Sensor-Free or Sensor-Full: A Comparison of Data Modalities in Multi-Channel Affect Detection

Luc Paquette
Teachers College Columbia University
525 West 120th Street
New York, NY 10027
paquette@tc.columbia.edu

Jonathan Rowe
North Carolina State University
Raleigh, NC 27695
jprowe@ncsu.edu

Ryan Baker
Teachers College Columbia University
525 West 120th Street
New York, NY 10027
ryanshaunbaker@gmail.com

Bradford Mott
North Carolina State University
Raleigh, NC 27695
bwmott@ncsu.edu

James Lester
North Carolina State University
Raleigh, NC 27695
lester@ncsu.edu

Jeanine DeFalco
Teachers College Columbia University
525 West 120th Street
New York, NY 10027
jad2234@tc.columbia.edu

Keith Brawner
US Army Research Lab
Orlando, FL, USA
keith.w.brawner@mail.mil

Robert Sottilare
US Army Research Lab
Orlando, FL, USA
robert.sottilare@us.army.mil

Vasiliki Georgoulas
Teachers College Columbia University
525 West 120th Street
New York, NY 10027
vasiliki.georgoulas@usma.edu

ABSTRACT

Computational models that automatically detect learners' affective states are powerful tools for investigating the interplay of affect and learning. Over the past decade, affect detectors—which recognize learners' affective states at run-time using behavior logs and sensor data—have advanced substantially across a range of K-12 and postsecondary education settings. Machine learning-based affect detectors can be developed to utilize several types of data, including software logs, video/audio recordings, tutorial dialogues, and physical sensors. However, there has been limited research on how different data modalities combine and complement one another, particularly across different contexts, domains, and populations. In this paper, we describe work using the Generalized Intelligent Framework for Tutoring (GIFT) to build multi-channel affect detection models for a serious game on tactical combat casualty care. We compare the creation and predictive performance of models developed for two different data modalities: 1) software logs of learner interactions with the serious game, and 2) posture data from a Microsoft Kinect sensor. We find that interaction-based detectors outperform posture-based detectors for our population, but show high variability in predictive performance across different affect. Notably, our posture-based detectors largely utilize predictor features drawn from the research literature, but do not replicate prior findings that these features lead to accurate detectors of learner affect.

Keywords

Affect detection, multimodal interaction, posture, serious games.

1. INTRODUCTION

Affect is critical to understanding learning. However, the interplay between affect and learning is complex. Some affective states, such as boredom, have been shown to coincide with reduced learning outcomes ([25]). Other affective states, such as confusion and engaged concentration, have been found to serve beneficial roles ([14], [24]). The ability to detect a learner's affective state while she interacts with an online learning environment is critical for adaptive learning technologies that aim to support and regulate learners' affect ([26]).

Research on affective computing has enabled the development of models that automatically detect learner affect using a wide variety of data modalities (see extensive review in [8]). Many researchers have focused on physical sensors, because of their capacity to capture physiological and behavioral manifestations of emotion, potentially regardless of what learning system is being used. Sensor-based detectors of affect have been developed using a range of physical indicators including facial expressions ([2], [7]), voice [35], posture ([11], [16]), physiological data [22] and EEG [1]. Despite this promise, deploying physical sensors in the classroom is challenging, and sometimes prohibitive [6], and efforts in this area are still ongoing, with some researchers arguing that this type of affect detection has not yet reached its full potential [13].

In recent years, efforts have also been made towards the development of complementary affect detection techniques that recognize affect solely from logs of learner interactions with an online learning environment ([2], [3], [24]). Initial results in this area have shown considerable promise. As both sensor-based and interaction-based affect detectors continue to mature, efforts are needed to compare the relative advantages of each approach. An early comparison was seen in D'Mello et al. [15], but considerable progress has been made in the years since.

In this paper, we compare the performance and the general process of developing models for affect detection using two different data modalities: learner interaction logs and posture data from a Microsoft Kinect sensor. Ground-truth affect data for detector development was collected through field observation [23] of learners interacting with vMedic, a serious game on tactical combat casualty care, integrated into the General Intelligent Framework for Tutoring (GIFT) [32]. Findings suggest that interaction-based affect detectors outperform posture-based detectors for our population. However, interaction-based detectors show high variability in predictive performance across different emotions. Further, our posture-based detectors, which utilize many of the same predictor features found throughout the research literature, achieve predictive performance that is only slightly better than chance across a range of affective states, a finding that is contrary to prior work on sensor-based affect detection.

2. DATA

Three sources of data were used in this work: 1) log file data produced by learners using the vMedic (a.k.a. TC3Sim) serious game, 2) Kinect sensor log data, and 3) quantitative field observations of learner affect using the BROMP 1.0 protocol [23]. This section describes those sources of data, by providing information on the learning environment, study participants, and research study method.

2.1 Learning System and Subjects

We modeled learner affect within the context of vMedic, a serious game used to train US Army combat medics and lifesavers on tasks associated with dispensing tactical field care and care under fire (Figure 1). vMedic has been integrated with the Generalized Intelligent Framework for Tutoring (GIFT) [32], a software framework that includes a suite of tools, methods, and standards for research and development on intelligent tutoring systems and affective computing.

