Automatic Detection of Learning-Centered Affective States in the Wild

Nigel Bosch, Sidney D'Mello
University of Notre Dame
384 Fitzpatrick Hall, Notre Dame, IN 46556, USA
{pbosch1, sdmello}@nd.edu

Ryan Baker, Jaclyn Ocumpaugh
Teachers College, Columbia University
525 W. 120th Street, New York, NY 10027, USA
baker2@exchange.tc.columbia.edu, jocumpaugh@wpi.edu

Valerie Shute, Matthew Ventura, Lubin Wang, Weinan Zhao
Florida State University
3205G Stone Building, 1114 West Call Street, Tallahassee, FL 32306-4453, USA
{vshute, mventura, lw10e}@fsu.edu, weinan.zhao@gmail.com

ABSTRACT
Affect detection is a key component in developing intelligent educational interfaces that are capable of responding to the affective needs of students. In this paper, computer vision and machine learning techniques were used to detect students' affect as they used an educational game designed to teach fundamental principles of Newtonian physics. Data were collected in the real-world environment of a school computer lab, which provides unique challenges for detection of affect from facial expressions (primary channel) and gross body movements (secondary channel)—up to thirty students at a time participated in the class, moving around, gesturing, and talking to each other. Results were cross validated at the student level to ensure generalization to new students. Classification was successful at levels above chance for off-task behavior (area under receiver operating characteristic curve or AUC = .816) and each affective state including boredom (AUC =.610), confusion (.649), delight (.867), engagement (.679), and frustration (.631) as well as a five-way overall classification of affect (.655), despite the noisy nature of the data. Implications and prospects for affect-sensitive interfaces for educational software in classroom environments are discussed.

Author Keywords
Affect detection; naturalistic facial expressions; classroom data; in the wild.

IUI 2015, March 29–April 1, 2015, Atlanta, GA, USA.

http://dx.doi.org/10.1145/2678025.2701397

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Learning from intelligent educational interfaces elicits frequent affective responses from students and wide variations in their behavior. A variety of affective states occur frequently in learning contexts, and can have both positive and negative effects on students' learning [12,38]. For example, students often encounter exercises that require information or techniques with which they are not familiar. Confusion, frustration, boredom, and other affective states are elicited in response to how these impasses are resolved [4,12,15]. These and other affective experiences are particularly important because they are inextricably bound to learning by coloring students' perceptions of a learning environment and changing how well they learn from it [9,19].

A human teacher can observe students' affect in a classroom or one-on-one tutoring situation (cf. [28]) and use that information to determine who needs help and to adjust the pace or content of learning materials. On the other hand, computerized learning environments used in school computer labs rarely incorporate such accommodations into their instructional strategies. One of the many challenges of creating intelligent educational interfaces is developing systems that can detect and respond to the affective states of students, though some initial progress has been made in laboratory settings (see [13] for a recent review). The goal of these interfaces is to provide a computerized learning environment that responds to the affective needs of students, whether by redirecting off-task behavior, providing encouragement, or altering learning materials to better suit the student. Much work remains to be done for effective affect-sensitivity in learning environments in the wild, however. At the core of such systems is the ability to detect or anticipate the affective state of students, a proposition considered in this paper.

Affect can be detected using various types of data. For example, interaction data (e.g., number of actions taken in an interface, speed of help requests) [3], facial expressions [5,26], posture [33], and other data sources have been used to detect the affective states of students (see [6,42] for recent reviews). Facial-feature based affect detection is particularly attractive because (a) there is a strong link between facial features and affective states [16], (b) it does not depend on learning environment or content, and (c) it does not require expensive hardware, as webcams are ubiquitous on laptops and mobile devices. For these reasons, we focus on vision-based techniques to detect affective states from facial features (primary channel) and body movements (secondary channel). Affect detection based on facial features has been the subject of considerable previous research (see [6,42] for reviews); albeit outside of learning contexts and mainly in laboratory settings (see exceptions discussed below). Laboratory environments have the advantage of relatively consistent lighting and freedom from distractions from other students, cell phones, walking around, and numerous other factors. Users of computer interfaces in the wild (e.g., students in a school computer lab) may be subject to such external distractions that make affective data far noisier than data collected in laboratory contexts. In many cases, motion, unusual head pose, and face-occluding gestures are so prevalent that facial feature detection is not possible and affect detection must be done using other modalities.

Much of the previous work in face-based affect detection has also focused on the so-called basic emotions of anger, fear, sadness, happiness, disgust, and surprise (see review in [6]). However, a recent review and meta-analysis of 24 studies indicated that these basic-emotions are quite infrequent in the context of learning with educational software [12]. Instead, students' affective experience mainly consists of learning-centered affective states (which we focus on in this study), such as engaged concentration, boredom, confusion, frustration, happiness, and anxiety. It is unclear if these states can be detected with similar fidelity as the basic emotions, where the links between emotion and expression have been carefully mapped out for decades [16,37]. Similar mappings for the learning-centered affective states are largely missing (see [30] for some initial work in this direction), and it is an entirely open question if such mappings even exist.

