressing labeled keys on the keyboard. Specifically, participants were asked to report a task-unrelated thought if they were “thinking about anything else besides the movie.” Participants were explicitly instructed to report a task-related interference if they were “thinking about the task itself but not the actual content of the movie.” A small beep sounded to register their response, but film play was not interrupted.

It is important to emphasize a couple of points on the self-caught method used to track MW. First, we chose to have participants self-report when they caught themselves MW instead of the more traditional probing method [3] because the probe method has the potential to interrupt the comprehension process (i.e., when participants are not MW and report "no" to the probes) [13]. This is particularly problematic as participants did not have control over the film presentation (i.e., no pausing or rewinding capabilities were available). Additionally, self-caught reports, as opposed to probe-caught reports, are likely to occur at the end of a MW episode when the student became aware that they were not attending to the task at hand. It is unclear, however, if a probe-caught report takes place at the onset or end of MW, or somewhere in between. Furthermore, although the method relies on self-reports, there is no clear alternative because MW is an internal phenomenon. Nevertheless, self-reported MW has been linked to predictable patterns in eye-gaze [14] and task performance [7], providing validity for this approach.

We obtained a total of 845 MW reports from the 108 participants. In this initial work, we do not distinguish between the two types of MW, instead merging the task-unrelated thoughts and the task-related interferences, both of which represent thoughts independent of the content of the film.

3 Machine Learning

Creating Instances of MW. MW reports were sparsely distributed throughout the 32.5 minute video. Our first task was to create data instances corresponding to short windows of time preceding MW reports. To ensure that we captured participants' faces while MW and not the act of reporting MW itself (i.e., the preparation and execution of the key press), we added a 3-second offset before each MW self-report. From observing participant videos, this appeared to be sufficient time to prevent detection of the key press. We chose not to use larger offsets because it is not known how long MW lasts and we aimed to avoid removing data from windows where the participant was MW prior to the report.

The next task was to extract instances corresponding to Not MW while ensuring a gap between the MW and Not MW instances to account for the fact that we do not know precisely when MW begins.

The procedure for creating instances was as follows:

1. Add a 3-second offset before the self-caught MW report to account for movement due to reporting.
2. Partition the video between consecutive MW reports into $(t_1 - t_0) / S$ segments, where t_0 and t_1 are the timestamps of consecutive MW reports and S is the segment size. The segment immediately preceding the MW report at t_1 is a MW segment. All other segments between t_0 and t_1 are Not MW segments.
3. Extract features from a window of data of size w, where $w < S$, at the end of each segment generated in the previous step. The remaining time ($S - w$ seconds) in the segment is the gap that is not analyzed.

In this study, we chose a 55 second segment length as it resulted in a MW rate of approximately 20% to 25%, which was consistent with previous research [3]. We explored various windows sizes within the 55-second segment (Section 4). The procedure described above is depicted in Fig. 2 using a 45-second windows as an example.

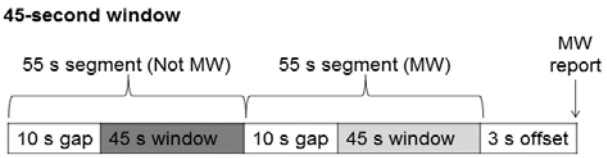

Fig. 2. Example of window segmentation approach, using a 45-second widow sizes. Features are extracted from the dark grey (Not MW) and light grey (MW) windows.

We generated a total of 3,370 segments in all. We excluded any instances in which the participants' face was occluded, yielding less than one second of data for the time window. Extreme head pose or position, occlusions from hand-to-face gestures, and rapid movements were common causes of face registration errors. We also experimented with various window sizes. The number of instances (after removing instances with too little valid data) varied as a function of window size (from 2,476 for 10 second windows to 2,734 for 45 second windows). Larger window sizes contained more instances because there was a higher probability that the face was registered for at least one second. MW rates were quite similar across window sizes although there was a slight increase for the longer windows (from .204 for 10 second windows to .221 for 45 second windows).

