Improving Affect Detection in Game-Based Learning with Multimodal Data Fusion

[1]Nathan Henderson, [1]Jonathan Rowe, [2]Luc Paquette, [3]Ryan S. Baker, and [1]James Lester

[1] North Carolina State University, Raleigh, NC, USA
{nlhender, jprowe, lester}@ncsu.edu

[2] University of Illinois at Urbana-Champaign, Champaign, IL, USA
lpaq@illinois.edu

[3] University of Pennsylvania, Philadelphia, PA, USA
rybaker@upenn.edu

Abstract. Accurately recognizing learner affect is critically important for enabling affect-responsive learning environments to support student learning and engagement. Multimodal affect detection combining sensor-based and sensor-free approaches has shown significant promise in both laboratory and classroom settings. However, important questions remain regarding which data channels are most predictive and how they should be combined. In this paper, we investigate a multimodal affect detection framework that integrates motion tracking-based posture data and interaction-based trace data to recognize the affective states of students engaged with a game-based learning environment for emergency medical training. We compare several machine learning-based affective models using competing feature-level and decision-level multimodal data fusion approaches. Results indicate that multimodal affect detectors induced using joint feature representations from posture-based and interaction-based data channels yield improved accuracy relative to unimodal models across several learner-centered affective states. These findings point toward implications for the design of multimodal affect-responsive learning environments that support learning and engagement.

Keywords: Affect detection, multimodal data fusion, game-based learning

1 Introduction

Affect is critical in student learning. Automatically recognizing learners' affective states is foundational to the development of affect-responsive learning environments that can support student emotion regulation and promote enhanced learning experiences [1]. Recent years have seen growing interest in the use of *sensor-based* approaches for capturing data on student affect within adaptive learning environments, and in particular, game-based learning environments [1, 2]. An important feature of sensor-based affect detection is its potential for generalizability across a range of domains and learning environments [3]. Sensor-based modalities such as facial expression [4] and posture [5] have been demonstrated to be highly indicative of learner-centered affective states.

An alternative to sensor-based affect detection is utilizing interaction trace log data to induce *sensor-free* affect detectors, which can be used in contexts where it may be difficult or prohibitive to use physical hardware sensors [6]. Sensor-free features are typically derived from trace data generated by learner interactions with adaptive learning environments [7]. Notably, sensor-free affect detectors typically avoid challenges inherent in the use of physical sensors, including calibration issues, hardware failure, and mistracking [8].

A subject of growing interest is the application of multimodal machine learning techniques to develop affect detectors using multiple complementary data sources. Multimodal affect detectors integrate sensor-free (i.e., interaction-based) and sensor-based approaches, capturing multiple simultaneous perspectives on student interactions with adaptive learning environments. Important questions remain about the predictive value of specific modalities and how they should be combined. Prior work has investigated multimodal affect detection across a range of educational subjects, including science [9], math [10], and introductory programming [11]. However, there is a need for continued research on how effectively multimodal affect detection techniques translate to alternative learning environments and educational subjects.

In this work, we present a multimodal affect detection framework that utilizes posture data and interaction-based trace data from college students engaged with a game-based learning environment for emergency medical training called TC3Sim. We extract both spatial and temporal posture features captured by a Microsoft Kinect sensor as well as interaction-based features depicting students' actions within the game-based learning environment. We compare several methods of multimodal data fusion to determine the optimal approach for detecting students' learner-centered affective states using a range of machine learning-based classification techniques. Results indicate that multimodal affect detectors that utilize a combination of posture-based and interaction-based feature representations outperform competing unimodal baseline models on classification accuracy across several affective states. In this work, we focus on variations of both decision-level and feature-level multimodal data fusion to determine the optimal method of combining modalities during the affective modeling process.

