Automatic Gaze-Based Detection of Mind Wandering during Narrative Film Comprehension

Caitlin Mills[a], Robert Bixler[a], Xinyi Wang, & Sidney K. D'Mello
University of Notre Dame
384 Fitzpatrick Hall, Notre Dame, IN, 46556, USA
[cmills4, rbixler, xwang24, sdmello]@nd.edu

ABSTRACT
Mind wandering (MW) reflects a shift in attention from task-related to task-unrelated thoughts. It is negatively related to performance across a range of tasks, suggesting the importance of detecting and responding to MW in real-time. Currently, there is a paucity of research on MW detection in contexts other than reading. We addressed this gap by using eye gaze to automatically detect MW during narrative film comprehension, an activity that is used across a range of learning environments. In the current study, students self-reported MW as they watched a 32.5-minute commercial film. Students' eye gaze was recorded with an eye tracker. Supervised machine learning models were used to detect MW using global (content-independent), local (content-dependent), and combined global+local features. We achieved a student-independent score (MW F_1) of .45, which reflected a 29% improvement over a chance baseline. Models built using local features were more accurate than the global and combined models. An analysis of diagnostic features revealed that MW primarily manifested as a breakdown in attentional synchrony between eye gaze and visually salient areas of the screen. We consider limitations, applications, and refinements of the MW detector.

Keywords
mind wandering; film comprehension; machine learning; eye gaze

1. INTRODUCTION
Mind wandering (MW) reflects an attentional shift from task-related to task-unrelated thoughts [31]. MW is estimated to consume half of our everyday thoughts [19] and can occur at almost any time – driving down the road, eating a meal, or during a classroom lecture. There are some benefits to our innate ability to MW, specifically with respect to planning and creativity [34]. However, MW has some detrimental effects as well, particularly in the realm of education [30]. A recent meta-analysis across 88 independent samples indicated that MW was negatively correlated with performance, and that the negative relationship was stronger for more complex tasks such as reading comprehension [26]. Given the negative impact of MW on learning [29, 30], it is important to develop attention-aware systems that can reorient attention when MW occurs [8]. However, these systems require reliable MW detection, which is the focus of this work.

MW detection can be particularly challenging since MW is an internal state with few overt markers (unlike some emotions per se). It can even be difficult for people to realize when they are MW, as it can occur without metacognitive awareness [30]. Moreover, the onset and duration of MW cannot be clearly demarcated as with other disengaged behaviors, such as gaming the system or WTF (Without Thinking Fastidiously) behaviors [1, 25].

In the present study, we focus on detecting MW in the novel educational context of narrative film comprehension – a more complex task than self-paced reading where most MW detection efforts have focused on. We chose this task for two reasons. First, a large number of students from all over the world watch educationally relevant films and recorded lectures daily, particularly in the advent of massive open online courses (MOOCs). Second, MW is quite frequent in online video lectures: students report MW around 40% of the time while viewing lectures [29, 33], so there is considerable promise to detecting and responding to MW in this context.

1.1 Background and Related Work
Only one study (to our knowledge) has attempted MW detection while students viewed dynamic visual scenes, such as the narrative film we consider here. Pham and Wang [25] detected MW while students watched video lectures on a smart phone with a MOOC-like application and responded yes or no to thought probes during the lectures. They used student heart rate (extracted via photoplethysmography) to train classifiers to detect MW. They achieved a 22% greater than chance detection accuracy, thereby providing some initial evidence that MW detection is feasible in this context.

Aside from [25], other MW detection efforts have been limited to self-paced reading. In one of the first MW detection studies [10], students read aloud and then paraphrased biology paragraphs. They were periodically asked to report zone outs during reading on a 1 (all the time) to 7 (not at all) scale. Supervised machine learning models trained on acoustic-prosodic features to classify between "high" (1-3 on the scale) versus "low" zone outs (5-7 on the scale) achieved a 64% accuracy. However, this study did not adopt a student-independent validation approach, so it is unclear how well their detector would generalize to new students.

Other research has utilized log-file information to detect MW during self-paced reading. In one study [23], MW reports were collected via pseudo-random thought probes during self-paced computerized reading. Students responded either "yes" or "no" about whether they were MW at the time of the probe. Using textual features and reading behaviors from log-files, supervised machine learning models were able to detect MW with a 21% above-chance accuracy. Similarly, [12] attempted to predict MW during reading using textual features (e.g., difficulty, familiarity, and reading time), but it is not clear if their method, which utilized researcher-pre-defined thresholds, would generalize more broadly.

[a] Denotes equal contribution by authors.