The present paper addresses these two challenges in an effort to detect naturalistic episodes of learning-centered affective states while students were using a computer interface in the setting of a school's computer lab. Students were in a context that was rich in interruptions, distractions, and conversations from fellow students. If successful, we will have shown that learning-centered affective states can be detected from facial expressions and body movements in the context most basic to education—a school.

RELATED WORK

There is a rich history on affect detection from facial features [6,42]. To keep scope manageable, we focus here on papers attempting to detect facial expressions in the wild and papers detecting learning-centered affective states from naturalistic facial expressions, as opposed to acted (posed) facial expressions. Furthermore, although we also consider gross body movements as an additional measure, the emphasis of the work and consequently the literature review is on facial features.

Facial Expression and Affect in the Lab

Kapoor et al. [25] developed the first system detecting affect in a 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 in an automated learning companion. They were able to predict when a user would self-report frustration with 79% accuracy (chance being 58%). Furthermore, using similar multimodal sensor fusion techniques including facial features, Kapoor et al. [26] were able to classify interest/disinterest with 87% accuracy (chance being 52%).

Hoque et al. [23] used facial features and temporal information in videos to classify smiles as either frustrated or delighted – two states that are related to 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 the fact that there can be large differences between naturalistic and posed data.

In a more recent affect detection effort, Whitehill et al. [41] 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 human judges. Four levels of engagement were annotated, ranging from complete disengagement (not even looking at the material) to strong 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.

Gabor features have also been used for detection of Action Units (AUs). Action Units are labels for specific facial muscle activations (e.g., lowered brow) [17]. As noted in [39], detecting action units can be a useful intermediate step in the process of detecting affective states. AUs provide a small set of features for use in affect detection efforts. A large database of AU-labeled data can be used to train AU detectors, which can then be applied to new data to generate AU labels. This can be particularly useful for datasets that

are difficult to collect (such as data in the wild) and may thus have few instances. The smaller set of features provided by the AU detector (compared to, for example, hundreds of Gabor features) can reduce overfitting.

The Computer Expression Recognition Toolbox (CERT) [29] 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 [29].

Grafsgaard et al. [20] 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 achieved 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 and higher 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.

In another study using CERT, Bosch and D'Mello [5] used machine learning to build fine-grained detectors for learning-centered affective states of novice programming students using the likelihoods of AUs provided by CERT. The students took part in the study in a laboratory setting. Students made retrospective judgments of their own affective states. Confusion and frustration were detected at levels above chance (22.1% and 23.2% better than chance, respectively), but performance was much lower for other states (11.2% above chance for engagement, 3.8% above chance for boredom).

Facial Expressions and Affect in the Wild

All of the aforementioned studies have been conducted during one-on-one interactions and with the high degree of control afforded by the laboratory, so generalizability to real-world contexts is uncertain. However, some relevant research has been done in real-world contexts including a classroom or school computer lab as reviewed below.

Facial expression data collected in the wild have been the subject of some research. The Affectiva-MIT Facial Expression Dataset (AM-FED) [31] contains videos of participants recorded on their own computers, in various settings while they watched Super Bowl commercials which were likely to elicit smiles. They provided baseline performance for smile detection (AUC = .90), AU2 (outer brow raise, AUC = .72), and AU4 (brow lower, AUC = .70). This dataset has been used for detection of whether viewers liked the commercials (AUC = .82) and wanted to view them again (AUC = .79) [32]. However, affect detectors have not been developed on this dataset to date.

Another study on smiles collected in the wild was conducted on a college campus using cameras set up in various university buildings. Hernandez et al. [21] used computer vision techniques to detect smiles and found expected patterns (e.g., more smiles on weekends and holidays). They demonstrated the feasibility of detecting smiles in the wild, but the question of whether data in the wild can be used for affect detection remains open.

The Emotion Recognition in the Wild Challenge [10] is an effort to unify detection of affect in the wild by creating a common benchmark for various state of the art audio and visual affect detection techniques. The data set used in this challenge was the *Acted Facial Expressions in the Wild* database, which was compiled by using clips from movies. This resulted in professionally acted affective expressions, rather than naturally experienced ones, which raises some concerns due to well-known differences between acted and naturalistic expressions (e,g., see [23]).

Perhaps the closest study to the current paper is one by Arroyo et al. [2]. They tracked emotions of high school and college mathematics students using self-reports and also recorded several modalities (interaction data from log-files, facial features, posture, skin conductance, and mouse movements). Their best models explained 52% of the variance (R^2) for confidence, 46% for frustration, 69% for excitement, and 29% for interest. Although this research suggests that it might be possible to perform automated affect detection in classroom, this conclusion should be interpreted with a modicum of caution. This is because the model was not validated with a separate testing set (i.e. no cross validation was performed), and the size of the data set was very small (20-36 instances depending on model) due to missing data. These issues raise concerns of overfitting to specific students and instances in the training data.