2 Related Work

Recent years have seen growing interest in both sensor-free and sensor-based affect detection in adaptive learning environments. Deep neural architectures have shown promise in sensor-free affect detection. Jiang et al. investigated tradeoffs between deep learning-based representation learning and expert feature-engineering in a sensor-free affect detection framework using interaction trace log data [7]. Botelho et al. explored the use of recurrent neural networks in a related sensor-free affect detection task [6]. More recent work has explored improvements in unimodal affect detection with the introduction of sensor-based modalities [12]. For example, Paquette et al. examined the predictive accuracy of several unimodal sensor-free and sensor-based affect detectors that utilized interaction trace log data as well as posture-based data [13], but did not explore multimodal models that integrate multiple modalities simultaneously.

Multimodal affect detection combining sensor-free and sensor-based data channels offers several benefits in terms of model accuracy and robustness. Grafsgaard et al. used multimodal posture and gesture data to model undergraduate students' affect as they engaged in computer-mediated tutoring sessions on introductory programming [14], with results indicating that more shifts in posture corresponds to increased frustration, while stationary posture may be predictive of engagement. Other multimodal affective computing work has investigated the predictive value of combining interaction-based modalities, such as keystroke data or text-based dialogues, with sensor-based modalities such as posture and gesture data [15]. The results of these prior efforts demonstrated the effectiveness and additive value sensor-based modalities contribute compared to unimodal dialogue-only models. Additional work has investigated student affective responses with facial expression data in combination with interaction patterns as a secondary modality [2]. Bosch et al. investigated the combination of facial expression and interaction log data to detect affect in students using an educational game to teach elementary physics, reporting that the facial expression modality was more predictive than the interaction-based modality [16]. However, important questions remain regarding how to most effectively combine independent modalities for affect detection in adaptive learning environments, such as student posture and interaction log data.

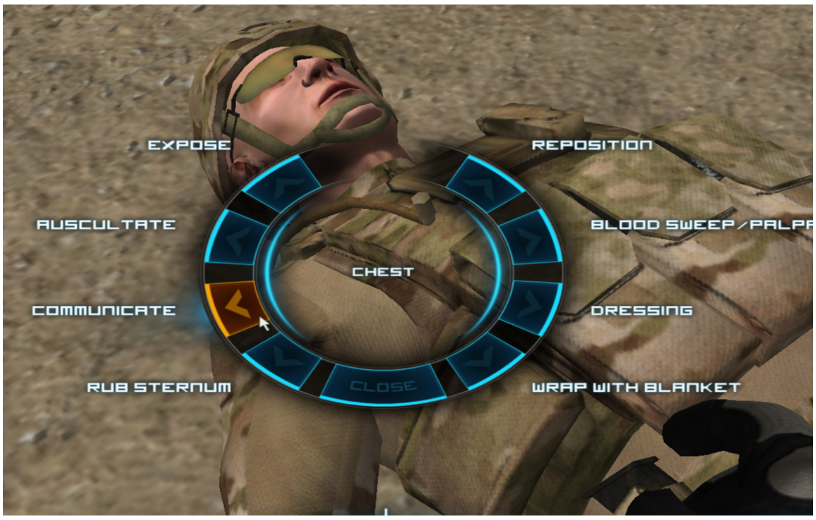

Fig. 1. TC3Sim game-based learning environment.

3 Multimodal Data Collection

To investigate multimodal affect detection, we utilized an existing dataset containing sensor-based and interaction-based log data from learners engaged with a game-based learning environment for emergency military medical training, *TC3Sim*. In TC3Sim, learners complete a series of simulated medical training scenarios and are tasked with providing adequate medical care to one or more wounded teammates. The dataset consisted of sensor data and interaction trace logs from 119 undergraduate students (i.e., cadets) from the United States Military Academy (83% male, 17% female). During the data collection, participants completed a series of four training scenarios in TC3Sim, which ranged from situations involving the simple application of a tourniquet to simulated scenarios involving severely injured characters expiring regardless of medical care administered (Fig. 1). Each learner engaged with TC3Sim for approximately one hour. Interaction trace log data was captured using *GIFT*, an open-source service-oriented software framework used to develop and deploy adaptive training environments [17]. To facilitate capture of learners' posture data, each participant sat at a laptop connected to a Microsoft Kinect motion-tracking sensor, which was positioned directly in front of the participant to capture skeletal vertex coordinates based on posture and upper-body movement during the session. For additional detail about the dataset, please see DeFalco et al [1]. We elect to use interaction data due to its ease of collection and cost effectiveness, while also utilizing posture-based data due to its low-cost, non-invasive method of capture.