Researchers have also adopted sensor-based approaches for MW detection during reading. Blanchard et al. [4] used an Affectiva Q sensor to record both galvanic skin response and skin temperature while participants read texts on research methods and periodically provided MW reports in response to thought probes. Their models attained a kappa value of .22 using a combination of peripheral physiology and contextual features (e.g., page numbers).

Eye gaze is perhaps one of the most promising modalities for MW detection due to the so called eye-mind link [27], which posits a coupling between eye movements and attentional focus. Several studies have thus built MW detectors using eye gaze features. The first study collected data from 84 students during self-paced reading of four texts on research methods [7]. MW reports were collected in response to thought probes triggered when gaze was fixated on predefined words on the screen. Supervised classification models were built from 27 gaze features and validated in a student-independent fashion. The authors achieved an accuracy of 60% after downsampling the data. Since downsampling was applied to both the training and test sets, it is unclear how the models would perform when presented with data that reflected the original skewed class distributions.

Their work was extended using a larger dataset of 178 students from two different universities and a wider array of 80 features, including blink and pupil features [2]. Students also read four texts on research methods, and MW reports were collected in response to nine pseudorandom probes that occurred between four to twelve seconds from the beginning of a page of text. Supervised models were built using an extended feature set and were cross-validated in a student-independent fashion. The models achieved an accuracy of 72% (31% above chance) when validated with a test set that maintained the original class distributions. Further, in [2], the authors provided evidence for the predictive validity of the model by showing that it predicted posttest scores at rates higher than self-reported MW, even after controlling for prior knowledge.

The results from this study indicate that MW can be detected from eye gaze during self-paced reading with moderate accuracy. However, there is an open question about the use of eye gaze to detect MW in additional contexts– in particular, for more complex stimuli like dynamic visual scenes. One study [35] provided evidence that eye movements can be predictive of attention while viewing short video clips. In this study, participants watched video clips in two different conditions: (1) without any distractions (attending) and (2) while performing a mental calculation (not attending). Results indicated that eye movements toward pre-determined salient locations in the scene could identify the watching condition (attending vs. not attending) with a 80.6% accuracy, albeit this is not quite the same as MW detection.

We should note that there is still some debate whether eye movements can be driven by salient features of the stimulus (*exogenous* control) or through conscious control (referred to as *endogenous* control). There is some research to suggest that eye movements are primarily driven by exogenous control. For example, previous research has shown that different viewers tend to fixate on the same locations [24], a phenomenon known as *attentional synchronicity*, which suggests exogenous control. However, other research pointed out that interesting objects are often the most visually salient [11]. Thus, it is possible that viewers fixate on the same locations because of top-down processes (endogenous control), as opposed to simply looking at what is salient. Additional evidence for endogenous control comes from a study which found that task instructions can have an effect on eye movements while viewing dynamic visual scenes [32]. The researchers found that participants looked at more peripheral and less visually salient areas of the scene when instructed in order to determine where the visual scenes were derived from compared to a general viewing task. Thus, eye movements related to endogenous control might be particularly revealing about MW. The current study utilizes this idea to compute features that capitalize on the relationship between eye movements and visually salient regions in the film.

1.2 Current Study and Novelty

In this paper we present one of the first attempts to automatically detect MW during narrative film viewing in a manner that generalizes to new students. We leverage what has been learned in previous work using eye gaze to detect MW during reading, while also developing theoretically-grounded features to improve detection accuracy in this novel context.

MW detection during film viewing poses unique challenges compared to reading, which has been the most common context for MW detection thus far. For one, eye movements are much more predictable during reading since the words on the screen are static. In addition, reading consists of fixations (periods where the gaze position is relatively stable) and saccades (rapid movements between fixations), while the dynamic nature of film also yields smooth pursuits (eye movements that follow a moving stimulus).

Second, the film played continuously without any clear breaks, presenting an additional challenge for MW detection. This is in contrast to reading tasks, which are segmented by page breaks. Thus, a novel method was devised to segment eye gaze data into instances for classification.

Finally, the dynamic nature of film allowed for novel content-dependent features that can be computed from dynamic areas of interest (AOI). Unlike reading, AOIs are particularly meaningful in a film viewing context because of the distinctive visual content areas that dynamically change throughout a film. In this study, AOIs were computed from both plot-related and visually salient regions.

2. DATA COLLECTION

This study utilized a subset of data reported by Kopp el. [21].

2.1 Participants

Eye gaze data was collected for 60 undergraduate students from a private Midwestern university. Students were 20.1 years old on average and 66% of the students were female.