Game-based learning environments, such as vMedic, enable learners to interact with virtual worlds, often through an avatar, and place fewer constraints on learner actions than many other types of computer-based learning environments ([3], [19], [24]). Some virtual environments place more constraints on learner behavior than others. For example, learning scenarios in vMedic are structured linearly, presenting a fixed series of events regardless of the learner's actions. In contrast, game-based learning environments such as EcoMUVE [20] and Crystal Island [29] afford learners considerable freedom to explore the virtual world as they please. While vMedic supports a considerable amount of learner control, its training scenarios focus participants' attention on the objectives of the game (e.g., administering care), implicitly guiding learner experiences toward key learning objectives.

To investigate interaction-based and sensor-based affect detectors for vMedic, we utilize data from a study conducted at the United States Military Academy (USMA). There were 119 cadets who participated in the study (83% male, 17% female). The participants were predominantly first-year students. During the data collection, all participants completed the same training module. The training module focused on a subset of skills for tactical combat casualty care: care under fire, hemorrhage control, and tactical field care. The study materials, including pre-tests, training materials, and post-tests, were administered through GIFT. At the onset of each study session, learners completed a content pre-test on tactical combat casualty care. Afterward, participants were presented with a PowerPoint presentation about tactical combat casualty care. After completing the PowerPoint, participants completed a series of training scenarios in the vMedic serious game where they applied skills, procedures, and knowledge presented in the PowerPoint. In vMedic, the learner adopts the role of a combat medic faced with a situation where one (or several) of her fellow soldiers has been seriously injured. The learner is responsible for properly treating and evacuating the casualty, while following appropriate battlefield doctrine. After the vMedic training scenarios, participants completed a post-test, which included the same series of content assessment items as the pre-test. In addition, participants completed two questionnaires about their experiences in vMedic: the Intrinsic Motivation Inventory (IMI) [30] and Presence Questionnaire [34]. All combined study activities lasted approximately one hour.

During the study, ten separate research stations were configured to collect data simultaneously; each station was used by one cadet at a time. Each station consisted of an Alienware laptop, a Microsoft Kinect for Windows sensor, and an Affectiva Q-Sensor, as well as a mouse and pair of headphones. The study room's layout is shown in Figure 2. In the figure, participant stations are denoted as ovals. Red cones show the locations of Microsoft Kinect sensors, as well as the sensors' approximate fields of view. The dashed line denotes the walking path for the field observers.

Kinect sensors recorded participants' physical behavior during the study, including head movements and posture shifts. Each Kinect sensor was mounted on a tripod and positioned in front of a participant (Figure 2). The Kinect integration with GIFT provided four data channels: skeleton tracking, face tracking, RGB (i.e., color), and depth data. The first two channels leveraged built-in tracking algorithms (which are included with the Microsoft Kinect for Windows SDK) for recognizing a user's skeleton and face, each represented as a collection of 3D vertex coordinates. The RGB channel is a 640x480 color image stream comparable to a standard web camera. The depth channel is a 640x480 IR-based image stream depicting distances between objects and the sensor.

Q-Sensors recorded participants' physiological responses to events during the study. The Q-Sensor is a wearable arm bracelet that measures participants' electrodermal activity (i.e., skin conductance), skin temperature, and its orientation through a built-in 3-axis accelerometer. However, Q-Sensor logs terminated prematurely for a large number of participants, necessitating additional work to determine the subset of field observations that are appropriate to predict with Q-Sensor-based features. Inducing Q-Sensor-based affect detectors will be an area of future work.

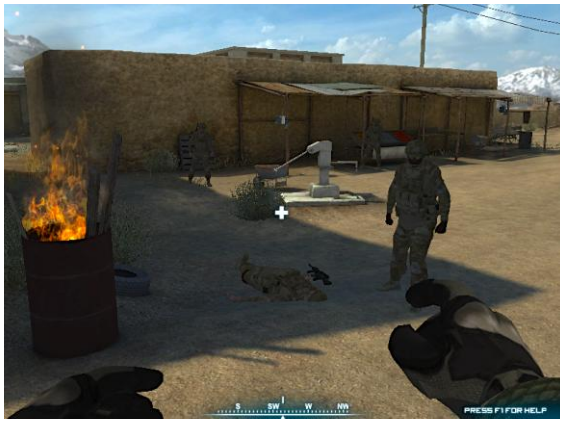

Figure 1. vMedic learning environment.

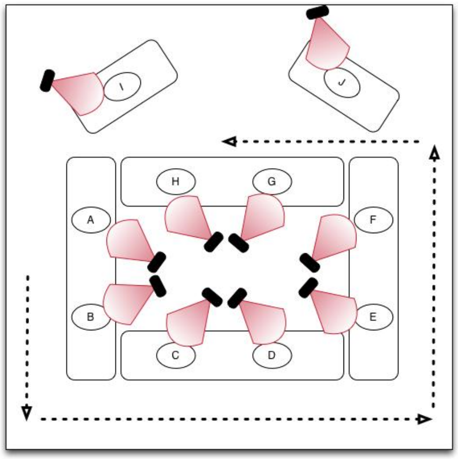

Figure 2. Study room layout.