To obtain ground-truth labels of affect, a pair of trained observers annotated students' affective states and learner behaviors in accordance with the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) [18]. The observations were recorded in 20-second intervals, and they were made using a small handheld device to allow annotations to be recorded discreetly. The two observers recorded an inter-rater agreement on a subset of the data (i.e., data from a single one-hour session) exceeding 0.6 in terms of Cohen's Kappa [19]. During this study, six affective states were recorded: *bored, confused, engaged concentration, frustration, surprise,* and *anxiety*, with 3,066 distinct BROMP observations captured between the two observers. During the post-processing stage, any observations for which the two observers did not agree were removed from the dataset, as were observations recorded when the students were not actually interacting with the game-based learning environment (i.e., viewing pre- and post-test material or receiving instruction on combat medic procedures). Following post-processing, there were 755 total BROMP observations captured during the study. 435 of the BROMP observations were labeled as *engaged concentration*, 174 as *confused*, 73 as *bored*, 32 as *frustrated*, 29 as *surprised*, and 12 as *anxious*. Due to the low number of observations for *anxious*, we do not consider instances of this affective state in this work.

4 Multimodal Affect Detection

Using the dataset containing synchronized posture data, interaction trace log data, and affect observation data, we induced binary affect detectors for the following learner-centered affective states: *bored, confused, engaged concentration, frustrated,* and *surprised.* We extracted three types of features—posture-based spatial features, posture-based temporal features, and interaction-based features—using feature engineering techniques. The data is upsampled (within the training set only) to resolve imbalances for specific affective states. The features are normalized, and they are utilized to train, validate, and test several machine learning-based models. Due to the multimodal nature of our dataset, we evaluated three variations of data fusion techniques to capture and model information from different modalities, including two feature-level fusion techniques and a decision-level technique [20].

4.1 Posture-Based Spatial Features

The Kinect motion-tracking sensor captures (x, y, z) coordinate information for 91 vertices. We selected the *head, top_skull,* and *center_shoulder* vertices to generate features based on prior work for similar posture-based affect detection tasks [14]. From these vertices, we extracted 73 distinct features to capture the spatial position of each participant's head and upper torso. For each of the three vertices, several positional features were extracted, including current Euclidean distance from the Kinect, current Z-coordinate, minimum observed distance, maximum observed distance, median distance, and variance in observed distance. These features were calculated for each BROMP observation. Additionally, summative features, such as the minimum, maximum, median, and variance in distance, were calculated for the preceding 5, 10, and 20 second time intervals prior to each BROMP observation. In addition to these 54 features, several features were extracted to capture net changes in posture and distance for time windows of 3 and 20 seconds. Finally, several features were calculated to determine whether a learner was leaning forward, backward, or sitting upright. These features were calculated using the *head* vertex. The learner was considered to be leaning forward or backward if the vertex was more than a single standard deviation from the median head position for that particular workstation. These three posture-based features were calculated over observed sequences of 5, 10, and 20 seconds, in addition to the entire gameplay session up to the current observation.

4.2 Posture-Based Temporal Features

While skeletal tracking functionalities of motion-tracking sensors, such as Microsoft Kinect, directly capture spatial information about upper body position, temporal information about torso movement is often left implicit despite having been shown to be informative and yield improved accuracy in affect recognition tasks [21]. To address this issue, we extracted several temporal features that capture the "velocity" of skeletal vertices tracked by the Microsoft Kinect sensor. Specifically, the distance between consecutive sensor readings was calculated for the head vertex's positional coordinates. The subsequent delta values were used to generate velocity features calculated across time windows of 3, 5, 10, and 20 seconds preceding each BROMP observation. Extracted statistical features included the mean, median, max, and variance of each corresponding velocity value, introducing an additional 48 features that served as a form of simulated temporal modality [22]. Because of the large number of features generated per vertex, additional velocity information was not calculated using the *top_skull* and *center_shoulder* vertices.

