A Comparison of Video-based and Interaction-based Affect Detectors in Physics Playground

Shiming Kai[1], Luc Paquette[1], Ryan S. Baker[1], Nigel Bosch[2], Sidney D'Mello[2], Jaclyn Ocumpaugh[1], Valerie Shute[3], Matthew Ventura[3]

[1]Teachers College Columbia University, 525 W 120th St. New York, NY 10027
[2]University of Notre Dame, 384 Fitzpatrick Hall, Notre Dame, IN 46556
[3]Florida State University, 3205G Stone Building, 1114 West Call Street, Tallahassee, FL 32306
{smk2184, paquette}@tc.columbia.edu, baker2@exchange.tc.columbia.edu, jo2424@tc.columbia.edu {pbosch, sdmello}@nd.edu, {vshute, mventura}@fsu.edu

ABSTRACT

Increased attention to the relationships between affect and learning has led to the development of machine-learned models that are able to identify students' affective states in computerized learning environments. Data for these affect detectors have been collected from multiple modalities including physical sensors, dialogue logs, and logs of students' interactions with the learning environment. While researchers have successfully developed detectors based on each of these sources, little work has been done to compare the performance of these detectors. In this paper, we address this issue by comparing interaction-based and video-based affect detectors for a physics game called Physics Playground. Specifically, we report on the development and detection accuracy of two suites of affect and behavioral detectors. The first suite of detectors applies facial expression recognition to video data collected with webcams, while the second focuses on students' interactions with the game as recorded in log-files. Ground–truth affect and behavior annotations for both face- and interaction-based detectors were obtained via live field observations during game-play. We first compare the performance of these detectors in predicting students' affective states and off-task behaviors, and then proceed to outline the strengths and weakness of each approach.

Keywords

Video-based detectors, interaction-based detectors, affect, behavior, Physics Playground

1. INTRODUCTION

The development of models that can automatically detect student affect now constitutes a considerable body of research [12,30], particularly in computerized learning contexts [1,33,34], where researchers have successfully built affect-sensitive learning systems that aim to significantly enhance learning outcomes [4,21,29]. In general, researchers attempting to develop affect detectors have developed systems falling into two categories: interaction-based detectors [9] and physical sensor-based detectors [12]. Many successful efforts to detect student affect in intelligent tutoring systems have used visual, audio or physiological sensors, such as webcams, pressure sensitive seat or back pads, and pressure-sensing keyboards and mice [3,27,36,40].

The development of sensor-based detectors has progressed significantly over the last decade, but one limitation to this research is that much of it has taken place in laboratory conditions, which may not generalize well to real-world settings [9]. While efforts are being made to address this issue [Arroyo et al, 2009], there are often serious obstacles to using sensors in regular classrooms. For example, sensor equipment may be bulky or otherwise obtrusive, distracting students from their primary tasks (learning); sensors may also be expensive and prone to malfunction, making large-scale implementation impractical, particularly for schools that are already financially strained. On the other hand, because physical sensors are external to specific learning systems, their use in affect detection creates the opportunity for them to be applied to entirely new learning systems, though this possibility has yet to be empirically tested.

Interaction-based detection [9] has also improved over the last decade. Unlike sensor-based detectors, which rely upon the physical reactions of the student, these detectors infer affective states from students' interactions with computerized learning systems [5,7,9,14,28,29]. The fact that interaction-based affect detectors rely on student interactions makes it possible for them to run in the background in real time at no extra cost to a school that is using the learning system. Their unobtrusive and cost-efficient nature also makes it feasible to apply interaction-based detectors at scale, leading to a growing field of research regarding discovery with models [8]. For example, interaction-based affect detection has been useful in predicting student long-term outcomes, including standardized exam scores [29] and college attendance [35]. Basing affect detection on student interactions with the system, however, give rise to issues with generalizing such detectors across populations [25] and learning systems. Because interaction-based detectors are highly dependent on the computation of features that captures the student's interactions with the specific learning platform, the type of features generated is contingent on the learning system itself, making it difficult to apply the same sets of features across different systems.