2.2 Quantitative Field Observations (QFOs)

We obtain ground-truth labels of affect using Quantitative Field Observations (QFOs), collected using the Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP) [23]. This is a common practice for interaction-based detection of affect (e.g. [3], [24]). Much of the work to date for video-based affect detection, by contrast, has focused on modeling emotion labels that are based on self-reports ([10], [16]), or labels obtained through retrospective judgments involving freeze-frame video analysis [11]. It has been argued that BROMP data is easier to obtain and maintain reliability for under real-world conditions than these alternate methods [23], being less disruptive than self-report, and easier to gain full context than video data.

To be considered BROMP-certified, a coder must achieve inter-rater reliability of Kappa >= 0.6 with a previously BROMP-certified coder. BROMP has been used for several years to study behavior and affect in educational settings ([3], [4], [27]), with around 150 BROMP-certified coders as of this writing, and has been used as the basis for successful automated detectors of affect ([3], [24]). Observations in this study were conducted by two BROMP-certified coders, the 2nd and 6th authors of this paper.

Within the BROMP protocol, behavior and affective states are coded separately but simultaneously using the Human Affect Recording Tool (HART), an application developed for the Android platform (and freely available as part of the GIFT distribution). HART enforces a strict coding order determined at the beginning of each session. Learners are coded individually, and coders are trained to rely on peripheral vision and side glances in order to minimize observer effects. The coder has up to 20 seconds to categorize each trainee's behavior and affect, but records only the first thing he or she sees. In situations where the trainee has left the room, the system has crashed, where his or her affect or behavior do not match any of the categories in the current coding scheme, or when the trainee can otherwise not be adequately observed, a '?' is recorded, and that observation is eliminated from the training data used to construct automated detectors.

In this study, the typical coding scheme used by BROMP was modified to accommodate the unique behaviors and affect that was manifest for this specific cadet population and domain. Affective states observed included frustration, confusion, engaged concentration, boredom, surprise and anxiety. Behavioral categories consisted of on-task, off-task behaviors, Without Thinking Fastidiously behavior [33], and intentional friendly fire (these last two categories will not be discussed in detail, as they were rare).

In total, 3066 BROMP observations were collected by the two coders. Those observations were collected over the full length of the cadets' participation in the study, including when they were answering questionnaires on self-efficacy, completing the pre and post-tests, reviewing PowerPoint presentations, and using vMedic. For this study, we used only the 755 observations that were collected while cadets were using vMedic. Of those 755 observations, 735 (97.35%) were coded as the cadet being on-task, 19 (2.52%) as off-task, 1 (0.13%) as Without Thinking Fastidiously, and 0 as intentional friendly fire. Similarly, 435 (57.62%) of the affect labels were coded as concentrating, 174 (23.05%) as confused, 73 (9.67%) as bored, 32 (4.24%) as frustrated, 29 (3.84%) as surprised and 12 (1.59%) as anxious.

3. INTERACTION-BASED DETECTORS

The BROMP observations collected while cadets were using vMedic were used to develop machine-learned models to automatically detect the cadet's affective states. In this section, we discuss our work to develop affect detectors based on cadets' vMedic interactions logs.

3.1 Data Integration

In order to generate training data for our interaction-based affect detectors, trainee actions within the software were synchronized to field observations collected using the HART application. During data collections, both the handheld computers and the GIFT server were synchronized to the same internet NTP time server. Timestamps from both the HART observations and the interaction data were used to associate each observation to the actions that occurred during the 20 seconds window prior to data entry by the observer. Those actions were considered as co-occurring with the observation.

3.2 Feature Distillation

For each observation, we distilled a set of 38 features that summarized the actions that co-occurred with or preceded that observation. Those features included: changes in the casualty, both recent and since injury, such as changes in blood volume, bleed rate and heart rate; player states in terms of attacker, such as being under cover and being with the unit; the number of time specific actions, such as applying a tourniquet or requesting a security sweep, were executed; and time between actions. (see [5] for a more complete list of features.)

3.3 Machine Learning Process

Detectors were built separately for each affective state and behavioral constructs. For example a detector was used to distinguish observations of boredom from observations that were not boredom. It is worth noting that the construct of engaged concentration, was defined during modeling as a learner having the affect of concentration and not being off-task, since concentrating while being off-task reflects concentration with something other than learning within the vMedic game. Only 2 such observations was found amongst the collected observations. Detectors were not developed for off-task behavior, Without Thinking Fastidiously behavior, and anxiety due to the low number of observations for those construct (19, 1 and 12 respectively).

Each detector was validated using 10-fold participant-level cross-validation. In this process, the trainees are randomly separated into 10 groups of equal size and a detector is built using data for each combination of 9 of the 10 groups before being tested on the 10th group. By cross-validating at this level, we increase confidence that detectors will be accurate for new trainees. Oversampling (through cloning of minority class observations) was used to make the class frequency more balanced during detector development. However, performance calculations were made with reference to the original dataset.

Detectors were fit in RapidMiner 5.3 [21] using six machine learning algorithms that have been successful for building similar detectors in the past ([3], [24]): J48, JRip, NaiveBayes, Step Regression, Logistic Regression and KStar. The detector with the best performance was selected for each affective state. Detector performance was evaluated using two metrics: Cohen's Kappa [9] and A' computed as the Wilcoxon statistic [18]. Cohen's Kappa assesses the degree to which the detector is better than chance at identifying the modeled construct. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly. A' is the probability that the algorithm will correctly identify whether an observation is a positive or a negative example of the construct (e.g. is the learner bored or not?). A' is equivalent to the area under the ROC curve in signal detection theory [18]. A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. A' was computed at the observation level.