Feature Extraction and Selection. We used FACET [15], a commercialized version of the CERT computer vision software for facial feature extraction. FACET provides likelihood estimates of the presence of 19 action units (AUs; specifically 1, 2, 4, 5, 6 ,7, 9, 10, 12, 14, 15, 17, 18, 20, 23, 24, 25, 26, and 28 [16]) as well as head pose (orientation), face position (horizontal and vertical within the frame), and face size (a proxy for distance to camera). Features were created by aggregating FACET estimates in a window of time leading up to each MW or Not MW instance using maximum, median, and standard deviation for aggregation. In all, there were 75 facial features (3 aggregation functions × [19 AUs + 3 head pose orientation axes + 2 face position coordinates + face size]).

We also computed gross body movement present in the videos using a validated motion estimation algorithm [17]. Body movement was calculated by measuring the proportion of pixels in each video frame that differed by a threshold from a continuously updated estimate of the background image generated from the four previous frames. We used the maximum, median, and standard deviation of gross body movement within each window, similar to the method used to compute FACET features.

In all, we extracted 78 features (75 facial features + 3 body movement features). We treated outliers, defined as values greater than three standard deviations away from the mean, with Winsorization, a common outlier handling technique [18]. This technique replaces outliers with the closest non-outlier value, allowing the retention of instances with outliers rather than discarding the entire instance.

We used tolerance analysis to eliminate features with high multicollinearity (variance inflation factor > 5) [19], after which 59 features remained. This was followed by RELIEF-F [20] feature selection (on the training data only) to rank features. Feature selection was performed using nested cross-validation on training data only. In particular, we ran 5 iterations of feature selection within each leave-one-participant-out cross-validation fold (discussed below), using data from a random 67% of students within the training set in each iteration. We retained a proportion of the highest ranked features (with rankings averaged across folds) for use in the models (proportions ranging from .05 to 1.0 were tested).

Classifier Selection and Validation. Informed by preliminary experiments, we selected nine classifiers for more tests (Naïve Bayes, Support Vector Machines, Simple Logistic Regression, LogitBoost, Random Forest, C4.5 trees, Stochastic Gradient Descent, Classification via Regression, and Bayes Net) using the WEKA toolkit [21].

We evaluated the performance of our classifiers using leave-one-participant-out cross-validation. This process runs multiple iterations of each classifier in which, for each fold, the instances pertaining to a single participant are added to the test set and the training set is comprised of the instances for the other participants. This process is repeated for each participant, and the classifications of all folds are weighted equally to produce the overall result. This cross-validation approach ensures that in each fold, data from the same participant is in the training set or testing set but never both, thereby improving generalization to new participants.

We considered the F_1 score for the MW class as our key accuracy measure as MW is the minority class of interest (compared to Not MW). Further, F_1 strikes a balance between precision and recall, and is less susceptible to skew from class imbalance (which is present in the current dataset) than simply measuring recognition rate.

4 Analyses and Results

Varying Window Size. We experimented with window sizes from 10 through 45 seconds in intervals of 5 seconds to empirically identify the window size that yielded the highest MW F_1. For the support vector machine (SVM) classifier (the most effective classifier – see below), there was a slight trend in performance of MW F_1 score in favor of larger window size (from .355 for 10-second windows to .390 for 45-second windows).. Therefore, all subsequent results focus on the 45-second window size.

Overall Classification Results. The results for the highest MW F_1 model were achieved with an SVM classifier using sequential minimal optimization (SMO) [22] on a data set with 45-second windows where the SMOTE technique [23] was used (on training data only). This model classified 45.1% of the instances as MW. We compared it to a chance (baseline) model that also assigned MW to 45.1% of all instances, but did so randomly. This process was repeated for 1,000 iterations and precision and recall were averaged across iterations. This chance-level method yielded a precision of .221 (i.e. the same as the MW base rate) and recall of .451 (i.e. the same as the predicted MW rate). We believe this chance model to offer a more appropriate comparison than a simple minority baseline that assigns MW to all instances, because a minority baseline

would results in an inflated recall (MW precision = .221, MW recall = 1, MW F_1 = .362). Additionally, a majority class baseline would result in a MW F_1 of 0, which is trivial to surpass.

Table 1 shows the results of the SVM classifier compared to the chance model. The key metrics are the precision, recall, and F_1 of the MW class. For completeness, we also provide results for Not MW class and a weighted average of the two (Overall).

