Detecting Student Engagement: Human Versus Machine

Nigel Bosch
University of Notre Dame
384 Fitzpatrick Hall
Notre Dame, IN 46556, USA
pbosch1@nd.edu

ABSTRACT
Engagement is complex and multifaceted, but crucial to learning. Computerized learning environments can provide a superior learning experience for students by automatically detecting student engagement (and, thus also disengagement) and adapting to it. This paper describes results from several previous studies that utilized facial features to automatically detect student engagement, and proposes new methods to expand and improve results. Videos of students will be annotated by third-party observers as mind wandering (disengaged) or not mind wandering (engaged). Automatic detectors will also be trained to classify the same videos based on students' facial features, and compared to the machine predictions. These detectors will then be improved by engineering features to capture facial expressions noted by observers and more heavily weighting training instances that were exceptionally-well classified by observers. Finally, implications of previous results and proposed work are discussed.

CCS Concepts
• Human-centered computing→Human computer interaction (HCI) • Computing methodologies→Supervised learning by classification

Keywords
Affective computing; engagement detection; facial expressions

1. INTRODUCTION
Most people can relate to the experience of becoming disengaged from almost any task where distractions occur or daydreams happen. For example, a student might spend nearly as much time in a lecture text messaging or plumbing the depths of Wikipedia as they do actually listening to the teacher. Similarly, while reading a textbook you might go through the motions of reading but soon find yourself thinking about something else entirely. This lack of engagement (in other words, disengagement) can be detrimental to performance in tasks such as learning [17].

Unsurprisingly, previous research has shown that engagement is positively related to learning [15]. Educational software can utilize this relationship to improve the learning experience for students and promote learning by adapting to a student's level of engagement and redirecting them toward the learning goal if necessary (intervening). Such interventions can be triggered by automated engagement detection systems. For example, in a fully automated learning environment a hint could be given to a student if the system detects that the student is confused or frustrated by the learning material and may soon become disengaged [9].

Accurate engagement detection is thus key to developing such learning systems and strategies. Many techniques have been employed to detect engagement and related constructs, including interaction log-files, eye gaze, physiological measurements, facial features, and others [7,16]. Each of these channels of data has their own advantages and disadvantages. For example, interaction log-files require no sensors, but are often highly context dependent. Gaze trackers measure the locus of attention precisely, but are not common in learning environments. Physiological sensors are increasingly popular in fitness trackers and related hardware, and can be used easily throughout the day. However, they are less accurate than other methods for detecting some components of engagement. Facial features are widely available via inexpensive and commonplace webcams, but are sensitive to various factors like lighting, occlusions, and movement. This paper focuses on face-based engagement detection methods, because it is potentially superior to other methods in terms of availability in various domains and potentially complementary to other modalities by focusing on visible features.

There are several facets to engagement that must also be discussed to properly situate the proposed work within the body of related work. This paper examines three components of engagement: affective, cognitive, and behavioral. Examples of these might be a student who is interested in a topic and enjoying learning about it (affective engagement), a student who is reading a book and thinking about how the material integrates with their previous knowledge of the topic (cognitive), or a student diligently typing an essay (behavioral). Various affective states play a role in engagement as well. For example, frustration can lead to boredom [9], which in turn indicates a lack of affective and cognitive engagement. Figure 1 illustrates the model of affective states and engagement that will be considered in this paper.