When fitting models, feature selection was performed using forward selection on the Kappa metric. Performance was evaluated by repeating the feature selection process on each fold of the trainee-level cross-validation in order to evaluate how well models created using this feature selection procedure perform on new and unseen test data. The final models were obtained by applying the feature selection to the complete dataset.

4. POSTURE-BASED DETECTORS

The second set of affect detectors we built were based on learner posture during interactions with vMedic. Kinect sensors produced data streams that were utilized to determine learner posture. Using machine learning algorithms, we trained models to recognize affective states based on postural features.

4.1 Data Integration

GIFT has a *sensor module* that is responsible for managing all connected sensors and associated data streams. This includes Kinect sensor data, which is comprised of four complementary data streams: face tracking, skeleton tracking, RGB channel, and depth channel data. Face- and skeleton-tracking data are written to disk in CSV format, with rows denoting time-stamped observations and columns denoting vertex coordinates. RGB and depth channel data are written to disk as compressed binary data files. To analyze data from the RGB and depth channels, one must utilize the GiftKinectDecoder, a standalone utility that is packaged with GIFT, to decompress and render the image data into a series of images with timestamp-based file names. Data from all four channels can be accessed and analyzed outside of GIFT. For the present study, we utilized only vertex data to analyze participants' posture. Each observation in the vertex data consisted of a timestamp and a set of 3D coordinates for 91 vertices, each tracking a key point on the learner's face (aka face tracking) or upper body (aka skeletal tracking). The Kinect sensor sampled learners' body position at a frequency of 10-12 Hz.

It was necessary to clean the Kinect sensor data in order to remove anomalies from the face and skeletal tracking. Close examination of the Kinect data revealed periodic, and sudden, jumps in the coordinates of posture-related vertices across frames. These jumps were much larger than typically observed across successive frames, and they occurred due to an issue with the way GIFT logged tracked skeletons: recording the *most recently* detected skeleton, rather than the *nearest* detected skeleton. This approach to logging skeleton data caused GIFT to occasionally log bystanders standing in the Kinect's field of view rather than the learner using vMedic. In our study, such a situation could occur when a field observer walked behind the trainee.

To identify observations that corresponded to field observers rather than participants, Euclidean distances between subsequent observations of a central vertex were calculated. The distribution of Euclidean distances was plotted to inspect the distribution of between-frame movements of the vertex. If the Kinect tracked field observers, who were physically located several feet behind participants, the distribution was likely to be bimodal. In this case, one cluster would correspond to regular posture shifts of a participant between frames, and the other cluster corresponded to shifts between tracking participants and field observers. This distribution could be used to identify a distance threshold for determining which observations should be thrown out, as they were likely due to tracking field observers rather than participants. Although the filtering process was successful, the need for this process reveals a challenge to the use of BROMP for detectors eventually developed using Kinect or video data.

In addition to cleaning the face and skeleton mesh data, we performed a filtering process to remove data that were unnecessary for the creation of posture-based affect detectors. A majority of the facial vertices recorded by the Kinect sensor were not necessary for investigating trainees' posture. Of the 91 vertices recorded by the Kinect sensor, only three were utilized for posture analysis: *top_skull*, *head*, and *center_shoulder*. These vertices were selected based on prior work investigating postural indicators of emotion with Kinect data [16].

Finally, HART observations were synchronized with the data collected from the Kinect sensor. As was the case for our interaction-based sensor, the Kinect data provided by GIFT was synchronized to the same NTP time server as the HART data. This allowed us to associate field observations with observations of face and skeleton data produced by the Kinect sensor.

4.2 Feature Distillation

We used the Kinect face and skeleton vertex data to compute a set of predictor features for each field observation. The engineered features were inspired by related work on posture sensors in the affective computing literature, including work with pressure-sensitive chairs ([10], [11]) and, more recently, Kinect sensors [16]. Several research groups have converged on common sets of postural indicators of emotional states. For example, in several cases boredom has been found to be associated with leaning back, as well as increases in posture variance ([10], [11]). Conversely, confusion and flow have been found to be associated with forward-leaning behavior ([10], [11]).

We computed a set of 73 posture-related features. The feature set was designed to emulate the posture-related features that had previously been utilized in the aforementioned posture-based affect detection work ([10], [11], [16], [17]). For each of three retained skeletal vertices tracked by the Kinect (*head*, *center_shoulder*, and *top_skull*), we calculated 18 features based on multiple time window durations. These features are analogous to those described in [16], and were previously found to predict learners' retrospective self-reports of frustration and engagement:

- Most recently observed distance

Table 1. Performance of each of the interaction-based and posture-based detectors of affect

Affect	Interaction-Based Detectors			Posture-Based Detectors		
	Classifier	**Kappa**	**A'**	**Classifier**	**Kappa**	**A'**
Boredom	Logistic Regression	0.469	0.848	Logistic Regression	0.109	0.528
Confusion	Naïve Bayes	0.056	0.552	JRip	0.062	0.535
Engaged Concentration	Step Regression	0.156	0.590	J48	0.087	0.532
Frustration	Logistic Regression	0.105	0.692	Support Vec. Machine	0.061	0.518
Surprise	KStar	0.081	0.698	Logistic Regression	-0.001	0.493

