4D Affect Detection: Improving Frustration Detection in Game-Based Learning with Posture-Based Temporal Data Fusion

Nathan L. Henderson[1], Jonathan P. Rowe[1], Bradford W. Mott[1], Keith Brawner[2],
Ryan Baker[3], and James C. Lester[1]

[1] North Carolina State University, Raleigh, NC, 27695, USA
{nlhender, jprowe, bwmott, lester}@ncsu.edu

[2]U.S. Army Combat Capabilities Development Command,
Orlando, FL, 32826 USA
keith.w.brawner.civ@mail.mil

[3]University of Pennsylvania, Philadelphia, PA, 19104, USA
rybaker@upenn.edu

Abstract. Recent years have seen growing interest in utilizing sensors to detect learner affect. Modeling frustration has particular significance because of its central role in learning. However, sensor-based affect detection poses important challenges. Motion-tracking cameras produce vast streams of spatial and temporal data, but relatively few systems have harnessed this data successfully to produce accurate run-time detectors of learner frustration outside of the laboratory. In this paper, we introduce a data-driven framework that leverages spatial and temporal posture data to detect learner frustration using deep neural network-based data fusion techniques. To train and validate the detectors, we utilize posture data collected with Microsoft Kinect sensors from students interacting with a game-based learning environment for emergency medical training. Ground-truth labels of learner frustration were obtained using the BROMP quantitative observation protocol. Results show that deep neural network-based late fusion techniques that combine spatial and temporal data yield significant improvements to frustration detection relative to baseline models.

Keywords: Affect detection, data fusion, posture, frustration, deep learning.

1 Introduction

Affect has a key role in shaping student learning outcomes [1]. Affective states such as *flow* tend to promote learning, while states such as *boredom* are not as conducive to learning. The affective state of *frustration* has a complex relationship with learning [2–5]. On the one hand, frustration often coincides with student efforts to overcome impasses, and it signifies situations in which students are grappling with a concept that is challenging [6]. On the other hand, frustration can lead to student disengagement, and it has been correlated with negative learning outcomes [7]. The ability to accurately detect student affect at run-time is critical to the development of affect-sensitive learning technologies that dynamically intervene to support engagement and emotion regulation [3, 8]. The complex relationship between frustration and learning underscores the importance of reliable frustration detection to inform how affect-sensitive pedagogical interventions are delivered within intelligent tutoring systems [3].

Several methods for detecting student frustration have been investigated in recent years. These include both *sensor-free* methods and *sensor-based* methods. Sensor-free methods leverage trace log data from student interactions with a learning environment to train machine learning-based models of affect [9, 10]. Results have shown that sensor-free affect detection, in combination with deep recurrent neural networks, can yield accurate models across several affective states [9]. Alternatively, sensor-based methods utilize physical sensors to capture trace-level data on learner behavior and physiology, including facial expression, eye gaze, electrodermal activity (EDA), electroencephalography (EEG), and posture [3, 4, 11]. Sensor-based methods show promise for enabling *generalized affect detection*, which eschews domain-specific input feature representations, instead leveraging sensor data that can be gathered across a range of educational domains and learning environments. Notably, sensor-based approaches to affect detection do not necessarily require specialized hardware because a growing number of sensors are built directly into computers and tablets, including webcams, motion-tracking cameras, and increasingly, eye trackers.

Sensor-based frustration detection has shown good results when targeting self-reported affect data [12] or deploying sensors in laboratory settings [4]. Specific data channels, such as facial expression, have also shown promise using student data from classrooms [13], but other data channels, such as posture, have received less attention. Sensor-based frustration detection outside of the lab raises significant challenges [3]. Physical sensor data can be affected by reliability issues, background noise, poor calibration, subject mistracking, data storage constraints, and inconsistent sensor configurations. Further, trace-level data generated by sensors is intrinsically temporal, yet the input feature representations that are distilled from these data streams often contain limited temporal information [3]. Spatiotemporal data has been demonstrated to significantly improve the performance of sensor-based classifiers for action recognition [14] and engagement intensity [15], and it is likely to benefit affect-sensitive learning technologies as well.