It has become clear that each modeling approach has its own utility; researchers have thus begun to speculate on effectiveness across the various approaches and the possible applications of multimodal detectors. However, the body of research that addresses this question is currently quite limited. Arroyo and colleagues [4] applied sensor-based detectors in a classroom setting, and compared performances between interaction-only detectors and detectors using both interaction and sensor data, in predicting student affect. They found that the inclusion of sensor data in the detectors improved performance and accuracy in

identifying student affect. However, a direct comparison between the two types of detectors was not made. Furthermore, the sample size tested was relatively small (26-30 instances depending on model), and the data was not cross-validated. Comparisons between types of detectors were made in D'Mello and Graesser's study [18], which compared interaction, sensor and face-based detectors in an automated tutor. They found face-based detectors to perform better than interaction and posture-based detectors at predicting spontaneous affective states. However, the study was conducted in a controlled laboratory setting, and the facial features recorded were manually annotated.

In this paper, we build detectors of student affect in classroom settings, using both sensor-based and interaction-based approaches. For feasibility of scaling, we limit physical sensors to webcams. For feasibility of comparison, the two types of detectors are built in comparable fashions, using the same ground truth data obtained from field observations that were conducted during the study. We conduct this comparison in the context of 8th and 9th grade students playing an educational game, Physics Playground, in the Southeastern United States. Different approaches were used to build each suite of detectors in order to capitalize on the affordances of each modality. However, the methods and metrics to establish accuracy were held constant in order to render the comparison meaningful.

2. PHYSICS PLAYGROUND

Physics Playground (formerly, Newton's Playground, see [38]) is a 2-dimensional physics game where students apply various Newtonian principles as they create and guide a ball to a red balloon placed on screen [37]. It offers an exploratory and open-ended game-like interface that allows students to move at their own pace. Thus, Physics Playground encourages conceptual learning of the relevant physics concepts through experimentation and exploration. All objects in the game obey the basic laws of physics, (i.e., gravity and Newton's basic laws of motion).

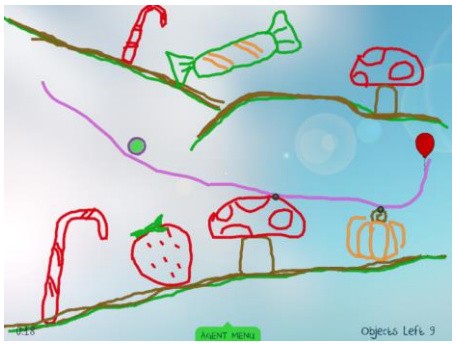

Figure 1: Screenshot of Physics Playground

Students can choose to enter one of seven different playgrounds, and then play any of the 10 or so levels within that playground. Each level consists of various obstacles scattered around the space, as well as a balloon positioned at different locations within the space (see Figure 1). Students can nudge the ball left and right, but will need to create simple machines (called "agents of force and motion" in the game) on-screen in order to solve the problems presented in the playgrounds. There are four possible agents that may be created: ramps, pendulums, levers and springboards. Students can also create fixed points along a line drawing to create pivots for the agents they create. Students use the mouse to draw agents that come to life after being drawn, and use them to propel the ball to the red balloon. Students control the weight and density of objects through their drawings, making an object denser, for example, by filling it with more lines.

Each level allows multiple solutions, encouraging students to experiment with various methods to achieve the goal and guide the ball towards the balloon. Trophies are awarded both for achieving the goal objective and for solutions deemed particularly elegant or creative, encouraging students to attempt each playground more than once. This unstructured game-like environment provides us with a rich setting in which to examine the patterns of students' affect and behavior as they interact with the game platform.

3. DATA COLLECTION

Students in the 8th and 9th grade were selected due to the alignment of the curriculum in Physics Playground to the state standards at those grade levels. The student sample consisted of 137 students (57 male, 80 female) who were enrolled in a public school in the Southeastern U.S. Each group of about 20 students used Physics Playground during 55-minute class periods over the course of four days.

An online physics pretest (administered at the start of day 1) and posttest (administered at the end of day 4), measured student knowledge and skills related to Newtonian physics. In this paper, our focus is on data collected during days 2 and 3, during which time students were participating in two full sessions of game play.

The study was conducted in a computer-enabled classroom with 30 desktop computers. Inexpensive webcams ($30 each) were affixed at the top of each computer monitor. 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 from the experimenters, who were available to answer any additional questions and to troubleshoot any problems.

3.1 Field Observations

Students were observed by two BROMP-certified observers while using the Physics Playground software. The Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP 2.0) is a momentary time sampling system that has been used to study behavioral and affective indicators of student engagement in a number of learning environments [9]. BROMP coders observe each student individually, in a predetermined order. They record only the first behavior and affect that the student displays, but they have up to 20 seconds to make a determination about what that might be.