2.2 Materials

Students watched "The Red Balloon," a 32.5 minute French film with few English subtitles (9 in all). The film was displayed on a computer screen with a resolution of 1920 × 1080. The film depicts the story of a young boy and a red balloon that follows him and can inexplicably move on its own. This film was chosen because it is unlikely that many students had previously seen it, which could have affected their propensity to mind wander. The film has also been used in previous film comprehension studies [36].

All data were collected using a Tobii TX 300 eye tracker that was attached to the bottom of the monitor. Eye gaze was recorded with a sampling frequency of 120 Hz for the first 14 participants (due to experimenter error), after which the sampling frequency was adjusted to 300 Hz. This difference was taken into account when filtering the gaze data as discussed below.

2.3 Mind Wandering Reports

Students were asked to self-report MW while they watched the film by pressing labeled keys on a standard keyboard. A short beep sounded to register their response, but the film was not otherwise interrupted. A self-caught MW report method was chosen as opposed to a probe-caught report method (where students are probed to report MW at pseudo-random intervals) in order to minimize disruption, which was critical as the film played without interruption.

Students were asked to differentiate between two different types of MW using separate keys: either task-unrelated thoughts (thoughts completely unrelated to the film such as upcoming vacation plans) or task-related interferences (thoughts related to the task but not the content of the film, such as "*This film is boring"*). For the present analyses, both task-unrelated thoughts and task-related interference were grouped as MW. There was a total of 616 MW reports. On average, students reported 10.3 instances of MW during the film (*SD* = 7.9l; *Min* = 1; *Max* = 31).

2.4 Procedure

Students were asked to sit comfortably at a desk in front of the monitor before beginning the eye-tracker calibration process. There were no restrictions on head movements, making the film viewing experience more ecologically valid than if a headrest was used. Students were randomly assigned to one of two conditions before the film started: in one condition, they read a short story explaining the movie plot [22] while students in the second condition read an unrelated baseball-themed story [1]. The experimental manipulations were part of a larger study and are not used here (more details can be found in [21]). Finally, students were given instructions for how to report MW and then the film began. Students completed a multiple choice comprehension assessment after viewing the film, but this data is not analyzed here.

3. MODEL BUILDING

3.1 Eye Movement Detection

Eye gaze was converted to eye movements (fixations, saccades, smooth pursuits, etc.) in order to filter out some of the inherent noise in raw eye gaze data. We first averaged the raw data from the right and left eyes. A simple moving average filter was then applied to the gaze points in order to smooth the signal while retaining the same sampling frequency. The filter used a window size of five samples for the 120 Hz data and seven samples for the 300 Hz data.

Eye movements were detected using a velocity based algorithm [18, 20]. These algorithms generally use thresholds to classify gaze points as fixations, saccades, or smooth pursuits. The algorithm first classified gaze points with a velocity greater than 110 degrees of visual angle/s as saccades. It then classified gaze points with a velocity lower than five degrees of visual angle/s as fixations. Any remaining gaze points were classified as smooth pursuits. The visual angle thresholds used were based on previous research [17].

3.2 Film Segmentation

Next, we segmented the continuous stream of eye gaze data into MW and non-MW segments. Each segment had three components: gap, window, and offset (see Figure 1). The *gap* was the number of seconds between adjacent segments and could be adjusted to change the ratio of MW to non-MW segments. The *window* was the portion of the segment used to compute features. The *offset* was the number of seconds between the MW report (the moment when the student pressed the key on the keyboard) and the end of the window. An offset was used in order to discard data affected by the student's motion to press the key when reporting MW. An offset size of three seconds was deemed appropriate based on observation of recorded videos.

The process began by creating a MW segment prior to each MW report (segment 2 in Figure 1). The data prior to the MW segment were then considered to be non-MW segments (segment 1) after accounting for the gap. There was no offset for non-MW segments as no key presses were involved.

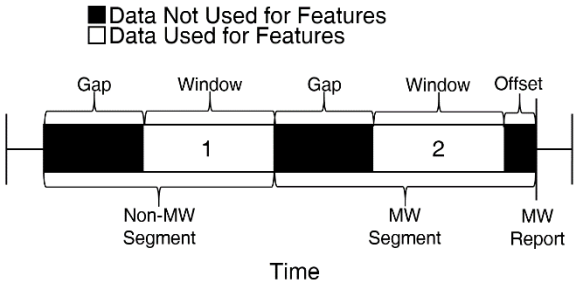

Figure 1. **Hypothetical example of segmented data**

There were several considerations when choosing the window and gap sizes. The segment size (sum of the window, offset, and gap sizes) determined both the number of available instances (segments) and the MW rate as shown in Table 1. Models were built with segment sizes of 45, 55, and 65 seconds, resulting in MW rates that ranged from .256 to .323 and number of instances from 2401 to 1626, thereby allowing us to explore how these two factors affected classification accuracy. For each of these segment sizes, the window size was also varied. In all, we considered window sizes of 10, 15, 20, and 25 seconds.