In this paper, we investigate sensor-based frustration detection using deep neural network-based data fusion techniques integrating spatial and temporal data on student posture captured by Microsoft Kinect cameras. The dataset was gathered from a study involving students using a game-based learning environment for emergency medical training, TC3Sim. Ground-truth labels for learner frustration were obtained using the BROMP quantitative observation protocol [16]. We compare the effectiveness of deep neural network-based early- and late-fusion techniques across several evaluation metrics. Results show that deep neural network-based late-fusion yields significant improvements to frustration detection compared to several baseline techniques.

2 Related Work

There is growing interest in sensor-based affective modeling in advanced learning technologies. Bosch et al. [13] utilized webcam recordings of students engaged in a physics-based learning game to construct feature vectors extracted from observed head positions and movement, brow position, and gross body movement. Ground truth data was obtained through the BROMP protocol, using trained observers to mark instances of certain affective states at set time intervals. Utilizing BROMP observations as a target label, a multitude of classifiers were trained, including Bayesian classifiers and C4.5 decision trees, to detect affective states such as frustration, boredom, confusion, delight, and engagement. Motion-tracking cameras, such as the Microsoft Kinect, have also been utilized in sensor-based affect detection [17]. Grafsgaard et al. utilized learner posture and gesture data gathered by a Microsoft Kinect as learners engaged in computer-mediated tutoring sessions for introductory programming [17]. Posture estimation vectors were distilled from the Kinect's depth-channel data, and the vectors were used to determine correlations between specific postures and self-reported frustration, engagement, and learning gains. DeFalco et al. [3] utilized posture data from a Kinect sensor to detect learner affect in a game-based learning environment for emergency medical training. Separate classifiers were induced for each of five affective states: boredom, confusion, concentration, frustration, and surprise. The affect detectors performed only slightly better than chance, yielding Kappa values between 0 and 0.11.

As an alternative to sensor-based affect detection, Jiang et al. [10] utilized interaction trace log data in an investigation of deep neural network-based representation learning versus expert feature-engineering for sensor-free affect detection using BROMP data. Time, frequency, and ratio-based features were calculated for each student based on his/her individual interaction with a game-based learning environment for physics education. Overall, deep neural network-based models achieved equal or better performance compared to feature engineering-based models, with a lone exception being frustration (i.e., the feature-engineering approach was slightly more accurate). Subsequent work showed that recurrent neural networks (RNNs) outperformed the previous classification algorithms in the same affect detection task [9].

Recent efforts in affect detection have started to explore usage of temporal data channels as an input modality. Yang et al. [15] used several feature extraction approaches on spatiotemporal face and posture data to train long short-term memory (LSTM) networks alongside regression fusion to approximate engagement intensity in individuals watching an educational video. Temporal information has also been used to develop rule-based models to classify affect through recognition of sequences of joint movement and repetition of certain motions [18]. The frustration detection framework presented in this paper builds on recent advances in deep neural network-based data fusion and introduces an artificial temporal data stream (i.e., a "fourth dimension") derived from spatial 3D posture data to enhance run-time detector accuracy during student interactions with a game-based learning environment.

3 TC3Sim Game-Based Learning Environment

We investigate automatic detection of student frustration in the context of a game-based learning environment for training military medical personnel, the Tactical Combat Casualty Care Simulation (TC3Sim). Developed by Engineering and Computer Simulations (ECS), TC3Sim (Fig. 1) is widely used by the U.S. Army to train soldiers in the essential procedures required of an Army Combat Medic or Combat Life Saver.

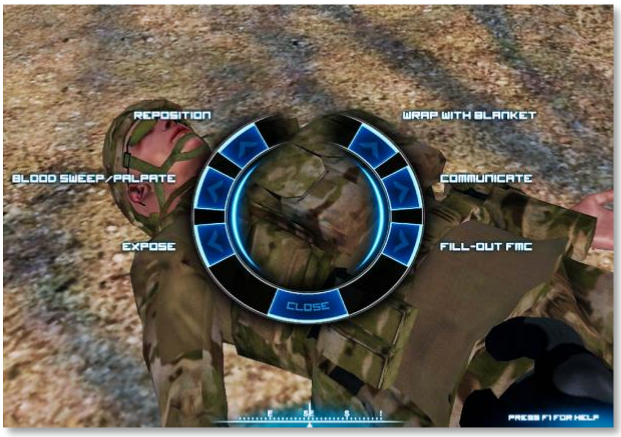

Fig. 1. Screenshot of injured soldier in TC3Sim.