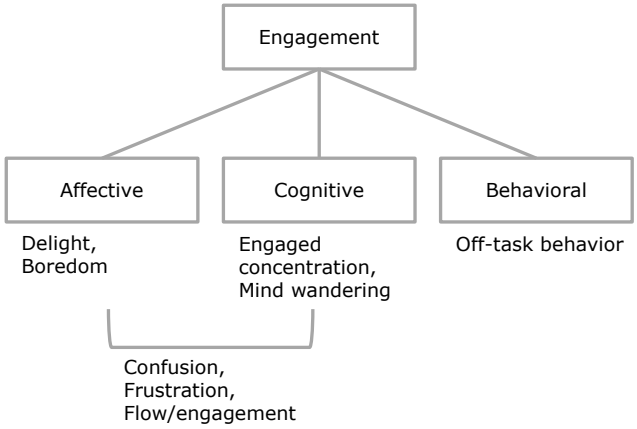

Figure 1. Conceptualized breakdown of engagement

The focus of proposed research in this paper is on mind wandering (MW), a type of cognitive disengagement. Students mind wander when they shift their focus from thinking about the learning task at hand toward unrelated thoughts [17]. For example, a student might mind wander thinking about a television show they recently saw. Such students are no longer cognitively

engaged in the task at hand, though they may not immediately realize it. Hence, it might be useful to detect MW and refocus students' attention. MW detection has been the subject of research efforts utilizing eye gaze and physiology [3,16], but face-based techniques are not yet well explored. Furthermore, it is not clear how well different techniques work for annotating MW.

There are various methods of engagement annotation, which typically fall into two broad categories: self-reports or third-party observations. One of the primary advantages of self-reported emotions is that the student making a self-report has access to their own internal state of mind, which an observer does not. On the other hand, observations made by a third party do not require interrupting students at all. However, observers lack access to the internal state of students, and must make their judgments based on external cues alone (which more closely matches the method an automatic detector must use).

This paper briefly reviews progress made toward the goal of automatic face-based engagement detection, including components of engagement and the affective states that manifest in relation to engagement (Figure 1). Both self-reported and observer labels of engagement have been employed for training the detectors as well. Proposed enhancements to engagement detection are then presented, focusing on a new study designed to capture cognitive disengagement and improve engagement detection via human knowledge.

Both self-reports and observations are also considered in related work, since the approaches have complementary strengths and weaknesses. Face-based approaches to automatic detection are discussed for various aspect of engagement mentioned (Figure 1), including affective states that manifest as a part of engagement.

2. RELATED WORK

Many methods have been used to detect engagements and its components [7]. Primarily face-based approaches are reviewed here, as the proposed work focuses on facial features.

2.1 Affective and cognitive state detection

Kapoor et al. [12] developed one of the first systems for detecting frustration in an automated learning environment. They used multimodal data channels including facial features (from video), a posture-sensing chair, a pressure-sensitive mouse, a skin conductance sensor, and interaction data to predict frustration. They were able to predict when a user would self-report frustration with 79% accuracy (chance being 58%).

Hoque et al. [11] used facial features and temporal information in videos to classify smiles as either frustrated or delighted – two states that are related to engagement and learning. They accurately distinguished between frustrated and delighted smiles correctly in 92% of cases. They also found differences between acted facial expressions and naturalistic facial expressions. In acted data only 10% of frustrated cases included a smile, whereas in naturally occurring frustration smiles were present in 90% of cases. These results illustrate that there can be large differences between naturalistic and posed data.

The Computer Expression Recognition Toolbox (CERT) [13] is a computer vision tool used to automatically detect AUs as well as head pose and head position information. CERT uses features extracted from Gabor filters as inputs to SVMs to provide likelihood estimates for the presence of 19 different AUs in any given frame of a video stream. It also supplies measures of unilateral (one side of the face only) AUs for three action units, as well as "Fear Brow" and "Distress Brow," which indicate the presence of combinations of AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), and AU4 (Brow Lowerer). CERT has been tested with databases of both posed facial expressions and spontaneous facial expressions, achieving accuracy of 90.1% and 79.9%, respectively, when discriminating between instances of the AU present vs. absent [13].

Grafsgaard et al. [10] used CERT to recognize the level of frustration (self-reported on a Likert scale) in a learning session and achieved modest results ($R^2 = .24$). Additionally, they found good agreement between the output of CERT AU recognition and human-coded ground truth measurements of AUs. After correcting for individual differences in facial feature movements they achieved Cohen's kappa >= .68 for several key AUs. They did not perform detection at a fine-grained level (i.e. specific affective episodes), instead detecting the presence of affect in the entire learning session. However, their work does provide evidence of the validity of CERT for automated AU detection.