Current Study

The literature review revealed that there are studies that focus on detection of naturalistic affective states, many of which go beyond the basic emotions by considering learning-centered affect. However, this work has been done within controlled lab contexts, so it is unclear if the detectors will generalize to the wild. On the other hand, researchers have started to make great strides towards moving to the wild, but these studies are limited in that they either focus on more atomic facial expressions rather than affective state detection [21,31], are still in need of cross-validation [2], or study acted instead of naturalistic affect [10].

To address these challenges, the present study, for the first time, considers the detection of learning-centered affective states in the wild. Videos of students faces' and affect labels (for supervised learning) was collected while students interacted with a game-based physics education environment called Physics Playground (formerly Newton's Playground; [40]) in their school's computer lab, which was regularly used by students for academic purposes.

METHOD

Data Collection

Participants. The sample consisted of 137 8th and 9th grade students (57 male, 80 female) who were enrolled in a public school in a medium-sized city in the Southeastern U.S. They were tested in groups of about 20 students per class period for a total of four periods on different days (55 minutes per period). Students in the 8th and 9th grades were selected because of the alignment of the Physics Playground content and the State Standards (relating to Newtonian Physics) at those grade levels.

Interface. 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 player can nudge the ball to the left and right (if the surface is flat) but the primary way to move the ball is by drawing/creating simple machines (which are called "agents of force and motion" in the game) on the screen that "come to life" once the object is drawn (example in Figure 1). Thus, the problems in Physics Playground require the player to draw/create four different agents (which are simple machine-like objects): inclined plane/ramps, pendulums, levers, and springboards. All solutions are drawn with colored lines using the mouse. Everything in the game obeys the basic laws of physics relating to gravity and Newton's three laws of motion.

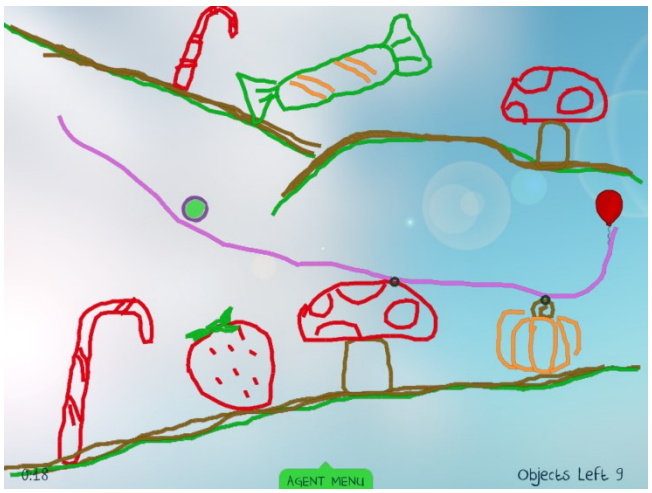

Figure 1. Ramp solution for a simple Physics Playground problem

Procedure. The study took place in one of the school's computer-enabled classrooms, which was equipped with about 30 desktop computers for schoolwork. Each computer was equipped with a monitor, mouse, keyboard, webcam, and headphones. Inexpensive webcams ($30) were affixed at the top of the monitor on each computer. At the beginning of each session, the webcam software displayed an interface that allowed students to position their faces in the center of the camera's view by adjusting the camera angle up or down. This process was guided by on-screen instructions and verbal instructions given by the experimenters, who were also available to answer any additional questions and to troubleshoot any problems.

We administered a qualitative physics pretest during the first day and a posttest at the end of the fourth day (both online). In this study we consider data from the second and third days (roughly two hours total) when students were only playing the game and not being tested. Students' affective states and on-task vs. off-task behaviors were observed during their interactions with Physics Playground using the Baker-Rodrigo Observation Method Protocol (BROMP) field observation system as detailed below [35]. These observations served as the ground truth labels used in training automated detectors.

The affective states of interest were boredom, confusion, delight, engaged concentration, and frustration. This list of states was selected based on previous research [12] and from observing students during the first day of data collection (this data was not used in the current models). In addition to affect, some basic student behaviors were observed. Students were coded as *on task* when looking at their own computer, *on-task conversation* when conversing with other students about what was happening on their own or others' screens, and *off task* in other situations (e.g., task-unrelated conversation, watching other students without conversation, using a cellphone).