Table 1. Results of the SVM classifier with chance values in parentheses

	Precision	**Recall**	**F_1MW**
MW	**.290 (.221)**	**.593 (.451)**	**.390 (.297)**
Not MW	.836 (.779)	.589 (.549)	.691 (.644)
Weighted Overall	.715 (.656)	.590 (.527)	.624 (.567)

The key result is that the SVM model detected MW at rates that were substantially (31%) greater than the chance model. The SVM model's recall was also double its precision. The model has a similar proportion of hits (.593) and correct rejections (.589). Similarly, we note the model makes the same proportions of misses (.407) and false positive (.411) errors. However, the effect of false positives are exemplified as the model predicts a much higher rate of MW (.451) than the true rate (.221).

Analysis of MW Threshold. SVMs provide an estimate of the model's confidence (on a 0 to 1 scale) that an instance reflects MW. This estimate needs to be converted into a binary decision. In the aforementioned results, any instance that exceeded a confidence of .500 was classified as MW. To determine the optimal threshold that would result in the highest MW F_1, we adjusted the threshold in increments of .100 and computed resultant F_1 scores for MW and Not MW classes (Fig. 3). We note that the MW and Not MW curves in Fig. 3 intersect at a threshold of .370, yielding an approximate equal F_1 scores of .380. However, the MW F_1 score, which is our primary metric of interest, peaked at a threshold of .500, which suggests that the default threshold was appropriate for this task.

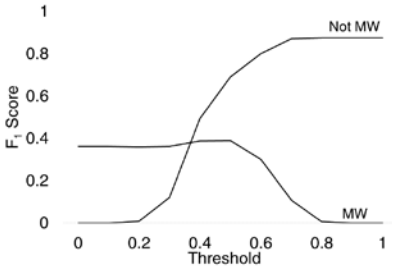

Fig. 3. F_1 scores for MW and Not MW across classification thresholds.

Feature Analysis. We examined the features used in the SVM model, focusing on the nine features most commonly selected by the RELIEF-F procedure as described in Section 3. Features were analyzed using Cohen's *d*, which measures the effect size of the difference of each feature across MW and Not MW instances divided by the pooled standard deviation [24]. Positive *d*-values for a feature indicate an increase in the value of that feature for MW compared to Not MW. We note that effect sizes for most of the features were in the small (d = .200) to medium (d = .500) range [24], suggesting that no one feature dominated, but a combination of features was needed for MW detection.

With respect to specific features, the median of the face's vertical position (d = .354) and size in the screen (d = .272) were quite predictive of MW. These two features suggest the participant was higher in the frame and closer to the screen. This could be due to participants dozing off and nodding their head when they were MW (based on examination of videos). With the exception of the median of AU7 (lid tightener, d = .169) and AU28 (lip suck, d = .010), there was less activity in facial features, such as the median of AU5 (upper eye lid raiser, d = -.305), AU10 (upper lip raiser, d = -.193), AU17 (chin raiser, d = -.224), and AU18 (lip puckerer, d = -.131), and the max value AU9 (nose wrinkle, d = -.203). This indicates that participants adopted more neutral facial expressions when MW, ostensibly because they were not reacting to the unfolding film as thoughts were concentrated inwards.

5 General Discussion

Main Findings. We expanded on previous MW research through our novel use of facial expressions and body movements to detect MW. We were able to detect MW with an F_1 score of .390, a notable 31% improvement over a chance-level model, which yielded an F_1 score of .297. Although we showed that the default threshold of .500 resulted the highest MW F_1, our model had higher recall (.593) than precision (.290), which suggests that it over predicts MW (i.e., more false positives). We should also note that it is possible that the self-reports underestimate the MW rate, either because participants choose not to report MW or because they are unaware that they are MW. Perhaps the truth lies somewhere between the self-reports and computer-estimates of MW.

Our model, shown in Table 1 achieved a 31% improvement over a chance classifier. As a point of comparison, Mills et. al. [11] achieved a comparable accuracy of 29% over chance. Note that both chance-level classifiers were computed using the same measure of chance (Section 4) and the method for partitioning data into MW and Not MW instances (Section 3) was similar, thus providing a basis for comparison. While the accuracy of models in both works is moderate, the improvement over chance demonstrates the feasibility of detecting MW from either eye gaze or facial features. However, Mills et. al. used research-grade eye trackers, which might prohibit widespread use of their method, particularly in online educational interfaces where students provide their own equipment. Additionally, their best models used content-dependent features, which might not easily generalize to new stimuli. Thus, our results are significant in that we were able to obtain results similar to the previous state of the art on this dataset, but by using a more scalable (and presumably generalizable) modality.