- Most recently observed depth (Z coordinate)
- Minimum observed distance observed thus far
- Maximum observed distance observed thus far
- Median observed distance observed thus far
- Variance in distance observed thus far
- Minimum observed distance during past 5 seconds
- Maximum observed distance during past 5 seconds
- Median observed distance during past 5 seconds
- Variance in distance during past 5 seconds
- Minimum observed distance during past 10 seconds
- Maximum observed distance during past 10 seconds
- Median observed distance during past 10 seconds
- Variance in distance during past 10 seconds
- Minimum observed distance during past 20 seconds
- Maximum observed distance during past 20 seconds
- Median observed distance during past 20 seconds
- Variance in distance during past 20 seconds

We also induced several *net_change* features, which are analogous to those reported in [11] and [10] using pressure-sensitive seat data:

$$net\_dist\_change[t] = \left| \begin{array}{c} head\_dist[t] - head\_dist[t-1] + \\ cen\_shldr\_dist[t] - cen\_shldr\_dist[t-1] + \\ top\_skull\_dist[t] - top\_skull\_dist[t-1] \end{array} \right| \qquad (1)$$

$$net\_pos\_change[t] = \left| \begin{array}{c} head\_pos[t] - head\_pos[t-1] + \\ cen\_shldr\_pos[t] - cen\_shldr\_pos[t-1] + \\ top\_skull\_pos[t] - top\_skull\_pos[t-1] \end{array} \right| \qquad (2)$$

These features were calculated from Kinect vertex tracking data, as opposed to seat pressure data. Specifically, the *net_dist_change* feature was calculated as each vertex's net change in distance (from the Kinect sensor) over a given time window, and then summed together. The *net_pos_change* feature was calculated as the Euclidean distance between each vertex's change in position over a given time window, and then summed together. Both the *net_dist_change* feature and *net_pos_change* feature were calculated for 3 second and 20 second time windows.

We also calculated several *sit_forward*, *sit_back*, and *sit_mid* features analogous to [10] and [17]. To compute these features, we first calculated the average median distance of participants' *head* vertex from each Kinect sensor. This provided a median distance for each of the 10 study stations (see Figure 1). We also calculated the average standard deviation of *head* distance from each sensor. Then, based on the station-specific medians and standard deviations, we calculated the following features for each participant:

$$sit\_forward = \begin{cases} 1 & if\ head\_dist \leq median\_dist\ \text{-}\ st\_dev \\ 0 & otherwise \end{cases} \qquad (3)$$

$$sit\_back = \begin{cases} 1 & if\ head\_dist \geq median\_dist + st\_dev \\ 0 & otherwise \end{cases} \qquad (4)$$

The *sit_mid* feature was the logical complement of *sit_forward* and *sit_back*; if a learner was neither sitting forward, nor sitting back, they were considered to be in the *sit_mid* state. We also computed predictor features that characterized the proportion of observations in which the learner was in a *sit_forward*, *sit_back*, or *sit_mid* state over a window of time. Specifically, we calculated these features for 5, 10, and 20 second time windows, as well as over the entire session to-date.

4.3 Machine Learning

Posture-based detectors of affect were built using a process analogous to the one used to build our interaction-based detectors. As such, separate detectors were, once again, built for each individual affective state and behavioral construct. All observations labeled as '?' were removed from the training set as they represent observations where the cadet's affective state or behavior could not be determined.

Each detector was validated using 10-fold participant-level cross-validation. Oversampling was used to balance class frequency by cloning minority class instances, as was the case when training our interaction-based detectors. RapidMiner 5.3 was used to train the detectors using multiple different classification algorithms: J48 decision trees, naïve Bayes, support vector machines, logistic regression, and JRip. When fitting posture-based affect detection models, feature selection was, once again, performed through forward selection using a process analogous to the one used for our interaction-based detectors.

5. RESULTS

As discussed above, each of the interaction-based and posture-based detectors of affect were cross-validated at the participant level (10 folds) and performance was evaluated using both Kappa and A'. Table 1 summarizes the performance achieved by each detector for both the Kappa and A' metrics.

Performance of our interaction-based detectors was highly variable across affective states. The detector of boredom achieved, by far, the highest performance (Kappa = 0.469, A' = 0.848) while some of the other detectors achieved very low performance. This was the case for the confusion detector that performed barely above chance level (Kappa = 0.056, A' = 0.552). Detectors of frustration and surprise achieved relatively low Kappa (0.105 and 0.081 respectively), but good A' (0.692 and 0.698 respectively). Performance for engaged concentration achieved a Kappa closer to the average (0.156), but below average A' (0.590).

In general, posture-based detectors performed only slightly better than chance, with the exception of the surprise detector, which actually performed worse than chance. The boredom detector, induced as a logistic regression model, achieved the highest predictive performance (Kappa = 0.109, A' = 0.528), induced as a logistic regression model.