BROMP. BROMP is a field coding protocol. In BROMP trained observers perform live affect and behavior annotations by observing students one at a time using a round-robin technique (observing one student until visible affect is detected or 20 seconds have elapsed and moving on to the next student). Observers use side glances to make a holistic judgment of the students' affect based on facial expressions, speech, body posture, gestures, and student interaction with the computer program (e.g., whether a student is progressing or struggling). Observers record students in a pre-determined order to maintain a representative sampling of students' affect, rather than focusing on the most interesting (but not most prevalent) things occurring in the classroom. Every BROMP observer was trained and tested on the protocol and achieved sufficient agreement (kappa $\geq$ .6) with a certified BROMP observer before coding the data.

The coding process was implemented using the HART application for Android devices [35], which enforces the protocol while facilitating data collection. Observation-codes recorded in HART were synchronized with the

videos recorded on the individual computers using Internet time servers.

It should be noted that there are many possible affect annotation schemes, each with their strengths and weaknesses, as recently reviewed in [36]. BROMP was selected for this study because it has been shown to achieve adequate reliability (among over 70 coders in over a dozen studies with a variety of learning environments [34]) in annotating affective states of a large number of students occurring in the "heat of the moment" and without interrupting or biasing students by asking them to self-report affect.

Instances of Affect Observed

It was not always possible to observe both affect and behavior in situations where students could not be easily observed (e.g., bathroom breaks, occlusions caused by hand to face gestures) or where the observer was not confident about an observation. Affect could not be observed in 8.1% of cases while on-task/off-task behavior could not be observed in 2.8% of cases. We obtained 1,767 successful observations of affective states and 1,899 observations of on-task/off-task behavior during the two days of data used in this study. The most common affective state observed was engaged concentration (77.6%), followed by frustrated (13.5%), bored (4.3%), delighted (2.3%), and confused (2.3%). On task behavior occurred 74.2% of the time, on-task conversation occurred 20.9% of the time, and off-task behavior occurred 4.9% of the time.

Model Building

Feature Engineering. We used FACET, a commercialized version of the CERT computer vision software, for facial feature extraction (http://www.emotient.com/products). Like CERT, FACET provides estimates of the likelihood estimates for the presence of nineteen AUs as well as head pose (orientation) and position information detected from video. Data from FACET was temporally aligned with affect observations in small windows. We tested five different window sizes (3, 6, 9, 12, and 20 seconds) for creation of features. Features were created by aggregating values obtained from FACET (AUs, orientation and position of the face) in a window of time leading up to each observation using maximum, median, and standard deviation. For example, with a six-second window we created three features from the AU4 channel (brow lower) by taking the maximum, median, and standard deviation of AU4 likelihood within the six seconds leading up to an affect observation. In all there were 78 facial features.

A quarter (25%) 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 game-like nature of the software and the active behaviors of the young students in this study. We also removed 9% of instances because the window of time leading up to the observation contained less than one second (13 frames) of data in which the face could be detected.

We also used features computed from gross body movement present in the videos as well. Body movement was calculated by measuring the proportion of pixels in each video frame that differed from a continuously updated estimate of the background image generated from the four previous frames (illustration in Figure 2). Previous work has shown that features derived using this technique correlate with relevant affective states including boredom, confusion, and frustration [11]. We created three body movement features using the maximum, median, and standard deviation of the proportion of different pixels within the window of time leading up to an observation, similar to the method used to create FACET features.

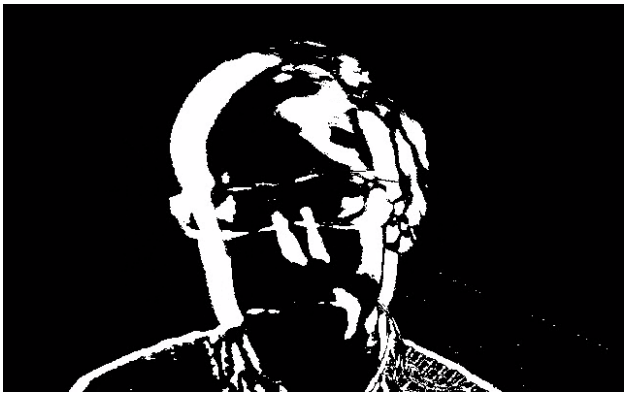

Figure 2. Silhouette visualization of motion detected in a video.

Tolerance analysis was used to eliminate features with high multicollinearity (variance inflation factor > 5) [1]. Feature selection was used to obtain a sparser, more diagnostic set of features for classification. RELIEF-F [27] was run on the *training* data in order to rank features. A proportion of the highest ranked features were then used in the models (.1, .2, .3, .4, .5, and .75 proportions were tested). Feature selection was performed using nested cross-validation on training data only. Ten iterations of feature selection were run on the training data, using data from a randomly chosen 67% of students within the training set for each iteration.