4.3 Interaction-Based Features

From the gameplay interaction logs, we extracted 39 distinct features centered around the participants' actions in the TC3Sim game-based learning environment, as well as information about the virtual patients treated in the game. Features summarizing the state of virtual patients in TC3Sim included changes in systolic blood pressure and heart rate, number of exposed wounds, lung volume, remaining blood volume, and bleed rate. Additionally, features were extracted based upon actions taken by the learner such as checking a patient’s vital signs, conducting a blood sweep, applying a bandage or tourniquet, communicating with the patient, or requesting an evacuation. The resulting interaction-based features were calculated over the 20 second duration prior to the current BROMP observation, using statistical measures such as the sum or current count of a gameplay action, or the standard deviation or average of a metric such as blood pressure.

4.4 Affect Model Evaluation

Following feature engineering, binary datasets were generated for each of the affective states with a variable (i.e., label) denoting whether the record was associated with a positive instance of that particular affective state (e.g., *bored, confused*, *engaged concentration*). Because of the imbalance between positive and negative instances of several affective states, including *frustrated* and *surprised*, each dataset underwent upsampling using the Synthetic Minority Oversampling Technique (SMOTE) [23], within training sets only. This process selects a positive instance of the minority class at random and linearly interpolates synthetic data points between the selected point and another minority sample chosen by a randomized K-nearest neighbor clustering approach. SMOTE is a common approach to resolving class imbalance issues by bringing the class distribution to a uniform balance while avoiding duplication of minority instances, which can lead to overfitting in affective models.

The datasets for each binary classification task were split into a training set and a held-out test set, containing approximately 80% and 20% of the total dataset, respectively. The datasets were sampled to ensure that the distributions between training and test data were relatively similar. The training set was used to evaluate each of the modeling approaches using 4-fold cross-validation. The splits for both the cross-validation and training/test sets were performed at a student level to avoid data leakage from a single session during either the training or evaluation phases.

Prior to training, each of the datasets was normalized and underwent forward feature selection to allow the models to train using only selected features, eliminating redundant or otherwise uninformative features. Forward feature selection involves iterating through combinations of features in a greedy fashion, beginning with feature vectors of size 1 and continuing until a certain number of features are selected or all combinations of features are exhausted. For this work, we selected 12 features per data channel. In cases involving multimodal input, 6 features were selected per feature type (i.e., spatial and temporal) for the posture-based feature representations, and 4 features were selected per feature type across both of the data channels (i.e., spatial, temporal, and interaction). We used a support vector machine (SVM) to guide feature selection due to its ability to efficiently perform non-linear classification. A feature is selected as “optimal” if its addition to a feature set yields improved accuracy for the SVM model. The computational efficiency of the SVM is important due to the number of times a model is trained during the feature selection process. Feature normalization, upsampling, feature selection, and model training took place within each cross-validation fold to prevent data leakage across the training and validation data.

4.5 Multimodal Data Fusion with Posture- and Interaction-Based Modalities

We investigated several approaches for integrating feature representations from the independent modalities using multimodal data fusion techniques. Two commonly used variations of data fusion include feature-level fusion (early fusion) and decision-level fusion (late fusion). *Early Fusion* (EF) involves the concatenation of features prior to training the models. We evaluated two different configurations of Early Fusion. Early Fusion 1 (EF1) implements feature selection (F. S.) following feature concatenation, and Early Fusion 2 (EF2) implements feature selection prior to feature concatenation. *Late Fusion* (LF) involves independently training a separate model on each modality and subsequently obtaining a single representative prediction through a voting scheme based on the predictions and confidence levels of each model. We used the highest average confidence across each class to determine the final representative prediction within Late Fusion. A visualization of the multimodal data fusion processes is shown in Fig. 2.