In this study, BROMP coding was done by the 5th author and the 3rd author. The 5th author, a co-developer of BROMP, has been validated to achieve acceptable inter-rater reliability (kappa >= 0.60) with over a dozen other BROMP-certified coders. The 3rd author achieved sufficient inter-rater reliability (kappa >= 0.60) with the 5th author on the first day of this study.

The coding process was implemented using the Human Affect Recording Tool (HART) application for Android devices [6], which enforces the protocol while facilitating data collection. The study used coding schema that had previously been used in several other studies of student engagement [e.g. 17], and included *boredom*, *confusion*, *engaged concentration*, and *frustration* (affective states) as well as *on task*, *on-task conversation*, and *off-task* (behavioral states). Consistent with previous BROMP research, "*?*" was recorded when a student could not be coded, when an observer was unable to identify the

student's behavior or affective state, or when the affect/behavior of the student was clearly a construct outside of the coding scheme (such as *anger*).

Modifications to the affective coding scheme were made on the third day of the study, when *delight* and *dejection* were added. *Delight* had been coded in previous studies, and was ultimately used to construct detectors. *Dejection*, defined as a state of being saddened, distressed, or embarrassed by failure [9], is likely the affect that corresponds with the experience of *stuck* [11,20]. Because it had not been coded in previous research, and because it was still quite rare in Physics Playground, it was not modeled for this study.

3.2 Affect and Behavior Incidence

An initial number of 2,374 observations were made across all 137 students during the course of the study. Only affect observations on the second and third days were used in the construction of the detectors, since the first and last days mostly consisted of pretests and posttests. Other observations were dropped as a result of two students who switched computers halfway through data collection, resulting in each student being logged under the other student's ID for part of the study. The remaining 2,087 observations recorded during the second and third days were used in the construction of both detectors. Of these 2087 observations, an additional 214 were removed prior to the construction of the interaction-based detectors and 863 were removed prior to the construction of the video-based detectors. Because the criteria for these exclusions were methodologically based, further details are provided in the sections describing the construction of each detector.

Within the field observations, the most common affective state observed was *engaged concentration* with 1293 instances (62.0%), followed by *frustration* with 235 instances (11.3%). *Boredom and confusion* were far less frequent despite being observed across both second and third days of observation: 66 instances (3.2%) for *boredom* and 38 instances (1.8%) for *confusion*. *Delight* was only coded on the third day, and was also rare (45 instances), but it still comprised 2.2% of the total observations.

The frequency of off-task behavior observations was 4.0% (84 instances), which was unusually low compared to prior classroom research in the USA using the same method with other educational technologies [26,32]. On-task conversation was seen 18.6% of the time (388 instances).

4. INTERACTION-BASED DETECTORS

To create interaction affect detectors, BROMP affect observations were synchronized to the log files of student interactions with the software. Features were then generated and a 10-fold student-level cross validation process was applied for machine learning, using five classification algorithms.

4.1 Feature Engineering

The feature engineering process for this study was based largely on previous research on student engagement, learning, and persistence. The initial set of features comprised 76 gameplay attributes that potentially contain evidence for specific affective states and behavior. Some attributes included:

- The total number of springboard structures created in a level
- The total number of freeform objects drawn in a level
- The amount of time between start to end of a level
- The average number of gold and silver trophies obtained in a level
- The number of stacking events (gaming behavior) in a level

Features created may be grouped into two broad categories. Time-based features focus on the amount of time elapsed between specific student actions, such as starting and pausing a level, as well as the time it takes for a variety of events to occur within each playground level. Other features take into account the number of specific objects drawn or actions and events occurring during gameplay, given various conditions.

Missing values were present at certain points in the dataset when a particular interaction was not logged. For example, a feature specifying the amount of time between the student beginning a level and his/her first restart of the level, would contain a missing value if the student manages to complete a level without having to restart it. A variety of data imputation approaches were used in these situations to fill in the missing values so that we could retain the full sample size. We used single, average and zero imputation methods to fill in the missing data, and ran the new datasets through the machine learning process to identify the best data imputation strategy for each affect detector. Zero imputations were performed where the missing values were replaced by the value 0, while average data imputations took place when the average value for the particular feature was computed, and the missing values replaced by this average value. In single data imputation, we used RapidMiner to build an M5' model [31], a tree-based decision model, to predict the values for each feature, and applied the model to compute a prediction of the missing value. We also ran the original dataset without any imputation through any of the classification algorithms that allowed it.