6. DISCUSSION

Across affective states, the posture-based detectors achieved lower predictive performance than the interaction-based detectors. In fact, the posture-based detectors performed only slightly better than chance, and in the case of some algorithms and emotions, worse than chance. This finding is notable, given that our distilled posture features were inspired largely from the research literature, where these types of features have been shown to predict learner emotions effectively in other contexts ([10], [11], [16], [17]). For example, D'Mello and Graesser found machine-learned classifiers discriminating affective states from neutral yielded kappa values of 0.17, on average [10]. Their work utilized posture features distilled from pressure seat data, including several features analogous to those used in our work. Grafsgaard et al. found that Pearson correlation analyses with retrospective self-reports of affect revealed significant relationships between posture and emotion, including frustration, focused attention, involvement, and overall engagement. Reported correlation coefficients ranged in magnitude from 0.35 to 0.56, which are generally considered moderate to large effects [19]. Cooper et al. found that posture seat-based features were particularly effective for predicting excitement in stepwise regression analyses ($R = 0.56$), and provided predictive benefits beyond log-based models across a range of emotions [10]. While the methods employed in each of these studies differ from our own, and thus the empirical results are not directly comparable, the qualitative difference in the predictive value of postural features is notable.

There are several possible explanations for why our posture-based predictors were not more effective. First, our use of BROMP to generate affect labels distinguishes our work from prior efforts, which used self-reports ([10], [16], [17]) or retrospective video freeze-frame analyses [11]. It is possible that BROMP-based labels of affect present distinct challenges for posture-based affect detection. BROMP labels are based on holistic judgments of affect, and pertain to 20-second intervals of time, which may be ill matched for methods that depend upon low-level postural features to predict emotion. Similarly, much of the work on posture-based affect detection has taken place in laboratory settings involving a single participant at a time [11], especially prior work using Kinect sensors ([16], [17]). In contrast, our study was performed with up to 10 simultaneous participants (see Figure 2), introducing potential variations in sensor positions and orientations. This variation may have introduced noise to our posture data, making the task of inducing population-general affect detectors more challenging than in settings where data is collected from a single sensor. If correct, this explanation underscores the challenges inherent in scaling and generalizing sensor-based affect detectors.

The study room's setup also limited how sensors could be positioned and oriented relative to participants. For example, it was not possible to orient Kinect cameras to the sides of participants, capturing participants' profiles, which would have made it easier to detect forward-leaning and backward-leaning postures. This approach has shown promise in other work, but was not a viable option in our study [31]. Had the Kinect sensors been positioned in this manner, the video streams would have been disrupted by other participants' presence in the cameras' fields of view.

Another possible explanation has to do with the population of learners that was involved in the study: U.S. Military Academy (USMA) cadets. Both BROMP observers noted that the population's affective expressiveness was generally different in kind and magnitude than the K-12 and civilian academic populations they were more accustomed to studying. Specifically, they indicated that the USMA population's facial and behavioral expressions of affect were relatively subdued, perhaps due to military cultural norms. As such, displays of affect via movement and body language may have been more difficult to recognize than would have otherwise been encountered in other populations.

In general, we consider the study population, BROMP affect labels, and naturalistic research setup to be strengths of the study. Indeed, despite the difference in how military display affect compared to the K-12 and civilian academic population, human observers were able to achieve the inter-rater reliability required by BROMP (Kappa >= 0.6) [23]. Thus we do not have plans to change these components in future work. Instead, we will likely seek to revise and enhance the data mining techniques that we employ to recognize learner affect, as well as the predictor features engineered from raw posture data. In addition, we plan to explore the predictive utility of untapped data streams (e.g., Q-Sensor data, video data).

It is notable that our interaction-based detectors had a more varied performance than had been seen in prior studies using this methodology; the detectors were excellent for boredom, and varied from good to just above chance for other constructs. It is possible that this too is due to the population studied, but may also be due to the nature of the features that were distilled in order to build the models. For example, the high performance of our detector of boredom can be attributed to the fact that one feature, whether the student executed any meaningful actions in the 20 second observation window, very closely matched the trainees' manifestation of this affective state. In fact, a logistic regression detector trained using this feature alone achieved higher performance than our detectors for any of the other affective state (Kappa = 0.362, A' = 0.680). It can be difficult to predict, a priori, which features will most contribute to the detection of a specific affective state. It is also possible that some of the affective states for which interaction-based detection was less effective (e.g., confusion) simply did not manifest consistently in the interactions with the learning environment across different trainees. It is thus difficult to determine whether poor performance of detectors for some constructs, such as our confusion detector, is due to insufficient feature engineering or inconsistent behaviors by the trainee. As such, the creation of interaction-based detectors is an iterative process, where features are engineered, and models are induced and refined, until performance reaches an acceptable level, or no improvement in performance is observed, despite repeated knowledge-engineering efforts.

We aim to identify methods to improve the predictive accuracy of posture-based detectors in future work. One advantage they possess relative to interaction-based detectors is that posture-based detectors may be more generalizable, since they pertain to aspects of learner behavior that are outside of the software itself. By contrast, much of the effort invested in the creation of interaction-based detectors is specific to the system for which the detectors are created. Features are built to summarize the learner's interaction in the learning environment and, as such, are dependent on the system's user interface. Much of the creation of interaction-based detectors must hence be replicated for new learning environments, though there have been some attempts to build toolkits that can replicate features seen across many environments, such as unitizing the time between actions by the type of action or problem step (e.g. [28]).