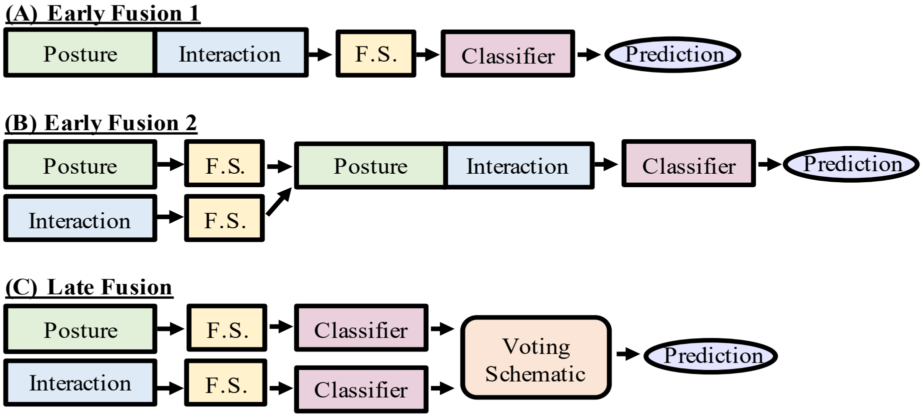

Fig. 2. Visualization of multimodal data fusion pipeline for Early Fusion 1 (2A), Early Fusion 2, (2B), and Late Fusion (2C).

5 Results

We compared five machine learning techniques for inducing detectors of each affective state: Support Vector Machine (SVM), Random Forest (RF), Gaussian Naive Bayes (NB), Logistic Regression (LR), and Multi-Layer Perceptron (MLP). To serve as baselines, we trained unimodal models using either interaction data or posture data, respectively. These models were based upon unimodal affect detectors induced in prior work [1], although we make several methodological refinements related to feature selection, upsampling, cross-validation, evaluation on a held-out test, and implementation of machine learning models. These modifications have a small impact on the results for the baseline models, but overall accuracy trends across affective states remained the same as in prior findings. The posture-only baselines were evaluated using both spatial and temporal modalities using data fusion techniques depicted in Fig. 2, but for this analysis, we consider these models to be "unimodal" because both the spatial and temporal features were extracted from the same sensor-based data channel.

Each model's predictive accuracy was examined under cross-validation on the training set to determine which model was "optimal" for the respective combination of feature set, data fusion method, and affective state. The model with highest performance during cross validation was evaluated using data from the held-out test set. Each model was evaluated with Cohen's Kappa as the primary metric, due to its ability to account for positive classifications occurring due to random chance or dataset-induced bias [19]. We also present results in terms of raw classification accuracy (Acc.) and F1. Results from this evaluation are shown in Table 1, with the highest-performing combination of data fusion technique and model for each affective state shown in bold.

Table 1. Optimal models for each combination of modalities and affective states.

	Bored				**Confused**			
Modality	**Model**	**Kappa**	**Acc.**	**F1**	**Model**	**Kappa**	**Acc.**	**F1**
Gameplay	RF	0.3788	0.8579	0.4476	MLP	0.0161	0.4581	0.3232
Posture (EF1)	MLP	0.3147	0.9074	0.3478	SVM	0.1336	0.6975	0.3288
Posture (EF2)	SVM	0.2941	0.9012	0.3334	**MLP**	**0.2206**	**0.7593**	**0.3607**
Posture (LF)	MLP	0.1107	0.8458	0.1935	MLP	0.1141	0.7531	0.2307
Multimodal (EF1)	LR	0.4328	0.8581	0.5106	MLP	0.1181	0.7099	0.2985
Multimodal (EF2)	**SVM**	**0.4664**	**0.9074**	**0.5161**	MLP	0.1023	0.5000	0.4000
Multimodal (LF)	MLP	0.4568	0.9135	0.5000	MLP	0.1329	0.5432	0.4127