Generalizability was a key design constraint that guided a number of our decisions. First, we used COTS webcams to afford eventual deployment of our models at scale, thereby allowing us to test generalizability in more diverse contexts. We also restricted ourselves to facial features and bodily movements that were independent of the specific content being displayed on the screen, suggesting that the models should generalize to additional films and perhaps even other interaction contexts. Further, our models were validated in a student-independent manner, which increases our models' ability to generalize to new students. We have even more confidence in the generalization of our

models as our data was collected from two universities with very different demographic characteristics. Taken together, these results increase our confidence that the models will generalize more broadly, though this claim requires further empirical validation.

Finally, given the paucity of research, it was unclear if MW manifests via facial expressions. It was therefore quite possible that our entire research endeavor would fail. Fortunately, our findings do indicate that there appear to be generalizable patterns between facial-expressions and MW. Specifically, we found that MW was characterized by vertical head movement and more neutral facial expressions.

Applications. The present findings are applicable to any user interface that involves viewing and comprehending videos. Monitoring MW in this context could greatly inform commercial or educational film makers as to how their films can be improved to better sustain viewers' attention. Segments of film with high rates of detected MW can be edited to better engage viewers.

Media, such as films and recordings of lectures, play a major role in online learning, so our MW detector, which only uses a webcam, can be quite beneficial in that context. One strategy is to assess comprehension of content associated with periods of high MW (as noted by the detector) by asking the student to answer a multiple-choice question or to self-explain the content. Both interleaved questions [25] and self-explanations [26] have been shown to be effective at focusing attention. Students who answer incorrectly will be encouraged to review the material associated with the questions and self-explanation prompts, and can optionally answer follow-up questions, thereby giving them multiple opportunities to correct comprehension deficits attributed to MW.

Our work also has applications in contexts apart from viewing videos. MW has been widely studied during reading using a variety of sensors [3, 8, 9, 14] but not facial features. Facial feature data could supplement existing features to improve MW detection. This also raises the possibility of multimodal MW detection.

Limitations and Future Work. There are a number of limitations with this study. First, our model had a MW F_1 of .390. Although it outperformed a chance model, this performance is moderate at best. The precision was also much lower than the recall, suggesting that caution should be taken when integrating the model into interfaces that sense and respond to MW. In future work, we will aim to improve precision by expanding the feature set and considering skew-sensitive classification methods.

Another limitation of this study is the self-report method which requires users to be mindful of when MW occurs and to respond accurately and honestly. Previous studies have validated the self-report method [7, 14], however, it is possible that some participants may not report MW accurately or honestly. One possibility would be to complement self-reports with observer annotations. However, this assumes that observers can identify when a person is MW, a question that we are investigating in our research.

Finally, although we provided some evidence of generalizability to new users, to further boost our claims of generalizability, data should be collected from more diverse populations apart from undergraduate students. It should also be collected more real-world environments, rather than the lab-setup used here. Generalizability could also be enhanced by studying video-based MW detection in other contexts such as playing interactive games, to better understand how the models generalize to other tasks. Training models on data from multiple domains is also likely to yield more general models.

Concluding Remarks. The ubiquity of webcams has opened up the possibility of advancing research in attentional state estimation, thereby enabling an entirely new generation of attention-aware interfaces, particularly in education. As a step in this direction, we demonstrated the feasibility of using facial features extracted from webcam video to record MW during a narrative film comprehension task. The next step is to close the loop by intervening when MW is detected.