Table 1. Effect of segment size on number of segments and MW rate

Seg. Size (secs)	Number of Segs.	MW Rate
45	2401	.256
55	1931	.297
65	1626	.323

3.3 Feature Engineering

A total of 143 features were computed from the window in each segment. We considered global features, which were independent of the film content, and local features, which were content specific.

3.3.1 Global Features

There were 88 total global features. Of these, 75 were computed from measures of the eye movements, including fixations, saccades, and smooth pursuits, as well as blinks and pupil diameter. Fixation features were computed from the fixation *durations* (ms). Saccade features were computed from the *saccade durations* (ms), *amplitudes* (degrees of visual angle), *velocities* (degrees of visual angle/s), *relative angle* (degrees of visual angle between two consecutive saccades), and *absolute angle* (degrees of visual angle between a saccade and the x-axis). Smooth pursuit features were computed from the *duration* (ms), *length* (degrees of visual angle), and *velocity* (degrees of visual angle/s) of smooth pursuits. The following descriptive statistics of the distributions were used as the features: minimum, maximum, mean, median,

standard deviation, skew, kurtosis, and range. Counts of each eye movement type were also included as features.

Eight global features were obtained from pupil diameters, which were first z-score standardized at the student-level. The minimum, maximum, median, standard deviation, skew, kurtosis, and range were computed for the standardized pupil diameter distributions from each window and used as features.

There were five additional global features: blink count, mean blink duration, the ratio of total fixation duration to total saccade duration, the proportion of horizontal saccades, and the fixation dispersion.

3.3.2 Local Features

We identified two types of areas of interest (AOIs), Red Balloon AOIs and Visual Saliency AOIs, and computed features based on the locations of the AOIs in each frame. Red Balloon AOIs were used because the red balloon is one of the main objects in the film and endogenous attentional control might direct students to focus on these AOIs despite competing content. OpenCV [4], an open source computer vision software library, was used to isolate the red balloon from the rest of the image using a red color mask. A bounding box was drawn around a contour of the resultant image for each frame in which the balloon appeared (as shown on the left in Figure 2). Local features related to the red balloon were only computed for frames where it was present (58.2% of frames).

We manually examined each frame to ensure that the AOIs were computed correctly. The red balloon was present in 27,262 out of the 46,851 frames. An AOI was constructed for 26,925 of those frames, yielding an accuracy of 98.7%. The frames where the red balloon was missed could be attributed to lighting conditions (making the red balloon appear darker and thus difficult to distinguish from other parts of the scene), the small size of the red balloon, or the majority of the red balloon being off screen or occluded. These frames were left untouched. An additional 8 frames incorrectly had an AOI around an object that was not the red balloon. The AOI was simply deleted from these frames.

Visual Saliency AOIs were used because visually salient areas are known to attract eye gaze [11]. Although, the visual saliency and red balloon AOIs overlap in some cases, as in Figure 2, the visual saliency AOI can be computed for frames without the red balloon. The MATLAB implementation of a Graph-Based Visual Saliency algorithm [16] was used to produce a visual saliency map for each frame based on color, intensity, orientation, contrast, and movement. An area of no more than 2,000 pixels (1.1% of the screen area) surrounding the most salient point were retained and the remaining pixels were set to an intensity of 0. Similar to above, a bounding box was drawn around the largest contour of the processed image.

Local features were computed based on the relationship between the AOIs and each type of eye movement. The features included: (1) *AOI distance*, (2) *AOI intersection*, and (3) *saccade landing*. There were 32 AOI distance features, which captured the distance between the AOI and gaze positions. AOI distance features were computed as the distance between each fixation point or smooth pursuit point and the center of the AOI for each frame in the window. Fixation points were generated for each frame at the centroid of the fixation. Smooth pursuit points were generated for each frame using linear interpolation from the onset to the offset of each smooth pursuit. The minimum, maximum, mean, median, standard deviation, skew, kurtosis, and range of the measured distances were then computed for each eye movement, resulting in 16 features for each type of AOI (32 in all).

There were 12 additional AOI intersection features. These were calculated as the proportion of frames in which a fixation or smooth pursuit point was within the AOI bounding box. Four of these features used the original dimensions of the AOI bounding box. An additional eight used a bounding box expanded by either one or two degrees of visual angle in order to account for inaccurate eye gaze or cases where the AOI was small in size.