	Engaged Concentration				**Frustrated**			
Modality	**Model**	**Kappa**	**Acc.**	**F1**	**Model**	**Kappa**	**Acc.**	**F1**
Gameplay	MLP	0.1047	0.5718	0.6046	MLP	0.0514	0.6643	0.1182
Posture (EF1)	SVM	0.1565	0.5679	0.5205	MLP	0.1492	0.9283	0.1667
Posture (EF2)	RF	0.1566	0.5864	0.6417	SVM	0.0825	0.9135	0.1250
Posture (LF)	MLP	0.1199	0.5741	0.6532	MLP	0.0825	0.9135	0.1250
Multimodal (EF1)	MLP	0.1199	0.5741	0.6532	NB	0.1124	0.7099	0.2034
Multimodal (EF2)	RF	0.1625	0.6049	0.7117	**MLP**	**0.2054**	**0.8951**	**0.2609**
Multimodal (LF)	**SVM**	**0.2544**	**0.6172**	**0.5694**	SVM	0.0028	0.3395	0.1157

	Surprised			
Modality	**Model**	**Kappa**	**Acc.**	**F1**
Gameplay	RF	0.0797	0.8362	0.1352
Posture (EF1)	MLP	0.0831	0.6605	0.1538
Posture (EF2)	SVM	0.0236	0.8642	0.0834
Posture (LF)	MLP	0.0053	0.0987	0.0875
Multimodal (EF1)	**MLP**	**0.1041**	**0.9259**	**0.1429**
Multimodal (EF2)	MLP	-0.0373	0.9259	0.0000
Multimodal (LF)	MLP	0.0803	0.9135	0.1250

We observe from the results that multimodal affect detectors utilizing a combination of interaction-based and posture-based modalities outperform posture-only baseline and interaction-only baseline models for four out of the five affective states, with the sole exception being the state of *confused*. For the four other affective states, Early Fusion 1 was the best fusion technique for *surprised*, and Early Fusion 2 was the most accurate method for *bored* and *frustrated.* Late Fusion achieved the highest performance for *engaged concentration*. The majority of the affective states produced a relatively high Kappa value (> 0.2), excluding *surprised*.

It is noteworthy that the MLP models were the optimal classification model for a majority of cases (60%), potentially due to their ability to robustly model complex, non-linear relationships between modalities. This capability is especially important when modeling data from multiple independent modalities such as Early Fusion and the posture-based models using both spatial and temporal modalities. SVM and RF models were occasionally the best-performing classification techniques for both unimodal and multimodal affect detection. NB and LR models were each selected once as the best-performing model for a certain multimodal configuration, although neither model was the optimally performing model for an entire affective state.

6 Discussion

To conduct a more in-depth analysis of the predictive value of each modality during multimodal data fusion, we recorded the frequency that each feature was selected during cross-validation for each data fusion variation. Although Early Fusion 2 and Late Fusion enforced an inherent balance between modality features, Early Fusion 1 combined all features into a single dataset prior to feature section, resulting in a majority of features being weighted towards the most predictive modality.

We find that the ratio of interaction-based features to posture-based features selected for all 4 folds (48 total features) is 25:23 for *bored*, 18:30 for *confused*, 22:26 for *engaged concentration*, 26:22 for *frustrated*, and 27:21 for *surprised*. The distribution of features was skewed towards interaction-based features for three of the affective states and toward posture-based features for two of the affective states, suggesting a comparable degree of predictive value between modalities across all affective states. This trend may explain why Early Fusion 2 and Late Fusion yielded the best-performing models for three of the five affective states examined (i.e., *bored*, *engaged concentration*, and *frustrated*). Both of these techniques allot equal emphasis on each modality and prevent individual modalities from monopolizing the feature set.

Results indicate that *confusion* was modeled most effectively using posture features only, which suggests that learner posture may be more indicative of confusion than interaction-based features extracted from TC3Sim log data. D'Mello and Graesser previously demonstrated a correlation between students' upright posture and instances of displayed confusion [24]. In aggregate, the results indicate that the predictive value of each modality varies across affective states, which in turn impacts the performance of Early Fusion and Late Fusion techniques. Utilizing dedicated models for each affective state, rather than inducing a single model to classify all affective states, enables the use of different modeling and data fusion techniques to yield improved detector performance. This also allows the multimodal framework to adapt to variance in feature balances for individual affective states.

