Detecting Student Emotions in Computer-Enabled Classrooms

Nigel Bosch,
Sidney K. D’Mello
University of Notre Dame,
Notre Dame, IN
pbosch1@nd.edu,
sdmello@nd.edu

Ryan S. Baker,
Jaclyn Ocumpaugh
Teachers College,
Columbia University,
New York, NY
rsb2162@tc.columbia.edu,
jo2424@tc.columbia.edu

Valerie Shute,
Matthew Ventura, Lubin Wang, Weinan Zhao
Florida State University,
Tallahassee, FL
{vshute, mventura,
lw10e}@fsu.edu,
weinan.zhao@gmail.com

Abstract
Affect detection is a key component of intelligent educational interfaces that can respond to the affective states of students. We use computer vision, learning analytics, and machine learning to detect students’ affect in the real-world environment of a school computer lab that contained as many as thirty students at a time. Students moved around, gestured, and talked to each other, making the task quite difficult. Despite these challenges, we were moderately successful at detecting boredom, confusion, delight, frustration, and engaged concentration in a manner that generalized across students, time, and demographics. Our model was applicable 98% of the time despite operating on noisy real-world data.

1 Introduction
Learning with educational interfaces elicits a range of affective states that can have both positive and negative connections to learning [D’Mello, 2013]. A human teacher or tutor can observe students’ affect in a classroom or one-on-one tutoring situation and use that information to determine when help is needed to adjust the pace or content of learning materials [Lepper, Woolverton, Mumme, & Gurtner, 1993]. However, computerized learning environments rarely consider student affect in selecting instructional strategies—a particularly critical omission given the central role of affect in learning [Calvo & D’Mello, 2011].

We believe that next-generation intelligent learning technologies should have some mechanism to respond to the affective states of students, whether by providing encouragement, altering materials to better suit the student, or redirecting the student to a different task when he/she becomes disengaged. Although some initial progress has been made in laboratory settings (see [D’Mello, Blanchard, Baker, Ocumpaugh, & Brawner, 2014] for a recent review), much work remains before affect-sensitive learning interfaces can be fielded at scale in the wild. A core challenge is the ability to detect student affect in real-world learning settings, which we address in this work by detecting the affective states that students naturally experience while interacting with an educational game in a computer-enabled classroom.

We focus on detecting the affective states that are common during interactions with technology, namely boredom, confusion, engagement/flow, frustration, happiness/delight, and anxiety [D’Mello, 2013]. We adapt a multimodal approach, combining facial features (primary) and interaction patterns (secondary). Facial features have long been linked to affect [Ekman, Freisen, & Ancoli, 1980], but are not robust for affect detection in the wild due to occlusion, movement, and lighting issues frequently encountered in computer-enabled classrooms. Interaction patterns are derived from logs of the student’s actions within the learning environment, and are thus less vulnerable to these factors. A multimodal approach is thus expected to capitalize on each approach’s benefits while mitigating their weaknesses.

We address several unique challenges of learning-centered affect detection in the wild: 1) detecting naturalistic (as opposed to acted) affective states in a relatively uncontrolled group setting (classroom); 2) testing generalization of these detectors across time and student demographics; and 3) exploring a tradeoff between more accurate affect detectors and those that are more robust to data availability issues that arise in the wild.

1.1 Related Work
Drawing from the ample body of research on affect detection [Calvo & D’Mello, 2010] we review key studies on affect detection with facial and multimodal features.

In one classic study, Kapoor et al. [2007] used multimodal data channels including facial features, a posture-sensing chair, a pressure-sensitive mouse, skin conductance, and interaction log-files to predict frustration while students interacted with an automated learning companion. They were able to predict when a student would self-report frustration with 79% accuracy, an improvement of 21% over chance.

Whitehill et al. [2014] used facial features to detect engagement as students interacted with cognitive skills training software. They were able to detect engagement with an
average area under the ROC curve (AUC) of .729 (AUC of .5 represents chance-level detection). More recently, Monkaresi et al. [in press] used facial features and heart rate estimated from face videos to detect engagement during writing. They achieved an AUC of .758, which we could consider to be state-of-the-art.

Bosch et al. [2014] built detectors of novice programming students’ affective states using facial features (action units, or AUs [Ekman & Friesen, 1978]) estimated by the Computer Expression Recognition Toolbox (CERT) [Littlewort et al., 2011]. They were able to detect confusion and frustration at levels above chance (22.1% and 23.2% better than chance, respectively), but accuracy was much lower for other states (11.2% above chance for engagement, 3.8% above chance for boredom).