Of the 2087 BROMP field observations that were collected, 214 instances were removed as most of these instances corresponded to times when the student was inactive. Additional instances were removed where the observer recorded a *?*, the code used when BROMP observers cannot identify a specific affect or behavior or when students are not at their workstation. As a result, these instances did not contribute to the building of the respective affect and behavior detectors.

4.2 Machine Learning

Data collection was followed by a multi-step process to develop interaction-based detectors of each affect. A two-class approach was used for each affective state, where that affective state was discriminated from all others. For example, engaged concentration 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, and 2) both on task behaviors and on task conversation related to the game.

4.2.1 Resampling of Data

Because observations of several of the constructs included in this study were infrequent, (< 5.0% of the total number of observations), there were large class imbalances in our data distributions. To correct for this, we used the *cloning* method for resampling, generating copies of respective positive affect on the training data, in order to make class frequency more balanced for detector development.

4.2.2 Feature Selection and Cross-Validation

Correlation-based filtering was used to remove features that had very low correlation with the predicted affect and behavior constructs (correlation coefficient > 0.04) from the initial feature

set. Feature selection for each detector was then conducted using forward selection.

Detectors for each construct were built in the RapidMiner 5.3 data-mining software, using common classification algorithms that have been previously shown to be successful in building affect detectors: JRip, J48 decision trees, KStar, Naïve-Bayes, step and logistic regression. Models were validated using 10-fold student-level batch cross-validation. The performance metric of A' was computed on the original, non-resampled, datasets.

4.3 Selected Features

From the forward selection process, a combination of features was selected in each of the affect and behavior detectors that provide some insight into the type of student interactions that predict the particular affective state or behavior.

The features for *boredom* involve a student spending more time between actions on average. A bored student would also expend less effort to guide the ball object to move in the right direction, as indicated by fewer nudges made on the ball object to move it, and more ball objects being lost from the screen.

The features that predict *confusion* are characterized by a student spending more time before his/her first nudge to make the ball object move, and drawing fewer objects in a playground level. A student who is confused may not have known how to draw and move the ball object towards the balloon, thus spending a long time within a certain level and resulting in a lower number of levels attempted in total.

From the features selected, *delight* appears to ensue from some indicator of success, such as a student who is able to achieve a silver trophy earlier on during gameplay, and who completes more levels in total. We can also portray the student who experiences *delight* as someone who was able to achieve the objective without having to make multiple attempts to draw the relevant simple machines (such as springboards and pendulums).

The features for *engaged concentration* would describe a student who is able to complete a level in fewer attempts but erases the ball object more often during each attempt, indicating that the student was putting in more effort to refine his/her strategies within a single attempt at the level. *Engaged concentration* would also depict a student who has experienced success during gameplay and achieved a silver trophy in a shorter than average time, perhaps because of his/her focused efforts during each attempt.

Table 1. Features in the final interaction-based detectors of each construct

Affect/ Behavior	Selected features
Boredom	Time between actions within a level
	Total number of objects that were "lost" (i.e. Moved off the screen)
	Total number of nudges made on the ball object to move it
Confusion	Amount of time spent before the ball object was nudged to move
	Total number of levels attempted
	Total number of objects drawn within the level
Delight	Number of silver trophies achieved
	Consecutive number of pendulums and springboards created
	Total number of levels attempted
	Total number of levels completed successfully
Engaged Concentration	Total number of silver trophies achieved in under the average time
	Total number of level re-starts within a playground
	Total number of times a ball object was erased consecutively
Frustration	Total number of silver trophies achieved in under the average time
	Total number of level re-starts within a playground
	Total number of levels completed successfully
	Total number of levels attempted
Off-task Behavior	Time spent in between each student action
	Total number of pauses made within a level
	Total number of times a student quits a level without completing the objective and obtaining a trophy

Unlike *engaged concentration*, a student who experiences *frustration* failed to achieve the objective and achieved fewer silver trophies within the average time taken. Student *frustration*, as seen in the features, would also result in the student having to make more attempts at a level due to repeated failure, thus resulting in fewer levels attempted in total.

Lastly, behavior that is *off-task* involves a student who spends more time pausing the level or between actions as a whole. It is also apparent in a student who draws fewer objects and quits more levels without completing them, implying that he or she did not put in much effort to complete the playground levels.