Supervised Learning. We built a detector for the overall five-way affect discrimination (bored, confused, delighted, engaged, and frustrated). In addition to the five-way classification, we also built separate detectors for each state. Building individual detectors for each state allows 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. For example, engaged was discriminated from all frustrated, bored, delighted, and confused instances combined (referred to as "all other"). Behaviors were grouped into two classes: 1) off task behaviors, and 2) both on task behaviors and on task conversation (i.e. not off task).

Classification	AUC	Accuracy	Classifier	No. Instances	No. Features	Window Size (secs)
Five-Way Affect	0.655	54%	Bayes Net	1209	38	9
Bored vs. Other	0.610	64%	Classification Via Clustering	1305	20	12
Confused vs. Other	0.649	74%	Bayes Net	1293	15	12
Delighted vs. Other	0.867	83%	Updateable Naïve Bayes	1003	24	3
Engaged vs. Other	0.679	64%	Bayes Net	1228	51	9
Frustrated vs. Other	0.631	62%	Bayes Net	1132	51	6
Off Task	0.816	81%	Logistic Regression	1381	15	12

Table 1. Details and results for classifiers.

The affective and behavior distributions lead to large class imbalances (e.g. .04 vs. .96 class priors in the bored vs. all other classification). Two different sampling techniques were used (on training data only) to compensate for class imbalance. These included downsampling (removal of random instances from the majority class) and synthetic oversampling (with SMOTE; [7]) to create equal class sizes. SMOTE creates synthetic training data by interpolating feature values between an instance and randomly chosen nearest neighbors. The distributions in the testing data were not changed, to preserve the validity of the results.

We built classification models for these seven discriminations (overall, five affective state models, and off task vs. on task), using 14 different classifiers including support vector machines, C4.5 trees, Bayesian classifiers, and others in the Waikato Environment for Knowledge Analysis (WEKA), a machine learning tool [22].

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 each trained and tested over 150 iterations with random students chosen each time, to help amortize random sampling errors. This helps models generalize to new learners since training and testing data sets are student-independent.

RESULTS

The best results for affect and off-task detection are presented in Table 1. The number of instances refers to the total number of instances that could be used to train the model, including negative examples. This number varies from model to model based on the window size because shorter windows have less data and are thus slightly less likely to contain at least one second of valid data (as stated earlier, windows with less than one second of valid data were not used).

Accuracy (recognition rate) for affect detectors varies widely in terms of percentage of instances correctly classified. However, accuracy is not a good performance metric for classification in situations where class distributions are highly skewed, as they are in this data. For example, delight occurs 2.3% of the time, which means a detector that simply guesses "Not delighted" for every single instance would have 97.7% accuracy. Other metrics such as Cohen's Kappa also provide unstable estimates when class distractions are widely skewed [24]. AUC is the recommended metric for skewed data and is used here as the primary measure of detection accuracy.

Area under the ROC curve (AUC) is only defined for binary classes, so we created an aggregate AUC by calculating AUC for each class versus all others within the five-way discrimination and averaging the results (see [18] for an overview of this and other methods of calculating AUC for multiple classes). The overall five-way discrimination between all affective states performed above chance with mean AUC = .655 (chance AUC = .500). For individual detectors, classification performed above chance for each affective state and off-task behavior.

Of particular note is the fact that classification was successful for infrequent states despite large class imbalances. Balanced classification performance is reflected in the confusion matrices for these states. For example, Table 2 shows the confusion matrix for delight, one of the most imbalanced affective states. Note that prior proportion of delighted is 2.9% rather than 2.3% as in the original observations because of instances removed due to face detection failures.

Actual	Classified		Priors
	Delighted	*All Other*	
Delighted	.685 (hit)	.315 (miss)	.029
All Other	.169 (false alarm)	.831 (correct rejection)	.971

Table 2. Confusion matrix for delighted affect.

The delight detector illustrated the effectiveness of using the SMOTE oversampling technique on the training data in order to improve model fit. In fact, SMOTE (as opposed to the alternatives, downsampling the training data or using no

balancing techniques) improved the results for all of the individual models except engaged concentration, which was already a relatively balanced discrimination. It comprises 78% of the affect observations, so it is not surprising that oversampling the training set did not improve performance for that particular affective state.

Figure 3 further illustrates the effect of using SMOTE on the affect and off-task detection models. Note that the mean hit (true positive) rate improves noticeably for models built without SMOTE, though with a slightly lower correct rejection rate. This effect arises because the detectors trained without SMOTE are biased towards recognizing the majority class (i.e. "all other" for all affective states except engagement). On the other hand, detectors built with SMOTE have equal numbers of both the affective state and the "other" instances, so they are better trained to recognize the affective state of interest.