Perhaps the study most relevant to the current work is [Arroyo et al., 2009]. The authors tracked emotions of high school and college mathematics students using both interaction features from log-files and facial features. Their best models explained 52% of the variance (R2) for confidence, 46% for frustration, 69% for excitement, and 29% for interest. However, these results should be interpreted with a modicum of caution, because the models were not cross-validated with a separate testing set. The dataset was also limited in size with as few as 20-36 instances in some cases, raising concerns of overfitting.

In summary, despite active research on affect detection, there is a paucity of research on learning-centered affect detection with naturalistic facial expressions in the wild. The present study addresses this challenge, and explores the generalizability of face-based affect detectors and the advantage afforded by a multimodal combination of face- and interaction-based affect detection.

2 Method

2.1 Data Collection

The sample consisted of 137 8th and 9th grade students (57 male, 80 female) enrolled in a public school in the Southwestern U.S. The study took place in one of the school’s computer-enabled classroom, which was equipped with about 30 desktop computers for schoolwork. Inexpensive webcams ($30) were mounted to the top of each computer monitor.

The main learning activity consisted of students interacting with the educational game Physics Playground [Shute, Ventura, & Kim, 2013] in groups of about 20 in 55 minute class periods over four days (data from two days is used here). Physics Playground is a two-dimensional game that requires the player to apply principles of Newtonian Physics in an attempt to guide a green ball to a red balloon in many challenging configurations (key goal). The primary way to move the ball is to draw simple machines (ramps, pendulums, levers, and springboards) on the screen that “come to life” once drawn (example in Fig. 1).

Students’ affective states were observed during their interactions with Physics Playground using the Baker-Rodrigo Observation Method Protocol (BROMP) field observation system [Ocumpaugh, Baker, & Rodrigo, 2015]. These observations served as affect labels for training detectors. In BROMP, trained observers use side glances to make a holistic judgment of students’ affect based on facial expressions, speech, posture, gestures, and interaction with the game (e.g., whether a student is progressing or struggling). We obtained 1,767 affect observations during the two days of data used in this study. The most common affective state observed was engagement (77.6%), followed by frustration (13.5%), boredom (4.3%), delight (2.3%), and confusion (2.3%).

2.2 Model Building

Video-based Features. We used FACET1, a commercialized version of CERT (see above), to estimate the likelihood of the presence of 19 AUs as well as head pose (orientation) and head position. Gross body movement was also estimated by the proportion of pixels in each video frame that differed from a constantly updated background image. Features were created by aggregating AUs, orientation, position, and body movement estimates in a window of time (3, 6, 9, 12, or 20 seconds) leading up to each BROMP observation using maximum, median, and standard deviation for aggregation. Feature selection was applied to obtain a sparser set of features for classification. RELIEF-F [Kononenko, 1994] was run on the training data in order to rank features and a set of the highest ranked features were then used in the models.

About a third (34%) of the instances were discarded because FACET was not able to register the face, and thus could not estimate the presence of AUs. Poor lighting, extreme head pose or position, occlusions from hand-to-face gestures, and rapid movements can all cause face registration errors; these issues were not uncommon due to the na-

---

1 Currently available as Emotient Module from iMotions (https://imotions.com)
ture of the game and the active behaviors of the young students in this study.

**Supervised Learning.** We built separate detectors for each affective state, which allowed the parameters (e.g., window size, features used) to be optimized for that particular affective state. A two-class approach was used for each affective state, where that affective state was discriminated from all others (e.g., confusion vs. all other). We experimented with supervised classifiers including C4.5 trees and Bayesian classifiers, using WEKA [Witten & Frank, 2000].

Models were cross-validated at the student level. Data from 66% of randomly-chosen students were used to train each classifier and the remaining students’ data were used to test its performance. Each model was trained and tested over 150 iterations. The students in the training and testing data for each iteration were chosen randomly with replacement to amortize random sampling errors. This approach ensures that the models are generalizable to new students since training and testing data sets are student-independent.

The affective distributions led to large class imbalances (e.g., .04 vs. .96 priors in the boredom vs. all other classification). Majority-class downsampling and synthetic over-sampling were used to equalize base rates in the training data to help combat this disadvantage.