5. VIDEO-BASED DETECTORS

The video-based detectors have been reported in a recent publication [10]. In the interest of completeness, the main approach is re-presented here. There are also small differences in the results reported here due to a different validation approach that was used to make meaningful comparisons with interaction-based detectors.

Video-based affect detectors were constructed using FACET (http://www.emotient.com/products), a commercialized version of the Computer Expression Recognition Toolbox (CERT) software (CITE). FACET is a computer vision tool used to automatically detect Action Units (AUs), which are labels for specific facial muscle activations (e.g. lowered brow). 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.

5.1 Feature Engineering

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 lowered) 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.

We 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 [17]. 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.

Of the initial 2087 instances available for us to train our video-based detectors on, about a quarter (25%) were discarded because FACET was not able to register the face and thus could not estimate the presence of AUs and computation of features. 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, culminating in 1224 instances where we had sufficient video data to train our affect models on.

5.2 Machine Learning

We also built separate detectors for each affective state similar to the interaction-based detectors. Building individual detectors for each state allows the parameters (e.g., window size, features used) to be optimized for that particular affective state.

5.2.1 Resampling of Data

Like the interaction-based detectors, there were large class imbalances in the affective and behavior distributions. Two sampling techniques, different from the one used in the building of interaction-based detectors, were used on the training data to compensate for this imbalance. These two techniques included downsampling (removal of random instances from the majority class) and synthetic oversampling (with SMOTE; [13]) 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.

5.2.2 Feature Selection and Cross-Validation

We used tolerance analysis to eliminate features with high multicollinearity (variance inflation factor > 5) [2]) for video-based detectors. Feature selection was then used to obtain a more diagnostic set of features for classification. RELIEF-F [24] 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).

We then built classification models 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 [23].

6. RESULTS

We evaluated the extent to which the detectors for each construct are able to identify their respective affect. Both detectors were evaluated using a 10-fold student-level batch cross-validation. In this process, students in the training dataset are randomly divided into ten groups of approximately equal size. A detector is built using data from all possible combinations of 9 out of the overall 10 groups, and finally tested on the last group. Cross-validation at this level increases the confidence that the affect and behavior detectors will be more accurate for new students. To ensure comparability between the two sets of detectors, the cross-validation process was carried out with the same randomly selected groups of students.

Table 2. A' performance values for affect and behavior using video-based and interaction-based detectors

Affect/Behavior Construct	Interaction-Based Detectors				Video-Based Detectors		
	Classifier	Data Imputation Scheme	A'	No. Instances	Classifier	A'	No. Instances
Boredom	Logistic regression	Zero	0.629	1732	Classification via Clustering	0.617	1305
Confusion	Step regression	Average	0.588	1732	Bayes Net	0.622	1293
Delight	Logistic regression	None	0.679	1732	Updateable Naïve Bayes	0.860	1003
Engaged Concentration	Naïve Bayes	Zero	0.586	1732	Bayes Net	0.658	1228
Frustration	Logistic regression	Average	0.559	1732	Bayes Net	0.632	1132
Off-Task behavior	Step regression	Zero	0.765	1829	Logistic Regression	0.780	1381

Detector performance was assessed using A' values that were computed as the Wilcoxon statistic [22]. A' is the probability that the given algorithm will correctly identify whether an observation is an example of a specific affective state. A' can be approximated by the Wilcoxon statistic and is equivalent to the area under the Receiver Operating Characteristic (ROC) curve in signal detection theory. A detector with a performance of A' = 0.5 is performing at chance, while a model with a performance of A' = 1.0 is performing with perfect accuracy.

Table 2 shows the performance of the two detector suites. Both interaction-based and video-based detectors' performance over all six affective and behavior constructs was better than chance (A' = 0.50). On average, the interaction-based detectors yielded an A' of 0.634 while the video-based detectors had an average A' of 0.695. This difference can be mainly attributed to the detection of delight, which was much more successful for the video-based detectors. Accuracy of the two detector suites was much more comparable for the other constructs, though the video-based detectors showed some advantages for engaged concentration and frustration, and were higher for 5 of the 6 constructs.

The majority of the video-based detectors performed the best when using the Bayes Net classifier, except for *boredom*, *delight* and *off-task behavior*. In comparison, logistic and step regression composed the classifiers that produced the best performance for most of the interaction-based detectors, with the exception of *engaged concentration*.