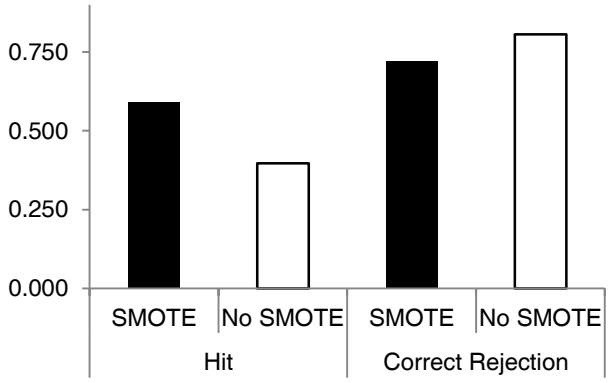

Figure 3. Comparison of mean hit rates and correct rejection rates for best models built with and without using SMOTE.

We also investigated the relationship between window size and classification performance in further detail. Figure 4 presents the performance of the best model for each affective state at different window sizes, which illustrates clear trends in the data.

Classification accuracy was not very dependent on the window size for boredom, engagement, and frustration (dotted lines in Figure 4), but it was clearly relevant for confusion, delight, and off task behavior (solid lines in Figure 4). The performance decrease for larger windows sizes in the detection of delight may be due to differences in the inherent temporal dynamics of expressive behavior for the different states [14]. For example, expressions of delight may last just a few seconds while confusion might be expressed for a longer period of time, though further research is needed to study this issue more thoroughly. Nevertheless, our results show that varying the window size between different detectors was an important consideration for some affective states. [1]

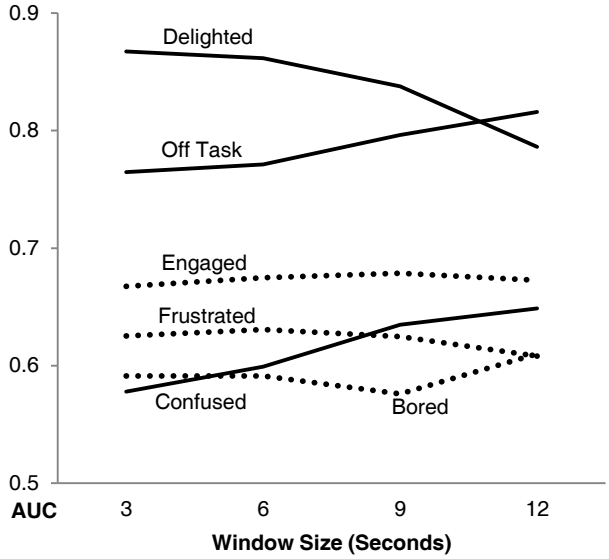

Figure 4. Detection results across window sizes.

DISCUSSION

Affect detection is a crucial component for affect-sensitive user interfaces, which aspire to improve students' engagement and learning by dynamically responding to affect. The inexpensive, ubiquitous nature of webcams on computers makes facial expression recognition an attractive modality to consider for affect detection. We expanded on the considerable body of vision-based affect detection research by building detectors for learning-centered affective states using data collected in the wild. Specifically, we have shown that affect detection is possible with data collected in a computer-enabled classroom environment in which students were subject to distractions, uncontrolled lighting conditions, and other factors which complicate affect detection. In this section, we discuss major findings, limitations of the present study, and implications of our affect detectors for future work with affect-sensitive interfaces.

Major Findings. Our key contribution was the development and validation of face-based detectors for learning-centered affect in a noisy school environment. We demonstrated that automatic detection of boredom, confusion, delight, engagement, frustration, and off-task behavior in the wild was possible for students using an educational game in a computer-enabled classroom environment—though many challenges exist for these classification tasks, such as classroom distractions and large imbalances in affective distributions.

[1] We also built models with a 20 second window. However, classification performance for those models was no better than other models so they were not further analyzed.

With respect to class distractions, students in the current study fidgeted, talked with one another, asked questions, left to go to the bathroom, and even occasionally used their cellphones (against classroom policy). In some situations multiple students crowded around the same screen to view something that another student had done. In short, students behaved as can be expected in a school computer lab. Furthermore, lighting conditions were inconsistent across students, in part due to placement of computers. Students' faces in some videos were well-illuminated, while they were barely visible in others. We were able to create detectors without excluding any of these difficult but realistic situations, except where faces could not be automatically detected at all in the video. In fact, we were unable to register the face in 34% of instances using modern computer vision techniques—an illustration of just how much uncontrolled lighting and the way students move, occlude, and pose their faces can make affect detection difficult in the wild.

Creating one set of parameters to use for all models is attractive for the sake of simplicity. However, we found differences in detection performance for some classifications with respect to window size. Confusion and off-task classifications worked better with larger window sizes, while delighted classification was better with a smaller window. Classifiers and feature selection parameters varied as well. This suggests that there are important differences in ideal parameters for different models, and that better performance can be achieved by tailoring models to their specific classification tasks.