Figure 2. An example frame with a bounding box around contours of the red balloon (left) and most visually salient region (right)

Finally, there were 12 saccade landing features. For each AOI, there was a single feature that captured the number of saccades onto, away from, or within the AOI bounding box, which resulted in six features (3 per AOI). An additional six features were computed using a bounding box expanded by one degree of the visual angle to accommodate gaze tracking errors or small AOIs.

In all, there were 56 local features (32 AOI distance, 12 AOI intersection, and 12 saccade landing).

3.4 Model Building

Twelve supervised machine learning algorithms from Weka [14] were used to build models that discriminated between MW and non-MW instances (windows). The following classifiers were used: Bayes network; naïve Bayes; logistic regression; SVM; *k*-nearest neighbors; decision table; JRip; C4.5 decision tree; random forest; random tree; REPTree; and REPTree with bagging.

We also varied four external parameters: (1) feature type; (2) window and segment size; (3) feature selection percentage; and (4) sampling method. With respect to feature type, models were built with global features, local features, or both global and local features using feature-level fusion.

The segment and window size(s) were varied because there are various tradeoffs at play. Specifically, a larger segment size resulted in fewer instances but a higher MW rate, thereby reducing class imbalance. A larger window size afforded more data for each instance, but it also reduced the number of instances available for segments with the same gap size (e.g., a window size of 30 and gap size of 15 resulted in fewer instances than a window size of 40 and gap size of 15). Thus, models were built with segment sizes of 45, 55, or 65 seconds, and window sizes of either 10, 15, 20, or 25 seconds.

Feature selection was used on the training set of each cross-validation fold (see below). Features were ranked using correlation-based feature selection (CFS) [15] from Weka and the top 30%, 50%, or 80% of features ranked were retained.

Class imbalance poses a well-known challenge for supervised classifiers. Hence, *training* sets were resampled using

downsampling or oversampling. Downsampling consisted of randomly removing instances from the majority class (non-MW) until the two classes were balanced. Oversampling consisted of using the Synthetic Minority Over-sampling Technique (SMOTE) algorithm [5]. We also built models without any resampling for comparison purposes.

Table 3. Confusion matrices for best models

Feature Type	Actual	Classified		Prior
		Yes	*No*	
Global	*Yes*	.65 (hit)	.35 (miss)	.25
	No	.55 (FA)	.45 (CR)	.75
Local	*Yes*	.67 (hit)	.33 (miss)	.26
	No	.47 (FA)	.53 (CR)	.74
Global + Local	*Yes*	.68 (hit)	.32 (miss)	.25
	No	.60 (FA)	.40 (CR)	.75

Note: Values are proportionalized by class label
FA = false alarm; CR = correct rejection

Tolerance analysis was performed to address multicollinearity prior to building each model [9]. This consisted of removing features with a tolerance below .2, which indicates highly collinear features (such as number of fixations and number of saccades).

3.5 Model Validation and Evaluation

The models were evaluated using leave-one-student-out cross-validation, which ensures that data from each student is exclusive to either the testing set or training set. Feature selection and resampling were performed on the training set only. Feature selection was performed with data from a random 66% of students in the training data in each fold. Feature rankings were summed over five different random selections. Resampling was also repeated for five iterations in each training fold.

Models were evaluated using the F_1 score for the target class (MW), which was compared to the MW F_1 score of a chance classifier. For example, if the actual model classified 52% of the instances as MW, the chance classifier would classify a random 52% of the instances as MW. This resulted in a chance precision equal to the actual base rate of MW and a chance recall equal to the predicted MW rate. We believe this chance model to offer a more stringent comparison than a simple minority baseline (assign MW to all instances).

4. RESULTS

4.1 MW Detection Accuracy

The overall best performing model achieved a MW F_1 score of .45, compared to a chance MW F_1 score of .35, which is consistent with a 29% improvement above chance (Table 2). The model was a decision table classifier that used local features and had a window size of 20 seconds, segment size of 65 seconds, 11 features, and a downsampled training set. The confusion matrix for the model (Table 3) shows that the model makes fewer misses than false alarms.