7. DISCUSSION

Affect detection is becoming an important component in educational software, which aims to improve student outcomes by dynamically responding to student affect. Affect detectors have been successfully built and implemented via different modalities [3,16,40], and each have their own advantages and disadvantages when implemented in a noisy classroom environment. This study is an extension of previous research conducted on both video-based and interaction-based detectors. Having been mostly built in controlled laboratory settings [12], we now test the performance for video-based detectors within an uncontrolled computer-enabled classroom environment that is more representative of an authentic educational setting. Although interaction-based detectors have been built to some degree of success in whole classroom settings [5,7,28], we now test the performance of these affect detectors in an open-ended and exploratory educational game platform.

In this paper, we compared the performances of six video-based and interaction-based detectors on student affect and behavior in the game-based software. We will discuss the implications of these comparisons in this section, as well as future work.

7.1 Main Findings

The performances of both detectors in the six affects and off-task behavior appear to be at similar levels above chance for five of the constructs, with video-based detectors performing slightly better than interaction-based detectors on the whole, and with video-based detector showing a stronger advantage for delight. Several factors may have help to explain the relative performances.

Performance of video detectors could be influenced by the uncontrolled whole-classroom setting in which video data is collected, where there are higher chances of video data being absent or compromised due to unpredictable student movement. While there were initially 2,087 instances of affect and behavior observed and coded, a moderate proportion of facial data instances were dropped from the final dataset when building the models. There were 44 instances of affect observation that were dropped either because the video was corrupted or incomplete, or because no video was recorded at all. In addition, there were 520 instances where video was recorded, but facial data were not detected for some reason, perhaps because the student had left the workstation, or when the face could not be detected in the video. An additional 211 instances were removed even though facial data was detected, because the facial data recorded was present for less than 1 second, such that no features could be calculated.

For interaction-based detectors, the exploratory and open-ended user-interface [39] constitutes a unique challenge in creating accurate models for student affect and behavior. The open-ended interface included multiple goals and several possible solutions that students could come up with to successfully complete each level. During gameplay, there are also multiple factors that could contribute to a student's failure to complete a level, such as conceptual knowledge as well as implementation of appropriate objects. A student with accurate conceptual knowledge of simple machines and Newtonian physics may still fail the level because of problems implementing the actions needed to guide the ball to the target. On the other hand, a student with misconceptions about the relevant physics topics may nevertheless be able to complete the level successfully through systematic experimentation. The possible combinations of student actions that result in failure or success in a playground level would hence contribute to the lower accuracy of interaction-based detectors on identifying students' affect based on their interactions with the software.

Another issue with the Physics Playground software could be that there are fewer indicators of success per unit of time, as compared to other learning software that have been studied previously, such as the Cognitive Tutors [e.g. 5]. During gameplay, the system is able to recognize when combinations of objects the student draws forms an eligible agent. However, this indicator of success or failure is not apparent to the student until after he or she creates the ball object and applies a relevant force to trigger a simulation. Since students often spend at least several minutes building agents and ball objects, this results in coarser-grained indicators and evaluations of success and failure. This is in comparison to affect detectors created in previous studies for the Cognitive Tutor software, in which there was regular evaluation of each question attempted, thus resulting in more indicators of success over a given time period. The combination of open-endedness and lack of success indicators per unit of time consequently leads to greater difficulty translating the semantics of student-software interactions into accurate affect predictions.

When comparing between the two sets of detectors, physical detectors make direct use of students' facial features and bodily movements captured by webcams and constitute embodied representations of students' affective states. On the other hand, interaction detectors were built based on student actions within the software, which serves as an indirect proxy of the students' actual affective states. These detectors rely, therefore on the degree to which student interactions with the software are influenced (or not) by the affective states they experience. Perhaps not surprisingly, video-based detectors perform somewhat better in predicting some affective states (e.g., delight, engaged concentration, and frustration). Although the video detectors are limited by missing data, interaction-based detectors can only detect something that causes students to change their behaviors within the software, which can be challenging given the issues arising from the open-ended game platform. Simply put, face-

based affect detectors appear to provide more accurate affect estimates but in fewer situations, while interaction-based affect detectors provide less accurate estimates, but are applicable in more situations. The two approaches thus appear to be quite complementary.

7.2 Limitations

In comparing the performances between interaction and video-based detectors, there exist several limitations in ensuring an equivalent set of methods for a fair comparison to be made.