Imbalance in affective state distributions is another challenge for affect detectors. This was a major concern with the present data, as three of the states occurred at rates less than 5%, while the most frequent state occurred at a rate nearing 80%. Despite this extreme skew, we were able to synthetically oversample the training data to create models that were not heavily biased against predicting the minority states. This is particularly important for future applications to affect-sensitive educational interfaces because detectors must be able to recognize relatively infrequent affective states that are important to learning (e.g., confusion) [12]. Infrequent does not mean inconsequential, however, since one or two episodes of intense frustration can disrupt an entire learning experience.

Limitations. This study is not without its limitations. First, the sample size is limited for some affective states, due in part to the difficulty of collecting data in the wild. This limitation was partially overcome by using SMOTE to create synthetic training data, but oversampling is not a perfect substitute for the diversity of genuine data. 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. A further study testing detectors on data with more variability in age and geographic distribution would be useful for determining the level to which results might generalize to interfaces targeted at a different group of students, such as elementary school students. Third, the distribution of affective states experienced may be dependent on the interface used. The interface in this study was game-based, which may increase engagement and decrease the incidence of other affective states compared to some other types of interfaces. Similarly, the observation method used (BROMP) requires observers to be in the room, which could influence students displays of affect similar to the Hawthorne effect [8]. This could be addressed by an additional study comparing the incidence of affective states experienced by students in the wild using a variety of different educational interfaces and affect collection methodologies.

Towards Affect-Sensitive Intelligent Interfaces. The detectors we created will be used to create intelligent instructional strategies towards developing an affect-sensitive version of Physics Playground. Separate strategies will be used for each affective state and off-task behavior. For example, when the detectors determine that a student is engaged or delighted, Physics Playground may not intervene at all. Confusion and frustration offer intervention opportunities in the form of hints or revisiting introductory material related to the concepts in the current problem. If the student has recently been frustrated and unable to complete problems, an easier problem might be suggested. Conversely, a more difficult problem might be appropriate if the student has not been challenged by recently completed problems. Boredom might be addressed by suggesting that the student attempt a new problem or by calibrating difficulty.

These aforementioned strategies have the goal of improving learning, but much work remains to be done in determining what types of interventions should be used in this context and how frequently they should be applied. Special considerations must also be given to the probability of spurious detections (false alarms) when designing and implementing these strategies since incorrect interventions could cause confusion or annoyance. In particular, interventions must be fail-soft so that learning is not negatively impacted by the strategies. Subtle strategies, such as re-ordering the problems to display an easier problem after a frustrating experience, may prove more effective. Future work will be needed to test a variety of intervention strategies in order to determine the most effective way to respond to the sensed affect.

It is our hope that affect detection will one day lead to intelligent, adaptive educational interfaces for use in computer-enabled classrooms. The research presented in this paper represents an important initial step toward accomplishing this goal by demonstrating techniques that can be used to detect students' affect in a noisy real-world environment.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF) (DRL 1235958) 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