6 Acknowledgements

This research was supported by the National Science Foundation (NSF) (DRL 1235958 and IIS 1523091). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF

7 References

1. D'Mello, S.K.: Giving Eyesight to the Blind: Towards Attention-Aware AIED. International Journal of Artificial Intelligence in Education. 1–15 (2016).
2. Killingsworth, M.A., Gilbert, D.T.: A wandering mind is an unhappy mind. Science. 330, 932–932 (2010).
3. Bixler, R., D'Mello, S.K.: Toward fully automated person-independent detection of mind wandering. In: Proceedings of the 22nd International Conference on User Modeling, Adaptation, and Personalization. pp. 37–48. Springer International Publishing, Switzerland (2014).
4. Kopp, K., Mills, C., D'Mello, S.K.: Mind wandering during film comprehension: The role of prior knowledge and situational interest. Psychonomic Bulletin & Review. 23, 842–848 (2015).
5. Franklin, M.S., Smallwood, J., Schooler, J.W.: Catching the mind in flight: using behavioral indices to detect mindless reading in real time. Psychon Bull Rev. 18, 992–997 (2011).
6. Szpunar, K.K., Moulton, S.T., Schacter, D.L.: Mind wandering and education: from the classroom to online learning. Frontiers in Psychology. 4, (2013).
7. Randall, J.G., Oswald, F.L., Beier, M.E.: Mind-Wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. Psychological bulletin. 140, 1411 (2014).
8. Bixler, R., D'Mello, S.K.: Automatic Gaze-Based Detection of Mind Wandering with Metacognitive Awareness. In: Ricci, F., Bontcheva, K., Conlan, O., and Lawless, S. (eds.) User Modeling, Adaptation and Personalization: 23rd International Conference. pp. 31–43. Springer International Publishing, Dublin, Ireland (2015).
9. Blanchard, N., Bixler, R., Joyce, T., D'Mello, S.K.: Automated physiological-based detection of mind wandering during learning. In: Intelligent Tutoring Systems. pp. 55–60. Springer, Honolulu, Hawaii, USA (2014).
10. Pham, P., Wang, J.: AttentiveLearner: Improving Mobile MOOC Learning via Implicit Heart Rate Tracking. In: Conati, C., Heffernan, N., Mitrovic, A., and

Verdejo, M.F. (eds.) Artificial Intelligence in Education. pp. 367–376. Springer International Publishing, Cham (2015).
11. Mills, C., Bixler, R., Wang, X., D'Mello, S.K.: Automatic Gaze-Based Detection of Mind Wandering during Film Viewing. In: Proceedings of the 9th International Conference on Educational Data Mining. International Educational Data Mining Society, Raleigh, NC, USA (2016).
12. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence. 31, 39–58 (2009).
13. Faber, M., Bixler, R., D'Mello, S.K.: An automated behavioral measure of mind wandering during computerized reading. Behavior Research Methods. 1–17 (2017).
14. Reichle, E.D., Reineberg, A.E., Schooler, J.W.: Eye movements during mindless reading. Psychological Science. 21, 1300–1310 (2010).
15. Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M.: The computer expression recognition toolbox (CERT). In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011). pp. 298–305. IEEE (2011).
16. Ekman, P., Friesen, W.V.: Facial action coding system. (1977).
17. Westlund, J.K., D'Mello, S.K., Olney, A.M.: Motion Tracker: Camera-based monitoring of bodily movements using motion silhouettes. PLoS ONE. 10, (2015).
18. Dixon, W.J., Yuen, K.K.: Trimming and winsorization: A review. Statistische Hefte. 15, 157–170 (1974).
19. Allison, P.D.: Multiple regression: A primer. Pine Forge Press (1999).
20. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Machine Learning: ECML-94. pp. 171–182. Springer (1994).
21. Holmes, G., Donkin, A., Witten, I.H.: Weka: A machine learning workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems. pp. 357–361. IEEE (1994).
22. Platt, J., others: Sequential minimal optimization: A fast algorithm for training support vector machines. (1998).
23. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 321–357 (2002).
24. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. 2nd edn. Hillsdale, New Jersey: L. Erlbaum (1988).
25. Szpunar, K.K., Khan, N.Y., Schacter, D.L.: Interpolated memory tests reduce mind wandering and improve learning of online lectures. Proceedings of the National Academy of Sciences. 110, 6313–6317 (2013).
26. Moss, J., Schunn, C.D., Schneider, W., McNamara, D.S.: The nature of mind wandering during reading varies with the cognitive control demands of the reading strategy. Brain research. 1539, 48–60 (2013).