On the other hand, posture-based detectors are built upon a set of features that are more independent of the system for which the detectors are designed. The process of creating the features itself requires considerable effort when compared with building a set of features for interaction-based detectors, such as elaborate efforts to adequately clean the data, but at least in principle, it is only necessary to develop the methods for doing so once. The same data cleaning and feature distillation procedures can be repeated for subsequent systems. This is especially useful in the context of a generalized, multi-system tutoring framework such as GIFT [32]. Although different posture-based affect detectors might need to be created for different tutoring systems—due to differences in the postures associated with affect for different populations of learners, environments and contexts—the posture features we computed from the data provided by Kinect sensors will ultimately become available for re-use by any tutor created using GIFT. This has the potential to considerably reduce the time required to build future posture-based affect detectors for learning environments integrated with the GIFT architecture.

7. CONCLUSION

Interaction-based and posture-based detectors of affect show considerable promise for adaptive computer-based learning environments. We have investigated their creation and predictive performance in the context of military cadets using the vMedic serious game for tactical combat casualty care. Interaction-based and posture-based detectors capture distinct aspects of learners' affect. Whereas interaction-based detectors capture the relationship between affect and its impact on the trainee's action in the learning environment, posture-based detectors capture learners' physical expressions of emotion.

In our study, we found that interaction-based detectors achieved overall higher performance than posture-based detectors. We speculate that the relatively weak predictive performance of our posture-based affect detectors may be due to some combination of the following: the interplay of high-level BROMP affect labels and low-level postural features, the challenges inherent in running sensor-based affect studies with multiple simultaneous participants, and population-specific idiosyncrasies in USMA cadets' affective expressiveness compared to other populations. The relative advantages and limitations of both interaction-based and posture-based detectors point toward the need for continued research on both types. Each type of detector captures different aspects of learners' manifestations of affective state, and many open questions remain about feature engineering and the predictive ability of each type of detector.

An important direction for future work will be the integration and combination of the two types of detectors presented here. In multiple cases, the combination of data modalities for the creation of affect detectors has been shown to produce detectors with better performance than single-modality detectors ([12], [13], [17]). As such, future work will focus on the study of how these two channels of information can be combined to produce more effective and robust detectors of affect.

Further research on effective, generalizable predictor features for posture-based affect detectors is also needed, as shown by the relatively weak predictive performance of existing features observed in this study. Complementarily, investigating the application of other machine learning algorithms, including temporal models, is likely to prove important, given the complex temporal dynamics of affect during learning. These directions are essential for developing an enhanced understanding of the interplay between affect detector architectures, learning environments, student populations, and methods for determining ground truth affect labels. While significant progress has been made toward realizing the vision of robust, generalizable affect-sensitive learning environments, these findings point toward the need for continued empirical research, as well as advances in educational data mining methods applicable to affective computing.

8. ACKNOWLEDGMENTS

We thank our research colleagues, COL James Ness and Dr. Michael Matthews in the Behavioral Science & Leadership Department at the United States Military Academy for their assistance in conducting the study. This research is supported by cooperative agreement #W911NF-13-2-0008 between the U.S. Army Research Laboratory, Teachers College Columbia University, and North Carolina State University. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Laboratory.

9. REFERENCES

[1] AlZoubi, O., Calvo, R.A., and Stevens, R.H. 2009. Classification of EEG for Emotion Recognition: An Adaptive Approach. *Proc. of the 22nd Australian Joint Conference on Artificial Intelligence*, 52-61.

[2] Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K., and Christopherson, R. 2009. Emotion Sensors Go to School. *Proc. of the 14th Int'l Conf. on Artificial Intelligence in Education*, 17-24.

[3] Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G.W., Ocumpaugh, J., and Rossi, L. 2012. Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *Proc. of the 5th Int'l Conf. on Educational Data Mining*, 126-133.

[4] Baker, R., D'Mello, S., Rodrigo, M.M.T., and Graesser, A. 2010. Better to be Frustrated than Bored: The Incidence and Persistence of Affect During Interactions with Three Different Computer-Based Learning Environments. *Int'l J. of Human-Computer Studies*, 68 (4), 223-241.

[5] Baker, R.S., DeFalco, J.A., Ocumpaugh, J., and Paquette, L. 2014. Towards Detection of Engagement and Affect in a Simulation-Based Combat Medic Training Environment. *2nd Annual GIFT User Symposium (GIFTSym2).*

[6] Baker, R.S., and Ocumpaugh, J. 2015. Interaction-Based Affect Detection in Educational Software. The Oxford Handbook of Affective Computing, 233-245.

[7] Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., and Zhao, W. 2015. Automatic Detection of Learning-Centered Affective States in the Wild. *Proc. of the 2015 Int'l Conf. on Intelligent User Interfaces.*

[8] Calvo, R.A., and D'Mello, S. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and their

Applications. *IEEE transactions on Affective Computing*, 1 (1), 18-37.

[9] Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational Psychological Measurement*, 20 (1), 37-46.

[10] Cooper, D.G., Arroyo, I., Woolf, B.P., Muldner, K., Burleson, W., and Christopherson, R. 2009. Sensors Model Student Self Concept in the Classroom. *Proc. of the 17th Int'l Conf. on User Modeling, Adaption, and Personalization,* 30-41.

[11] D'Mello, S., and Graesser, A. 2009. Automatic detection of learners' affect from gross body language. *Applied Artificial Intelligence*, 23, 2, 123-150.

[12] D'Mello, S., Kory, J. 2012. Consistent but Modest: Comparing Multimodal and Unimodal Affect Detection Accuracies from 30 Studies. *Proc. of the 14th ACM International Conf. on Multimodal Interaction*, 31-38.