Although both types of detectors were built based on the same ground truth data, varying sets of limitations exist that are unique to each set of detectors. A smaller proportion of instances were retained to build video-based detectors due to missing video data, which may influence performance comparison. Interaction-based detectors, on the other hand, are relatively more sensitive to the type of educational platform it is built upon, as compared to video-based detectors. The type of learning platform thus affects the variety of features that are relevant and useful in building the affect and behavior detectors, which in turn impacts its performance relative to previous work.

For both detectors, the sample size available for some of the affective states was quite limited, which made it necessary to oversample the training data in order to compensate for the class imbalances. However, because each detector was built on different platforms, different methods were used in oversampling the datasets. The need to conduct data imputations was also unique to interaction-based detectors due to the nature of some of the computed features, and not required for video-based detectors. The difference in these methods may in turn affect performance comparison between the two types of detectors.

7.3 Concluding Remarks

Given the various advantages and limitations to each type of detector in accurately predicting student affect, it may be beneficial for affect detection strategies to include a combination of video-based and interaction-based detectors. While video-based detectors provide more direct measures of student affect, practical issues may lead to video data being absent or unusable in detecting affect, simply because there is no facial data available to detect affect in. These situations may be alleviated by the presence of interaction data that are recorded automatically during students' use of the software. On the other hand, video-based facial data would be able to provide support to interaction data and boost the accuracy in which affective states are detected among students. This form of late-fusion or decision-level fusion can also be complemented by early-fusion or feature-level fusion, where features from both modalities are combined prior to classification. Whether this leads to improved accuracy, as routinely documented in the literature on multimodal affect detection [15,16] awaits future work.

8. 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.

9. REFERENCES

[1] Ai, H., Litman, D.J., Forbes-Riley, K., Rotaru, M., Tetreault, J., and Purandare, A. 2006. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 797–800.

[2] Allison, P.D. 1999. *Multiple regression: A primer.* Pine Forge Press.

[3] AlZoubi, O., Calvo, R. a., and Stevens, R.H. 2009. Classification of EEG for affect recognition: An adaptive approach. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5866 LNAI*, 52–61.

[4] Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K., and Christopherson, R. 2009. Emotion sensors go to school. *Frontiers in Artificial Intelligence and Applications*, 17–24.

[5] Baker, R., Gowda, S., and Wixon, M. 2012. Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *Proceedings of the 5th International Conference on Educational Data Mining*, 126–133.

[6] Baker, R., Gowda, S., Wixon, M., et al. 2012. Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *Proceedings of the 5th International Conference on Educational Data Mining*, 126–133.

[7] Baker, R.S., Ocumpaugh, J., Gowda, S.M., Kamarainen, A.M., and Metcalf, S.J. 2014. Extending Log-Based Affect Detection to a Multi-User Virtual Environment for Science. *22nd Conference on User Modeling, Adaptation and Personalization (UMAP 2014)*, 290–300.

[8] Baker, R.S.J.D. and Yacef, K. 2009. The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining 1*, 1, 3–16.

[9] Baker, R.S.; Ocumpaugh, J. 2015. Interaction-Based Affect Detection in Educational Software. In R.A. Calvo, S.K. D'Mello, J. Gratch and A. Kappas, eds., *Handbook of Affective Computing*. Oxford University Press, Oxford, UK, 233–245.

[10] Bosch, N., Mello, S.D., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., Zhao, W. Automatic Detection of Learning - Centered Affective States in the Wild. *In Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015). New York, NY, USA: ACM.*

[11] Burleson, W. and Picard, R.W. 2004. Affective agents: Sustaining motivation to learn through failure and a state of stuck. *Proceedings of the Workshop on Social and Emotional Intelligence in Learning Environments in conjunction with the seventh International Conference on Intelligent Tutoring Systems*.

[12] Calvo, R.A.. and D'Mello, S.K. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and their Application to Learning Environments. *IEEE Transactions on Affective Computing 1*, 1, 18–37.

[13] Chawla, N. V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. 2011. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16*, 321–357.

[14] D'Mello, S., Jackson, T., Craig, S., et al. 2008. AutoTutor Detects and Responds to Learners Affective and Cognitive States. *Proceedings of the Workshop on Emotional and*

Cognitive issues in ITS in conjunction with the 9th International Conference on Intelligent Tutoring Systems, 31–43.

[15] D'Mello, S. and Kory, J. A Review and Meta-Analysis of Multimodal Affect Detection. *ACM Computing Surveys*.