It was observed that the most frequently selected features across all of the affective states were *sitmid_freq, sit_forward_freq, Sum of isSafe, CENTER_SHOULDER_max, sitmid_freq_20sec,* and *Min of HeartRate*. This indicates that each modality contributed relatively equally to the performance of the optimal multimodal classifiers. The two most frequent features (*sitmid_freq, sit_forward_freq*) were representative of the frequency that a learner adjusted their posture, while the two interaction-based features (*Sum of isSafe, Min of HeartRate*) were representative of the student's in-game actions and states, respectively. The remaining two features were also posture-based features: *CENTER_SHOULDER_max* focused on the furthest distance of the *CENTER_SHOULDER* vertex from the Kinect sensor over the entire session, and *sitmid_freq_20sec* focused on the learners' frequency of sitting upright for the preceding 20 seconds. A possible explanation for the improvement of the multimodal models' performance over the unimodal baselines is that the multimodal models were able to obtain a more thorough, comprehensive picture of the learners' behavior, as the most frequently used features were widely varied in the information provided.

7 Conclusion

Accurately detecting learner affect is a critical component of student modeling and has significant potential for guiding adaptive learning environments that support student learning and engagement. In this work, we have demonstrated the effectiveness of a multimodal affect detection framework that integrates interaction-based and posture-based data channels captured from undergraduate students engaging with a game-based learning environment for emergency military medical training. Results indicate that use of multiple independent modalities yields improved performance from multimodal detectors of five affective states compared to unimodal detectors that utilize only interaction-based or posture-based feature representations. We also explored several methods of multimodal data fusion to combine the two modalities and found that feature-level data fusion yielded the greatest predictive accuracy for three of the five affective states.

These results suggest several promising directions for future work. Investigating recent advances in multimodal machine learning techniques, including multimodal neural architectures, has strong potential to yield further improvements to the accuracy of multimodal affect detectors in adaptive learning environments. More sophisticated methods of data upsampling can be explored, as this might have a significant impact on classifier performance due to the pronounced imbalance and relatively small size of many learner-centered affective datasets. Generative models such as generative adversarial networks and variational autoencoders are upsampling methods that show particular promise. Finally, it will be important to investigate the run-time integration of multimodal affect detectors into game-based learning environments to enable adaptive features such as user-sensitive feedback or tailored scaffolding to improve learner engagement and support greater learning outcomes.

Acknowledgements. We wish to thank Dr. Jeanine DeFalco, Dr. Benjamin Goldberg, and Dr. Keith Brawner at the U.S. Army Combat Capabilities Development Command, Dr. Mike Matthews and COL James Ness at the U.S. Military Academy, and Dr. Robert Sottilare at SoarTech for their assistance in facilitating this research. The research was supported by the U.S. Army Research Laboratory under cooperative agreement #W911NF-13-2-0008. 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.