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

[14] D'Mello, S., Lehman, B., Pekrun, R., and Graesser, A. 2014. Confusion can be Beneficial for Learning. *Learning and Instruction*, 29, 153-170.

[15] D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., Person, N., Kort, B., el Kaliouby, R., Picard, R., and Graesser, A. 2008. AutoTutor Detects and Responds to Learners Affective and Cognitive States. *Workshop on Emotional and Cognitive Issues at the 9th Int'l Conf. on Intelligent Tutoring Systems.*

[16] Grafsgaard, J., Boyer, K., Wiebe, E., and Lester, J. 2012. Analyzing Posture and Affect in Task-Oriented Tutoring. *Proc. of the 25th Florida Artificial Intelligence Research Society Conference*, 438-443.

[17] Grafsgaard, J., Wiggins, J., Boyer, K.E., Wiebe, E., and Lester, J. 2014. Predicting Learning and Affect from Multimodal Data Streams in Task-Oriented Tutorial Dialogue. *Proc. of the 7th Int'l Conf. on Educational Data Mining*, 122-129.

[18] Hanley, J., and McNeil, B. 1982. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143, 29-36.

[19] Litman, D.J., and Forbes-Riley K. 2006. Recognizing Student Emotions and Attitudes on the Basis of Utterances in Spoken and Tutoring Dialogue with Both Humans and Computer-Tutors. *Speech Communication*, 48, 559-590.

[20] Metcalf, S., Kamarainen, A., Tutwiler, M.S., Grotzer, T., and Dede, C. 2011. Ecosystem Science Learning via Multi-User Virtual Environments. *Int'l J. of Gaming and Computer-Mediated Simulations*, 3 (1), 86-90.

[21] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. 2006. YALE: Rapid Prototyping for Complex Data Mining Tasks. *Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining*, 935-940.

[22] Nasoz, F., Alvarez, K., Lisetti, C.L., and Finkelstein, N. 2004. Emotion from Physiological Signals Using Wireless Sensors for Presence Technologies. *Cognition, Technology and Work*, 6, 4-14.

[23] Ocumpaugh, J., Baker, R.S.J.d, and Rodrigo, M.M.T. 2012. *Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0.* Technical Report.

[24] Pardos, Z., Baker, R.S.J.d., San Pedro, M.O.Z., Gowda, S.M., and Gowda, S. 2014. Affective States and State Tests: Investigating how Affect and Engagement During the School Year Predict End of Year Learning Outcomes. *J. of Learning Analytics*, 1 (1), 107-128.

[25] Pekrun, R., Goetz, T., Daniels, L.M., Stupnisky, R.H., and Perry, R.H. 2010. Boredom in Achievement Settings: Exploring Control-Value Antecedents and Performance Outcomes of a Neglected Emotion. *J. of Educational Psychology*, 102, 531-549.

[26] Rai, D., Arroyo, I., Stephens, L., Lozano, C., Burleson, W., Woolf, B.P., and Beck, J.E. 2013: Repairing Deactivating Emotions with Student Progress Pages. *Proc. of the 16th Int'l Conf. on Artificial Intelligence in Education*, 795-798.

[27] Rodrigo, M.M.T., and Baker, R.S.J.d. 2009. Coarse-Grained Detection of Student Frustration in an Introductory Programming Course. *Proc. of the 5th Int'l Workshop on Computing Education Research Workshop*, 75-80.

[28] Rodrigo, M.M.T., Baker, R.S.J.d., McLaren, B., Jayme, A., and Dy, T. 2012. Development of a Workbench to Address the Educational Data Mining Bottleneck. *Proc. of the 5th Int'l Conf. on Educational Data Mining,* 152-155.

[29] Rowe, J., Mott, B., McQuiggan, J., Robison, S., and Lester, J. 2009. Crystal Island: A Narrative-Centered Learning Environment for Eighth Grade Microbiology. *Workshop on Educational Games at the 14th Int'l Conf. on Artificial Intelligence in Education*, 11-20.

[30] Ryan, R.M. 1982. Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory. *J. of Personality and Social Psychology*, 43, 450-461.

[31] Sanghvi, J., Castellano, G., Leite, I., Pereria, A., McOwan, P., and Paiva, A. 2011. Automatic Analysis of Affective Postures and Body Motion to Detect Engagement with a Game Companion. *Proc. of the 6th ACM/IEEE Int'l Conf. on Human-Robot Interaction*, 305-311.

[32] Sottilare, R.A., Golberg, B. and Holden, H. 2012. *The Generalized Intelligent Framework for Tutoring (GIFT).*

[33] Wixon, M., Baker, R.S.J.d., Gobert, J., Ocumpaugh, J., and Bachmann, M. 2012. WTF? Detecting Students who are Conducting Inquiry Without Thinking Fastidiously. *Proc. of the 20th Int'l Conf. on User Modeling, Adaptation and Personalization*, 286-298.

[34] Witmer, B.G., Jerome, C. J., & Singer, M. J. (2005, June). The factor structure of the presence questionnaire. *Presence*, Vol. 14(3), 298 312. MIT Press, Cambridge MA.

[35] Zeng, Z., Pantic, M., Roisman, G.I., and Huang, T.S. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (1), 39-58.