In TC3Sim, trainees complete a series of 3D simulated combat missions alongside a group of computer-controlled teammates. Each story-driven training scenario includes a series of simulated combat events that lead to the eventual injury of one or more teammates. Trainees must administer tactical combat casualty care in real-time, which includes securing the area, assessing casualties, performing triage, administering treatment, and preparing for medical evacuation. Trainees encounter opportunities to handle a wide range of injuries, including cuts, puncture wounds, blocked airways, amputations, and burns. In the present work, we focus on learner interactions with four training scenarios from TC3Sim, including a tutorial scenario, a leg injury scenario, a narrative scenario involving a squad of soldiers on patrol, and a final scenario that is impossible to complete successfully—the patient expires regardless of treatment. Prior work with TC3Sim has found evidence of a negative relationship between frustration and learning, and further, motivational feedback interventions that target frustration can positively impact learning outcomes [3]. We seek to improve the effectiveness of generalizable frustration detectors to enable affect-sensitive pedagogical support with enhanced effectiveness and reliability.

4 Detecting Frustration with Posture-Based Temporal Data Fusion

The primary goal of this work was to induce machine learning-based classifiers for run-time frustration detection using student posture data collected by a Microsoft Kinect sensor. The detector's objective was to classify whether a student was frustrated or not given an input feature vector consisting of spatial and/or temporal posture data.

4.1 Dataset

We utilized a previously published dataset containing data from 119 students (83% male, 17% female) at the United States Military Academy. The training materials were administered using the Generalized Intelligent Framework for Tutoring (GIFT), an open-source software framework for building and deploying adaptive training systems [19]. All participants worked individually at laptops and received the same materials; there were no experimental conditions. Study sessions lasted approximately 1 hour.

The study procedure was as follows. First, learners completed a brief demographic questionnaire and content pre-test. Next, they viewed a PowerPoint presentation about tactical combat casualty care. Afterward, participants completed a series of training scenarios in TC3Sim, each working at their own pace. The session concluded with a brief post-test, which included the same knowledge assessment items that were presented on the pre-test. Utilizing identical items on both the pre- and post-tests reduced the challenge of identifying items with matching difficulty for counterbalancing the assessments. Further, no feedback was given about student performance on the pre-test during the study.

During the study sessions, each participant was instrumented with a tripod-mounted Microsoft Kinect for Windows 1.0 sensor. The Kinect sensor was positioned in front of each participant to capture all head movements, body movements, and gestures throughout participants' interactions with TC3Sim using built-in skeletal-tracking features supported by GIFT. Kinect sensor data was recorded at approximately 10-12 Hz. The data consisted of a series of timestamped feature vectors containing 3D coordinate data for 91 vertices, each corresponding to a facial or body joint tracked by the Kinect. In addition to the Kinect, learners were equipped with a wireless Affectiva Q-Sensor bracelet, and their interaction trace log data was recorded by GIFT. The Q-Sensors captured timestamped data on learners' skin temperature, learners' electrodermal activity, and sensor 3D coordinates as measured by built-in accelerometers. However, the Q-Sensor data contained significant recording gaps for a large number of participants, and therefore it was not utilized in the current work. The interaction trace log data was not relevant to devising sensor-based frustration detection, so it was also not utilized.

To obtain ground-truth labels of learner affect, two field observers recorded learners' affect and behavior using the BROMP quantitative field observation protocol throughout the study [16]. The field observers, who were both BROMP-certified coders, walked around the perimeter of the classroom and used a hand-held Android device running the HART field-observation software to discreetly record each learner's affect and behavior at 20-second intervals in round robin sequence. The following emotional states were recorded: *Concentrating*, *Confused*, *Boredom*, *Surprised*, *Frustrated*, *Contempt*, and *Other*.