References

1. DeFalco, J., Rowe, J., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B., Baker, R., Lester, J.: Detecting and addressing frustration in a serious game for military training. International Journal of Artificial Intelligence in Education 28 (2), 152–193 (2018).
2. Bosch, N., D'Mello, S., Dame, N., Baker, R., Shute, V., Ventura, M., Wang, L., Zhao, W.: Detecting student emotions in computer-enabled classrooms. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, pp. 4125–4129 (2016).
3. Girardi, D., Lanubile, F., Novielli, N.: Emotion detection using noninvasive low cost sensors. In: Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction, pp. 125–130. IEEE (2017).
4. Soleymani, M., Asghari-Esfeden, S., Fu, Y., Pantic, M.: Analysis of EEG signals and facial expressions for continuous emotion detection. IEEE Transactions on Affective Computing 7 (1), 17–28 (2015).
5. Grafsgaard, J., Wiggins, J., Boyer, K., Wiebe, E., Lester, J.: Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue. In: Proceedings of the 7th International Conference on Educational Data Mining, pp. 122–129. International Educational Data Mining Society, London, UK (2014).
6. Botelho, A., Baker, R., Heffernan, N.: Improving sensor-free affect detection using deep learning. In: Proceedings of the International Conference on Artificial Intelligence in Education, pp. 40–51. Springer, Cham (2017).
7. Jiang, Y., Bosch, N., Baker, R., Paquette, L., Ocumpaugh, J., Andres, J., Moore, A., Biswas, G.: Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection? In: Proceedings of the International Conference on Artificial Intelligence in Education, pp. 198–211. Springer, Cham (2018).
8. Baker, R., Gowda, S., Wixon, M., Kalka, J., Wagner, A., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., Rossi, L.: Towards sensor-free affect detection in cognitive tutor algebra. In: Proceedings of the 5th International Conference on Educational Data Mining, pp. 126–133 (2012).
9. Bosch, N., D’Mello, S., Baker, R., Ocumpaugh, J., Shute, V.: Temporal Generalizability of face-based affect detection in noisy classroom environments. In: Proceedings of the International Conference on Artificial Intelligence in Education, pp. 44–53. Springer, Cham (2015).
10. Arroyo, I., Cooper, D., Burleson, W., Woolf, B., Muldner, K., Christopherson, R.: Emotion sensors go to school. In: Proceedings of the 14th International Conference on Artificial Intelligence In Education, pp. 17–24 (2009).
11. Grafsgaard, J., Wiggins, J., Boyer, K., Wiebe, E., Lester, J.: Automatically recognizing facial expression: Predicting engagement and frustration. In: Proceedings of the 6th International Conference on Educational Data Mining, pp. 43–50 (2013).
12. D’Mello, S., Kory, J.: A review and meta-analysis of multimodal affect detection systems. ACM Computing Surveys (CSUR) 47 (3), 43 (2014).
13. Paquette, L., Rowe, J., Baker, R., Mott, B., Lester, J., DeFalco, J., Brawner, K.,

Sottilare, R. and Georgoulas, V.: Sensor-Free or Sensor-Full: A comparison of data modalities in multi-channel affect detection. In: Proceedings of the 8th International Conference on Educational Data Mining, pp. 93–100 (2015).

14. Grafsgaard, J., Boyer, K., Wiebe, E., Lester, J.: Analyzing posture and affect in task-oriented tutoring. In: Proceedings of the International Conference of the Florida Artificial Intelligence Research Society, pp. 438–443 (2012).

15. Grafsgaard, J., Wiggins, J., Vail, A., Boyer, K., Wiebe, E., Lester, J.: The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In: Proceedings of the Sixteenth ACM International Conference on Multimodal Interaction, pp. 42–49. ACM (2014).

16. Bosch, N., Chen, H., Baker, R., Shute, V., D'Mello, S.: Accuracy vs. Availability heuristic in multimodal affect detection in the wild. In: Proceedings of the 17th ACM International Conference on Multimodal Interaction, pp. 267–274 (2015).

17. Sottilare, R., Baker, R., Graesser, A., Lester, J.: Special Issue on the Generalized Intelligent Framework for Tutoring (GIFT): Creating a stable and flexible platform for innovations in AIED Research. International Journal of Artificial Intelligence in Education 28 (2), 139–151 (2018).

18. Baker, R., Ocumpaugh, J., Andres, J.: BROMP Quantitative Field Observations: A Review. In: Learning Science: Theory, Research, and Practice. McGraw-Hill, New York, NY (2018).

19. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement 20 (1), 37–46 (1960).

20. Baltrušaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2), 423–443 (2018).

21. Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P., Paiva, A.: Automatic analysis of affective postures and body motion to detect engagement with a game companion. In: Proceedings of the 6th International Conference on Human-robot Interaction, pp. 305–312. ACM (2011).

22. Henderson, N., Rowe, J., Mott, B., Brawner, K., Baker, R., Lester, J.: 4D affect detection : Improving frustration detection in game-based learning with posture-based temporal data fusion. In: Proceedings of the 20th International Conference on Artificial Intelligence in Education, pp. 144–156 (2019).

23. Chawla, N. V., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16 (1), 321–357 (2002).

24. D'Mello, S., Graesser, A.: Mining bodily patterns of affective experience during learning. In: Proceedings of the 3rd International Conference on Educational Data Mining, pp. 31–40 (2010).