Table 2. Performance metrics (F_1) for best models

Feature	F_1MW (Chance)	F_1 MW	F_1 Non MW	F_1 Overall
Global	.35	.39	.57	.53
Local	**.35**	**.45**	**.64**	**.59**
Global+ Local	.36	.39	.54	.50

The best global and global + local models were SVMs with a window size of 15 seconds, a segment size of 65 seconds, and a downsampled training set. The global model contained 5 features, while the global + local model contained 11 features. Both models achieved a lower MW F_1 score than the local feature model, due to much higher false alarm rates (see Table 3 and Figure 3)

With respect to the external parameters, no clear trends were observed for window size, segment size, or proportion of features selected, but downsampling and SMOTEing the training set outperformed no resampling method.

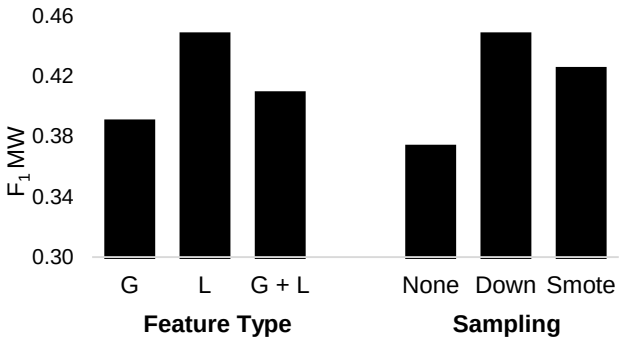

Figure 3. MW F_1 score for the best model by feature Type and resampling method. G = Global, L = Local, G + L = Global + Local; Down = Downsampling

4.2 Feature Analysis

We compared the mean values of each feature (computed per participant) for MW vs. non-MW instances with a two-tailed paired-samples *t*-test. We focused on the 16 global and 21 local features that were included in the best local and global models. Table 4 shows the effect size (Cohen's *d* – with positive values of *d* denoting higher values for MW compared to non-MW instances) for the significantly different ($p < .05$) features. We did not perform adjustments for multiple comparisons as the present analysis is exploratory in nature. Further, the number of significant findings (18%) is far greater than what we could achieve if we were capitalizing on chance alone.

We note that students were less likely to focus on the AOIs when they were MW. This is evidenced by a fewer number of frames where the smooth pursuit points intersected with the red balloon AOI or the most visually salient AOI. Further, there were fewer saccades onto and off of the most visually salient region during MW. Third, smooth pursuits had a longer range, but less variability in velocity during MW. Finally, there were fewer saccades during MW, which is consistent with previous findings of eye movements during MW while reading [2, 28]. Taken together, these results reflect a decoupling between salient regions

on the screen and eye movements, essentially signaling a breakdown in attentional synchronicity during MW.

Table 4. Effect size of difference in feature value between MW and non-MW instances

Feature	Cohen's *d*
Smooth Pursuit with Balloon AOI (frames)	-.37
Smooth Pursuit within 2° Saliency AOI (frames)	-.38
Number of Saccades away from Saliency AOI	-.39
Number of Saccades nearly onto Saliency AOI	-.35
Smooth Pursuit Duration Range (ms)	.30
Smooth Pursuit Velocity SD (°/s)	-.28
Number of Saccades	-.31

Note: SD = Standard Deviation; All tests were significant at $p < .05$ $df = 53$ for local features and $df = 50$ for global features.

5. DISCUSSION

There is a growing interest in assuaging the negative effects of MW during learning [6, 8]. Reliable MW detection is likely required to realize this goal. Although efforts in MW detection have had some success in the context of reading, MW detection in more media-rich contexts has been unexplored. As a step in this direction, this paper presents a student-independent detector of MW during narrative film comprehension, a context which is both timely and relevant given the increasing use of film and video lectures as educational resources.

5.1 Key Findings and Contributions

Our primary contribution is the computation of novel local gaze features that are based on the dynamic visual content of the film. Using these features, we were able to detect MW with a F_1 of .45 reflecting, a 29% improvement over chance. Furthermore, models built with local features outperformed models built with global features, or a combination of both global and local features. This suggests that taking the dynamic visual content into account (local features) can be more effective than merely tracking overall gaze patterns (global features), which has been the common method for MW detection during reading.

The local features likely performed better in the present context (narrative film viewing) compared to reading, because the unfolding visual stream provides cues as to where attention should be directed. Reading, in contrast, does not provide such explicit cues, so there is likely more variability in gaze patterns. This would explain why the global gaze features outperformed the local features during reading.

We also found that local features outperformed a combined local + global model, but we adopted a rather simplistic feature-level fusion strategy. It is an open question as to whether performance of the combined model could be boosted with more advanced fusion strategies.

Our results also provide insight into eye movements related to MW during film viewing. The key finding was that eye movements during MW were decoupled from the visually salient and important (balloon AOI) components of the visual stream, suggesting a breakdown in attentional control.