2.2 Behavioral engagement detection

In a recent engagement detection effort, Whitehill et al. [18] used Gabor features (appearance-based features capturing textures of various parts of the face) with a support vector machine (SVM) classifier to detect engagement as students interacted with cognitive skills training software. Labels used in their study were obtained from retrospective annotation of videos by third-party observers. Four levels of engagement were annotated, ranging from complete disengagement (not even looking at the material) to strong engagement. This type of engagement annotation primarily captures behavioral engagement. They were able to detect engagement with an Area Under the ROC Curve (AUC, averaged across all four levels of engagement) of .729 where AUC = .5 is chance level detection.

Finally, off-task behavior detection in learning environments is perhaps the most clearly behavioral type of engagement detection. Off-task behavior has been detected with interaction log-file clickstream data [1]. However, face-based approaches in learning contexts are not yet well established.

3. CURRENT RESULTS

The results described briefly in this paper extend the related work to demonstrate the feasibility of automatic face-based detection of components of engagement.

Some of the completed results used texture-based facial features and heart rate detected from changes in skin color to detect behavioral engagement [14]. Students wrote essays and self-reported engagement (i.e., if they were working on the essay) both during the writing task and afterword. Engagement was detected with accuracy of AUC = .758 (versus chance = .5), using a fusion of both types of features extracted.

Completed work also included detection of components of cognitive and affective engagement using facial features. In one study confusion and frustration were detected at 22.1% and 23.2% above chance respectively [4]. In another study, both confusion (AUC = .637) and frustration (AUC = .609) were detected above chance levels. Importantly, engagement detectors fit to specific learning scenarios were more effective (average AUC = .595) than general detectors (average AUC = .554), indicating that the learning task can have an appreciable effect on detector accuracy.

Finally, research so far also explored some engagement detection aspects in the wild, using facial features extracted from face

videos in a computer-enabled classroom. Boredom (AUC = .610), confusion (.649), delight (.867), engaged concentration (.679), frustration (.631), and off-task behavior (.816) were all detected at levels above chance. These results demonstrated the feasibility of detecting various facets of engagement in the wild, despite the noisy nature of data collected in a classroom environment.

4. PROPOSED WORK

Completed work has primarily focused on affective, behavioral, and related facets of engagement. As discussed previously, MW is a cognitive component of engagement that has not been well researched in terms of facial features or automatic face-based detection. The proposed work (not yet completed) focuses on answering questions about human observer perception of MW, automated MW detection, and if observers can improve detection.

Data collection will consist of obtaining observer ratings of face videos of humans MW or not MW. These video clips come from a study in which 98 participants read an instructional text and self-reported MW whenever they realized they had been MW. Video clips for observer annotation will then be extracted in 12-second windows leading up to MW self-reports. 12 seconds was chosen to correspond to the average MW report time within a page, as a compromise between shorter windows (less data to use) and longer windows (fewer windows can be extracted because they won't fit between page start and MW report). Non-MW video clips will be similarly extracted from periods of time where there were no MW self-reports, with windows ending at the average MW report position. In total, there are 3,272 such clips available from the original dataset collected for proposed work.

Clips will be rated by observers on Amazon Mechanical Turk, which has been used in many previous studies (e.g., [6]). Observers will be shown a sequence of 10 clips asked to judge each clip as MW or non-MW and provide a confidence rating regarding their observation. Additionally, observers will be asked to describe the reason for each of their observations in a text box. They will be provided with detailed descriptions of MW to aid them in determining what constitutes MW and what does not.