### 3 Results

**Baseline Results.** The best results for baseline face-based affect detection (student-level cross-validation) are presented in Table 1. The number of instances refers to the total number of instances used to train the model, including negative examples. This number varied based on the window size because shorter windows represent fewer video frames, and thus have a lower probability of containing valid video data.

Accuracy (recognition rate) is not an ideal metric for evaluation when base rates are highly skewed, as they were here. For example, delight occurred 2.3% of the time, so a detector that always predicted “Not delight” would have a 97.7% recognition rate. AUC is recommended for skewed data and is used here as the primary metric of detection accuracy [Jeni, Cohn, & de la Torre, 2013].

Classification accuracy was better than chance for all affective states including the infrequently-occurring states with large class imbalances. Delight was detected best, likely due to overt facial features that often accompany it. Classification accuracy may have been lower for other affective states because they manifest more gradually and less dra-

**Generalization.** We also tested the generalizability of face-based detectors across days, time of day, gender, and ethnicity [Bosch, 2015] (as annotated by researchers). Detector accuracy was not greatly influenced by these differences. We found less than 4% decrease relative to within-group baseline accuracies. Fig. 2 illustrates the effect of generalization across key dimensions.

![Fig. 2. AUC change when generalizing across time and demographics](image)

**Availability.** The face-based affect detection results discussed thus far are derived from 65% of instances. The remaining instances are unclassifiable due to factors such as hand-to-face occlusion, rapid movement, and poor lighting. To increase availability (proportion of all instances from which features could be extracted), we developed multimodal affect detectors including features from the log-files recorded while students interacted with the game [Bosch, Chen, Baker, Shute, & D’Mello, 2015]. Interaction features were distilled from log-files and comprised 76 gameplay attributes theoretically linked to affect. Example features included the amount of time between start and end of level and the total number of objects. See [Kai et al., 2015] for additional details of interaction features computed for Physics Playground. Interaction-based detection was available in 94% of instances, but was less accurate than the face-based detectors (see Figure 3). Fusing these detectors at the decision-level using logistic regression yielded a multimodal detector that was nearly as accurate as the face-based detectors, but available in 98% of instances.

<table>
<thead>
<tr>
<th>Classification</th>
<th>AUC</th>
<th>Accuracy</th>
<th>Classifier</th>
<th>No. Instances</th>
<th>No. Features</th>
<th>Window Size (secs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boredom</td>
<td>0.610</td>
<td>64%</td>
<td>Classification Via Clustering</td>
<td>1305</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>Confusion</td>
<td>0.649</td>
<td>74%</td>
<td>Bayes Net</td>
<td>1293</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Delight</td>
<td>0.867</td>
<td>83%</td>
<td>Updateable Naïve Bayes</td>
<td>1003</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>Engagement</td>
<td>0.679</td>
<td>64%</td>
<td>Bayes Net</td>
<td>1228</td>
<td>51</td>
<td>9</td>
</tr>
<tr>
<td>Frustration</td>
<td>0.631</td>
<td>62%</td>
<td>Bayes Net</td>
<td>1132</td>
<td>51</td>
<td>6</td>
</tr>
</tbody>
</table>
We also showed that our approach generalized across days, time of day, gender, and ethnicity (as perceived by researchers). Accuracy decreased by less than 4% for each of these generalization tests, demonstrating that detectors can be applied quite broadly.

Despite these encouraging findings, this study is not without its limitations. First, the number of instances was limited for some affective states. Second, though the students in this study varied widely across some demographic variables, they were all approximately the same age and in the same location. Further research is needed to test detectors on a larger dataset and with more demographic variability.

A next step is to use the detectors to guide intelligent instructional strategies in an affect-sensitive version of Physics Playground. Given the moderate detection accuracy, the interventions must be fail-soft so that they are not harmful if delivered due to detection errors. Subtle strategies, such as re-ordering the problems to display an easier problem after a frustrating experience, may be used.

In summary, affect-sensitive interfaces offer the exciting possibility of endowing computers with the ability to sense and respond to student emotions, just like a gifted human teachers. This research takes an important step toward making this vision a reality, by demonstrating the feasibility of automated detection of student affect in a noisy real-world environment: a school.

Acknowledgements

This research was supported by the National Science Foundation (NSF) (DRL 1235958 and IIS 1523091) and the Bill & Melinda Gates Foundation. 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 or the Bill & Melinda Gates Foundation.

References