[16] D'Mello, S. and Kory, J. 2012. Consistent but Modest: A Meta-Analysis on Unimodal and Multimodal Affect Detection Accuracies from 30 Studies. *ACM International Conference on Multimodal Interaction*, 31–38.

[17] D'Mello, S. 2011. Dynamical emotions: bodily dynamics of affect during problem solving. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

[18] D'Mello, S.K. and Graesser, A. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modelling and User-Adapted Interaction 20*, 2, 147–187.

[19] D'Mello, S.K. 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.

[20] D'Mello, S.K.; Graesser, A. 2012. Emotions During Learning with AutoTutor. In *Adaptive Technologies for Training and Education*. 169–187.

[21] Dragon, T., Arroyo, I., Woolf, B.P., Burleson, W., El Kaliouby, R., and Eydgahi, H. 2008. Viewing student affect and learning through classroom observation and physical sensors. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5091 LNCS*, 29–39.

[22] Hanley, J.A. and Mcneil, B.J. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology 143*, 29–36.

[23] Holmes, G., Donkin, A., and Witten, I.H. 1994. WEKA: a machine learning workbench. *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference*, 357–361.

[24] Kononenko, I. 1994. Estimating attributes: Analysis and extensions of RELIEF. In F. Bergadano and L. De Raedt, eds., *Machine Learning: ECML-94*. Springer, Berlin Heidelberg, 171–182.

[25] Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., and Heffernan, C. 2014. Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology 45*, 3, 487–501.

[26] Ocumpaugh, J., Baker, R.S.J., Gaudino, S., Labrum, M.J., and Dezendorf, T. 2013. Field Observations of Engagement in Reasoning Mind. *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, 624–627.

[27] Pantic, M., Pantic, M., Rothkrantz, L.J.M., and Rothkrantz, L.J.M. 2003. Toward an Affect-Sensitive Multimodal Human Computer Interaction. *Proceedings of the IEEE 91*, 9, 1370–1390.

[28] Paquette, L., Baker, R.S.J. d., Sao Pedro, M., et al. 2014. Sensor-Free Affect Detection for a Simulation-Based Science Inquiry Learning Environment. *Proceedings of the 12th International Conference on Intelligent Tutoring Systems, ITS 2014*, 1–10.

[29] Pardos, Z. a., Baker, R.S.J. d., San Pedro, M.O.C.Z., Gowda, S.M., and Gowda, S.M. 2014. Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. *Journal of Learning Analytics 1*, 1, 107–128.

[30] Picard, R.W. 1997. *Affective Computing.* MIT Press, Cambridge, MA.

[31] Quinlan, J.R. 1992. Learning with continuous classes. *Machine Learning 92*, 343–348.

[32] Rodrigo, M., Baker, R., and Rossi, L. 2013. Student Off-Task Behavior in Computer-Based Learning in the Philippines: Comparison to Prior Research in the USA. *Teachers College Record 115*, 10, 1–27.

[33] Rodrigo, M.M.T. and Baker, R.S.J. d. 2009. Coarse-grained detection of student frustration in an introductory programming course. *Proceedings of the fifth international Computing Education Research Workshop - ICER 2009*.

[34] Sabourin, J., Mott, B., and Lester, J. 2011. Modeling Learner Affect with Theoreticallly Grounded Dynamic Bayesian Networks. *Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction*, 286–295.

[35] San Pedro, M.O.Z., Baker, R.S.J. d., Bowers, A.J., and Heffernan, N.T. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. *Proceedings of the 6th International Conference on Educational Data Mining*, 177–184.

[36] Sebe, N., Cohen, I., Gevers, T., and Huang, T.S. 2005. Multimodal Approaches for Emotion Recognition: A Survey. *Proceedings of SPIE – The International Society for Optical Engineering*, 56–67.

[37] Shute, V., Ventura, M., and Kim, Y.J. 2013. Assessment and Learning of Qualitative Physics in Newton ' s Playground Newton ' s Playground. *The Journal of Educational Research 29*, 579–582.

[38] Shute, V. and Ventura, M. 2013. *Measuring and Supporting Learning in Games Stealth Assessment.* MIT Press, Cambridge, MA.

[39] Shute, Valerie; Ventura, Matthew; Kim, Y.J. 2013. Assessment and Learning of Qualitative Physics in Newton's Playground. *Journal of Educational Research 106*, 423–430.

[40] Zeng, Z., Pantic, M., Roisman, G.I., and Huang, T.S. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 39–58.