5.2 Applications

MW impedes comprehension by diverting a student's attention from the task at hand toward task-unrelated thoughts. Educational activities that involve comprehension from dynamic visual scenes, such as video clips or short instructional lectures, could benefit from pairing a MW detector with interventions that direct attention toward the learning task.

Beyond educational interfaces, detectors built from dynamic visual scenes have applications in entertainment and safety contexts. For example, they could be used to determine when viewers are more likely to MW while viewing entertainment films. The scenes could then be improved to increase viewer engagement.

Attentional focus is especially important for safety-critical tasks that require vigilance, such as air traffic control. MW detectors built for dynamic visual scenes might be more suitable for these types of tasks. However, empirical evidence is needed to determine the extent to which models built from narrative film viewing would generalize to these other contexts.

5.3 Limitations and Future Work

There were also some limitations with this study. The first limitation is the detection accuracy, which is moderate at best. It would be fruitful to explore improvements to the detector. Some possibilities include considering additional features based on other aspects of the visual content, such as faces or attempting more sophisticated modeling approaches that capture the unfolding temporal dynamics in eye gaze.

The segmentation method used in the study reflects yet another limitation as it rather arbitrarily segments the visual stream based on temporal windows. It would be worthwhile to explore content-based segmentation, such as scene transitions and event boundaries. This would also ensure consistent segments across students in lieu of the current method, which segments the film at different locations depending on the MW reports.

It is also unclear if the detector would generalize beyond the current film. "The Red Balloon" is a commercially produced film that employs cinematic devices to draw attention to the viewer [3]. In contrast, many instructional videos consist of an instructor lecturing to students [13] or lecturing over power point, which reflect rather different visual content.

Another limitation is the cost of eye tracking technology. The eye tracker used for this study was a cost-prohibitive Tobii TX300 that will not scale out of the laboratory. Fortunately, cost-effective eye tracking alternatives are becoming available, such as the Eye Tribe and Tobii EyeX, so replication with these trackers is warranted.

Finally, other limitations include a limited student sample (i.e. undergraduates from a private Midwestern college) and a laboratory setup. It is possible that the detector would not generalize to a more diverse student population or in more ecological environments. Retraining our model with data from more diverse populations and environments would be a suitable next step to increase its ecological validity.

5.4 Conclusion

We built the first student-independent gaze-based MW detector in the context of film viewing. The detector could be used to trigger interventions aimed at counteracting the negative effects of MW for an array of tasks involving dynamic visual scenes (e.g., watching instructional films, historic documentaries, or video lectures). Taken together, this work takes us closer to the goal of developing next-generation intelligent educational interfaces that "attend to attention" [6].

6. ACKNOWLEDGMENTS

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] Bernie the Early Bloomer - Bedtime Bedtime: *http://www.bedtime.com/bernie-the-early-bloomer/*. Accessed: 2016-02-09.

[2] Bixler, R. and D'Mello, S. 2015. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*. 26, 1 (Sep. 2015), 33-68.

[3] Bordwell, D. 2013. *Narration in the fiction film*. Routledge, New York, NY.

[4] Bradski, G. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*. 25, 11 (2000), 120-126.

[5] Chawla, N.V. et al. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*. 16, 1 (Jun. 2002), 321–357.

[6] D'Mello, S. et al. 2016. Attending to Attention: Detecting and Combating Mind Wandering during Computerized Reading. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, CA, May 07 - 12, 2016). *CHI EA '16*. ACM, New York, NY, 1661-1669.

[7] D'Mello, S. et al. 2013. Automatic Gaze-Based Detection of Mind Wandering during Reading. *Proceedings of the 7th International Conference on Educational Data Mining*. (Memphis, TN, Jul. 06 - 09, 2013) *EDM* '13. IEDMS, 364-365.

[8] D'Mello, S.K. 2016. Giving Eyesight to the Blind: Towards Attention-Aware AIED. *International Journal of Artificial Intelligence in Education*. 26, 2 (Jun. 2016), 645–659.

[9] Domingos, P. 2012. A few useful things to know about machine learning. *Communications of the ACM*. 55, 10 (Oct. 2012), 78–87.

[10] Drummond, J. and Litman, D. 2010. In the zone: Towards detecting student zoning out using supervised machine learning. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (Pittsburgh, PA, Jun. 14 - 18, 2010). *ITS '10*. Springer Berlin Heidelberg, 306–308.

[11] Elazary, L. and Itti, L. 2008. Interesting objects are visually salient. *Journal of Vision*. 8, 3 (Mar. 2008), 3–3.