1. Allison, P.D. *Multiple regression: A primer.* Pine Forge Press, 1999.
2. Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K., and Christopherson, R. Emotion sensors go to school. *AIED*, (2009), 17–24.
3. Baker, R., Gowda, S.M., Wixon, M., et al. Towards sensor-free affect detection in cognitive tutor algebra. *Proceedings of the 5th International Conference on Educational Data Mining*, (2012), 126–133.
4. Baker, R., Rodrigo, M.M.T., and Xolocotzin, U.E. The dynamics of affective transitions in simulation problem-solving environments. In A.C.R. Paiva, R. Prada and R.W. Picard, eds., *Affective Computing and Intelligent Interaction*. Springer, Berlin Heidelberg, 2007, 666–677.
5. Bosch, N., Chen, Y., and D'Mello, S. 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 (2014), 39–44.
6. Calvo, R.A. and D'Mello, S. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing 1*, 1 (2010), 18–37.
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16*, (2011), 321–357.
8. Cook, D.L. The Hawthorne effect in educational research. *Phi Delta Kappan*, (1962), 116–122.
9. Craig, S., Graesser, A., Sullins, J., and Gholson, B. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media 29*, 3 (2004), 241–250.
10. Dhall, A., Goecke, R., Joshi, J., Wagner, M., and Gedeon, T. Emotion recognition in the wild challenge 2013. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ACM (2013), 509–516.
11. D'Mello, S. Dynamical emotions: bodily dynamics of affect during problem solving. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, (2011).
12. D'Mello, S. A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology 105*, 4 (2013), 1082–1099.
13. D'Mello, S., Blanchard, N., Baker, R., Ocumpaugh, J., and Brawner, K. I feel your pain: A selective review of affect-sensitive instructional strategies. In R. Sottilare, A. Graesser, X. Hu and B. Goldberg, eds., *Design Recommendations for Intelligent Tutoring Systems - Volume 2: Instructional Management*. 2014, 35–48.
14. D'Mello, S. and Graesser, A. The half-life of cognitive-affective states during complex learning. *Cognition & Emotion 25*, 7 (2011), 1299–1308.
15. D'Mello, S. and Graesser, A. Dynamics of affective states during complex learning. *Learning and Instruction 22*, 2 (2012), 145–157.
16. Ekman, P., Freisen, W.V., and Ancoli, S. Facial signs of emotional experience. *Journal of Personality and Social Psychology 39*, 6 (1980), 1125–1134.
17. Ekman, P. and Friesen, W.V. Facial action coding system. *Consulting Psychologist Press*, (1978), Palo Alto, CA.
18. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters 27*, 8 (2006), 861–874.
19. Frenzel, A.C., Pekrun, R., and Goetz, T. Perceived learning environment and students' emotional experiences: A multilevel analysis of mathematics classrooms. *Learning and Instruction 17*, 5 (2007), 478–493.
20. Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., and Lester, J.C. Automatically recognizing facial indicators of frustration: A learning-centric analysis. (2013).
21. Hernandez, J., Hoque, M. (Ehsan), Drevo, W., and Picard, R.W. Mood meter: Counting smiles in the wild. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ACM (2012), 301–310.
22. Holmes, G., Donkin, A., and Witten, I.H. WEKA: a machine learning workbench. *Proceedings of the Second Australian and New Zealand Conference on Intelligent Information Systems*, (1994), 357–361.
23. Hoque, M.E., McDuff, D., and Picard, R.W. Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing 3*, 3 (2012), 323–334.
24. Jeni, L.., Cohn, J.F., and de la Torre, F. Facing imbalanced data–Recommendations for the use of performance metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, (2013), 245–251.
25. Kapoor, A., Burleson, W., and Picard, R.W. Automatic prediction of frustration. *International Journal of Human-Computer Studies 65*, 8 (2007), 724–736.
26. Kapoor, A. and Picard, R.W. Multimodal affect recognition in learning environments. *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ACM (2005), 677–682.
27. Kononenko, I. Estimating attributes: Analysis and extensions of RELIEF. In F. Bergadano and L.D. Raedt, eds., *Machine Learning: ECML-94*. Springer, Berlin Heidelberg, 1994, 171–182.

28. Lepper, M.R., Woolverton, M., Mumme, D.L., and Gurtner, J. Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. *Computers as cognitive tools 1993*, (1993), 75–105.
29. Littlewort, G., Whitehill, J., Wu, T., et al. The computer expression recognition toolbox (CERT). *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, (2011), 298–305.
30. McDaniel, B.T., D'Mello, S., King, B.G., Chipman, P., Tapp, K., and Graesser, A. Facial features for affective state detection in learning environments. *Proceedings of the 29th Annual Cognitive Science Society*, (2007), 467–472.
31. McDuff, D., El Kaliouby, R., Senechal, T., Amr, M., Cohn, J.F., and Picard, R. Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected in-the-wild. *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2013), 881–888.
32. McDuff, D., El Kaliouby, R., Senechal, T., Demirdjian, D., and Picard, R. Automatic measurement of ad preferences from facial responses gathered over the Internet. *Image and Vision Computing 32*, 10 (2014), 630–640.
33. Mota, S. and Picard, R.W. Automated posture analysis for detecting learner's interest level. *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '03)*, (2003), 49–56.
34. Ocumpaugh, J., Baker, R., Kamarainen, A., and Metcalf, S. Modifying field observation methods on the fly: Creative metanarrative and disgust in an environmental MUVE. *Proceedings of the 4th International Workshop on Personalization Approaches in Learning Environments (PALE), held in conjunction with the 22nd International Conference on User Modeling, Adaptation, and Personalization (UMAP 2014)*, (2014), 49–54.
35. Ocumpaugh, J., Baker, R., and Rodrigo, M.M.T. *Baker-Rodrigo observation method protocol (BROMP) 1.0. Training manual version 1.0.* Technical Report. New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences, 2012.
36. Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., and Baker, R. Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education 22*, 3 (2013), 107–140.
37. Reisenzein, R., Studtmann, M., and Horstmann, G. Coherence between emotion and facial expression: Evidence from laboratory experiments. *Emotion Review 5*, 1 (2013), 16–23.
38. Schutz, P. and Pekrun, R., eds. *Emotion in Education.* Academic Press, San Diego, CA, 2007.
39. Senechal, T., Bailly, K., and Prevost, L. Impact of action unit detection in automatic emotion recognition. *Pattern Analysis and Applications 17*, 1 (2014), 51–67.
40. Shute, V.J., Ventura, M., and Kim, Y.J. Assessment and learning of qualitative physics in Newton's Playground. *The Journal of Educational Research 106*, 6 (2013), 423–430.
41. Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., and Movellan, J.R. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing 5*, 1 (2014), 86–98.
42. Zeng, Z., Pantic, M., Roisman, G.I., and Huang, T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 1 (2009), 39–58.