After all videos have been coded by observers, the text responses will be examined to identify common themes (e.g., selected MW because participant yawned). The most frequent themes will then be made into checkboxes in the observation interface, and the text response will be removed. Clips will then be rated on Mechanical Turk repeatedly to improve reliability of ratings.

4.1 Observer-based MW classification

Maximum Likelihood Estimation (MLE) will be used to calculate the set of clip labels that is most consistent with the observer ratings. Observers who tend to disagree with other observers are likely less reliable and thus will have all of their ratings weighted lower, and vice versa. In the event of ties the clip label will be randomly assigned. This analysis will be the first measure of how well third-party observers can detect MW from facial expressions.

4.2 Automatic MW classification

One third of students (33 students) will be randomly chosen and their clips will be reserved as the evaluation set. The rest of the clips (roughly 2,200) will compose the development set which will be used to create automatic MW detectors.

A unique set of high- and low- level facial features will be extracted for MW classification. High level features will be based on action units extracted using EmotiontSDK for each video frame. AUs will then be aggregated across the duration of each 12 second clip to obtain the mean and standard deviation of each AU. Prior work has shown that various time scales are effective for different classification tasks [5]. Thus, AUs will also be aggregated for 3, 6, and 9 second subsections of each clip to create a set of multiscale features. Relationships between AUs will be captured by features measuring the Jensen-Shannon divergence (JSD) of AU pairs. Finally, temporal features of AUs will be encoded by applying 1-dimensional Gabor filters to each detected AU signal and measuring the patterns of presence and absence of each AU within the clip [2].

Low-level facial features will be extracted using Local Binary Patterns in Three Orthogonal Planes (LBP-TOP) and 2-dimensional Gabor filters. These features have been shown to be effective for engagement classification [14,18]. LBP-TOP features capture texture patterns, which can be indicative of facial expression changes. For example, if a student smiles the texture pattern near the mouth will change from ordinary skin texture to a lip texture as the mouth widens. Gabor filters are particularly well suited for detecting edges, which can capture not only the edges of facial features such as eyes and eyebrows, but also skin wrinkles that occur in some facial expressions (e.g., on the nose when the brow is furrowed).

Support vector machine (SVM) classifiers will be trained using leave-one-student-out cross validation on the development data. SVMs will be used because they have been used successfully in previous engagement detection research [18], and because they lend themselves to modification for improving predictions using human observations of MW (section 4.3). Individual SVMs will be trained for each group of related features (e.g., LBP-TOP features) and combined with a logistic regression. The feature set will be reduced using feature selection.

This analysis will provide a baseline for improvement of MW classification. It is also novel in that student MW during reading has not been well studied before. Additionally, timescale-invariant features have not been explored for MW detection.

4.3 Improving automatic predictions

Knowledge gained from observer ratings may be useful for improving the accuracy of automatic detectors. The first step will be to compare the accuracies of observer and automatic predictions. This will be done on the evaluation dataset. F1 score of MW will be the primary accuracy metric, area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) will also be considered as they are common metrics.

There are several potential avenues for improving the automatic detectors by integrating observer knowledge. Even if humans prove less accurate than computers at classifying MW, their input could prove valuable if they are able to classify different instances well than the automatic methods. First, instances in the development set that were poorly classified by the detectors but classified well by observers will be isolated. Then, the observers justifications for their ratings (both text responses and check boxes) will be examined to determine if there are facial cues that should be added as features. For example, LBP-TOP features could be engineered to capture features from a very specific part of the face, or more heavily weight features from that part of the face. Second, weights will be assigned to instances during training so this set of important isolated instances will be more influential in training, and thus influence the position of the SVM hyperplane. Similarly, instances that cannot be well classified by either humans or computers will be weighted lower as they are

likely unhelpful for classification. Third, examples from this set of isolated instances will be annotated by researchers to determine if there are additional clues observers may have used to accurately classify these instances. All of these methods will utilize the development training set only, to avoid overfitting to characteristics of the evaluation dataset.