[12] Franklin, M.S. et al. 2011. Catching the mind in flight: using behavioral indices to detect mindless reading in real time. *Psychon Bull Rev*. 18, 5 (Oct. 2011), 992–997.

[13] Guo, P.J. et al. 2014. How video production affects student engagement: an empirical study of MOOC videos. *Proceedings of the first ACM conference on Learning@ scale conference* (Atlanta, GA, Mar. 04 - 05, 2014), ACM, 41–50.

[14] Hall, M. et al. 2009. The WEKA Data Mining Software: an Update. *ACM SIGKDD Explorations Newsletter*. 11, 1 (2009), 10–18.

[15] Hall, M.A. 1999. *Correlation-Based Feature Selection for Machine Learning*. Doctoral Thesis. Department of Computer Science, The University of Waikato.

[16] Harel, J. et al. 2006. Graph-based visual saliency. In *Proceedings of the Advances in Neural Information Processing Systems* (Vancouver, BC, Dec. 04 - 06, 2006), *NIPS* '06. MIT Press, Cambridge, MA, 545–552.

[17] Holmqvist, K. et al. 2011. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, Oxford, UK.

[18] Karpov, A.V. and Komogortsev, O. 2013. Automated Classification and Scoring of Smooth Pursuit Eye Movements in Presence of Fixations and Saccades. *Journal of Behavioral Research Methods*. 45,1 (Mar. 2013) 203-215.

[19] Killingsworth, M.A. and Gilbert, D.T. 2010. A Wandering Mind is an Unhappy Mind. *Science*. 330, 6006 (Nov. 2010), 932–932.

[20] Komogortsev, O.V. et al. 2010. Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *Biomedical Engineering, IEEE Transactions on*. 57, 11 (Jul. 2010), 2635–2645.

[21] Kopp, K. et al. 2015. Mind wandering during film comprehension: The role of prior knowledge and situational interest. *Psychonomic Bulletin & Review*. (Sep. 2015), 1-7.

[22] Lamorisse, A. 1956. *The Red Balloon*. Penguin Random House LLC, New York, NY.

[23] Mills, C. and D'Mello, S. 2015. Toward a Real-time (Day) Dreamcatcher: Sensor-Free Detection of Mind Wandering During Online Reading. *Proceedings of the 8th International Conference on Educational Data Mining*. (Madrid, Spain, Jun. 26 - 29, 2015) *EDM* '15. IEDMS, 69-76.

[24] Mital, P.K. et al. 2011. Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion. *Cognitive Computation*. 3, 1 (Mar. 2011), 5–24.

[25] Pham, P. and Wang, J. 2015. AttentiveLearner: Improving Mobile MOOC Learning via Implicit Heart Rate Tracking. In *Proceedings of the 17th International Conference on Artificial Intelligence in Education*. (Madrid, Spain, Jun. 22 - 26, 2015). *AIED '15*. Springer International Publishing, 367–376.

[26] Randall, J.G. et al. 2014. Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin*. 140, 6 (Nov. 2014), 1411–1431.

[27] Rayner, K. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*. 124, 3 (Nov. 1998), 372.

[28] Reichle, E.D. et al. 2010. Eye Movements During Mindless Reading. *Psychological Science*. 21, 9 (Aug. 2010), 1300–1310.

[29] Risko, E.F. et al. 2012. Everyday attention: variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology*. 26, 2 (Apr. 2012), 234–242.

[30] Smallwood, J. et al. 2007. Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic Bulletin & Review*. 14, 2 (Apr. 2007), 230–236.

[31] Smallwood, J. and Schooler, J.W. 2006. The Restless Mind. *Psychological Bulletin*. 132, 6 (Nov. 2006), 946–958.

[32] Smith, T.J. and Mital, P.K. 2013. Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision*. 13, 8 (Jul. 2013), 16–16.

[33] Szpunar, K.K. et al. 2013. Mind wandering and education: from the classroom to online learning. *Frontiers in psychology*. 4 (Aug. 2013), 1-7.

[34] Tan, T. et al. 2015. Mind Wandering and the Incubation Effect in Insight Problem Solving. *Creativity Research Journal*. 27, 4 (Nov. 2015), 375–382.

[35] Yonetani, R. et al. 2012. Multi-mode saliency dynamics model for analyzing gaze and attention. *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, CA, Mar. 28 - 30, 2012). ETRA ' 12, ACM, 115–122.

[36] Zacks, J.M. 2010. The brain's cutting-room floor: segmentation of narrative cinema. *Frontiers in Human Neuroscience*. 4, 168 (Oct. 2010), 1-15.