Finally, observer judgments will be augmented with the automatic predictions by training a model using observer judgments and automatic predictions as features. This model will serve as a further comparison to determine if observers are utilizing important features that are not captured by automatic detectors.

This analysis will be the first to compare third-party observations with automatic face-based predictions of MW. It will also be the first to explore the possibility of improving face-based MW detection using knowledge gleaned from human observers.

5. CONCLUSIONS

Engagement is important for learning [8,17]. Engagement detection thus offers opportunities for improving learning through automated engagement evaluation and targeted interventions. This paper describes prior work laying the ground for automatic face-based detection of various aspects of engagement. Face-based detection is particularly attractive for practical applications due to its potential for context generalizability. MW is a relatively unexplored facet of engagement detection, especially with face-based approaches. This paper proposed work to address questions of how well third-party observers can detect MW and what can be done to improve automatic detection. The proposed work, if effective, will provide a powerful addition to computerized learning environments in the future by automatically detecting MW from faces and improving detection by incorporating observer annotations of MW.

6. ACKNOWLEDGEMENTS

I would like to thank my advisor, Sidney D'Mello, for his guidance on this research. 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 author(s) and do not necessarily reflect the views of the NSF.

7. REFERENCES

1. Ryan Shaun Baker, Albert T. Corbett, Kenneth R. Koedinger, and Angela Z. Wagner. 2004. Off-task behavior in the cognitive tutor classroom: When students "game the system." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 383–390.
2. Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, and Kang Lee. 2014. Automatic decoding of facial movements reveals deceptive pain expressions. *Current biology: CB* 24, 7: 738–743.
3. Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. 2014. Automated physiological-based detection of mind wandering during learning. *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)*, Switzerland: Springer International Publishing, 55–60.
4. Nigel Bosch, Yuxuan Chen, and Sidney D'Mello. 2014. It's written on your face: Detecting affective states from facial expressions while learning computer programming. *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)*, Switzerland: Springer International Publishing, 39–44.
5. Nigel Bosch, Sidney D'Mello, Ryan Baker, et al. 2015. Automatic detection of learning-centered affective states in the wild. *Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015)*, New York, NY: ACM, 379–388.
6. Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6, 1: 3–5.
7. Rafael A. Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing* 1, 1: 18–37.
8. Sidney D'Mello. 2013. A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology* 105, 4: 1082–1099.
9. Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2: 145–157.
10. Joseph F. Grafsgaard, Joseph B. Wiggins, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. 2013. Automatically recognizing facial expression: Predicting engagement and frustration. *Proceedings of the 6th International Conference on Educational Data Mining*.
11. Mohammed (Ehsan) Hoque, Daniel McDuff, and Rosalind W. Picard. 2012. Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing* 3, 3: 323–334.
12. Ashish Kapoor, Winslow Burleson, and Rosalind W. Picard. 2007. Automatic prediction of frustration. *International Journal of Human-Computer Studies* 65, 8: 724–736.
13. Gwen Littlewort, J. Whitehill, Tingfan Wu, et al. 2011. The computer expression recognition toolbox (CERT). *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, 298–305.
14. Hamed Monkaresi, Nigel Bosch, Rafael A. Calvo, and Sidney K. D'Mello. in press. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*.
15. Zachary A. Pardos, Ryan S. J. D. Baker, Maria O. C. Z. San Pedro, Sujith M. Gowda, and Supreeth M. Gowda. 2013. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, ACM, 117–124.
16. Erik D. Reichle, Andrew E. Reineberg, and Jonathan W. Schooler. 2010. Eye movements during mindless reading. *Psychological Science* 21, 9: 1300–1310.
17. Jonathan Smallwood, Daniel J. Fishman, and Jonathan W. Schooler. 2007. Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic Bulletin & Review* 14, 2: 230–236.
18. J. Whitehill, Z. Serpell, Yi-Ching Lin, A Foster, and J.R. Movellan. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* 5, 1: 86–98.