In total, the study yielded 3,066 BROMP observations by the two field observers. For the purpose of the current analysis, we utilize a subset of 755 observations coinciding with the time period during which participants interacted with the TC3Sim game-based learning environment and on which there was no disagreement between BROMP coders about the occurrence of a target affective state. The distribution of affective states across these observations were the following: 435 (57.6%) were coded as *Concentrating*, 174 (23.1%) as *Confused*, 73 (9.7%) as *Boredom*, 32 (4.2%) as *Frustrated*, 29 (3.8%) as *Surprised*, and 12 (1.6%) as *Contempt*.

To prepare the data for training posture-based frustration detectors, we re-coded the data into binary categories, yielding 32 instances of *Frustrated* and 723 instances of *Not-Frustrated*. The Kinect data was cleaned to remove instances of tracking anomalies and extraneous vertex data. Sessions containing fewer than 3 BROMP observations were also removed. Of the 91 vertices tracked by the Kinect, 3 were utilized for posture-based frustration detection: top_skull, head, and center_shoulder. These vertices were selected based on prior efforts investigating affect detection from Kinect data [17]. Next, Kinect and BROMP data were integrated and temporally aligned. A set of 73 posture-related features were computed for each BROMP observation after the initial data collection, serving as input features for frustration detection. These features captured spatial information about student posture, and they included summary statistics (e.g., median, variance, min, max) calculated over time windows of 5, 10, and 20 seconds preceding the BROMP coding event. These time window sizes are similar to prior work on affect detection, including a maximum window size that corresponds to the targeted maximum time between BROMP observations [3, 13]. In addition, features capturing aggregate changes in learner posture, as well as forward/backward lean behaviors, were computed. In aggregate, these features provided a detailed view of the spatial orientation of learners' posture.

4.2 Temporal Feature Engineering

The spatial features that were distilled from the Kinect posture data had ranges that varied widely, so feature scaling was performed. Each student's data was normalized using Z-score standardization: for each session, the difference between a single data point and session mean was divided by the session standard deviation. Temporal posture features were computed from the spatial posture feature vectors using the first derivative of each observation's posture coordinates [20]. Using the head vertex, for each set of (x, y, z) posture coordinates, the coordinate deltas across two consecutive sensor readings were calculated. The deltas were used to calculate *velocity features* averaged across time windows of 3, 5, 10, and 20 seconds. For each posture coordinate, the mean, median, max, and variance of the average corresponding velocity were calculated. This process provided an additional 48 temporally-related posture features. Due to the large number of additional features calculated per vertex, velocity information was not calculated for center_shoulder and top_skull. The temporal data was normalized using the same Z-score normalization described previously.

4.3 Feature Selection

Given the large number of available posture features, automated feature selection was utilized to reduce the size of the final feature sets for training frustration detectors. Forward feature selection was performed to investigate alternate configurations of feature vectors up to length 10. Greedy feature selection was performed using RapidMiner 9.0, and it was guided by classification performance with the sequential minimal optimization (SMO) variation of a polynomial-kernel support vector machine (SVM) [21]. We utilize RapidMiner because it is a convenient toolkit for processing and modeling data using a range of supervised learning algorithms, and it has been used widely in prior work on affect detection [3]. Forward feature selection is a common approach in prior work on affect detection, and SVMs trained with SMO have previously been found to outperform competing algorithms for frustration detection with learner data from the TC3Sim environment [3, 13].

4.4 Deep Neural Network Architecture

Each dataset produced by the feature-selection algorithm was used to train a multi-layer perceptron neural network. Each network was comprised of feed-forward layers containing 800, 800, 500, 100, 50, and 2 nodes, respectively. Each hidden layer utilized a Rectified Linear Unit (ReLU) activation function. The networks were trained for 10 epochs and used an ADADELTA [22] adaptive learning rate to help prevent overfitting. All deep neural network models were implemented using RapidMiner 9.0 [23].

4.5 Data Fusion

To investigate alternate approaches for integrating spatial and temporal posture features, we compared several classifiers induced with both early- and late-fusion techniques. Early fusion is based on the concept of "feature-level" fusion, or concatenation of multiple feature vectors to form a single vector prior to supervised learning [24]. To determine the best sequence of feature selection and feature-level fusion, we implemented two variants of early fusion. The first method, EarlyFusion1, performs feature selection after concatenating the spatial and temporal feature vectors. (Fig. 2A). The second method, EarlyFusion2, performs feature selection on spatial and temporal features separately. After feature selection, feature-level fusion is performed

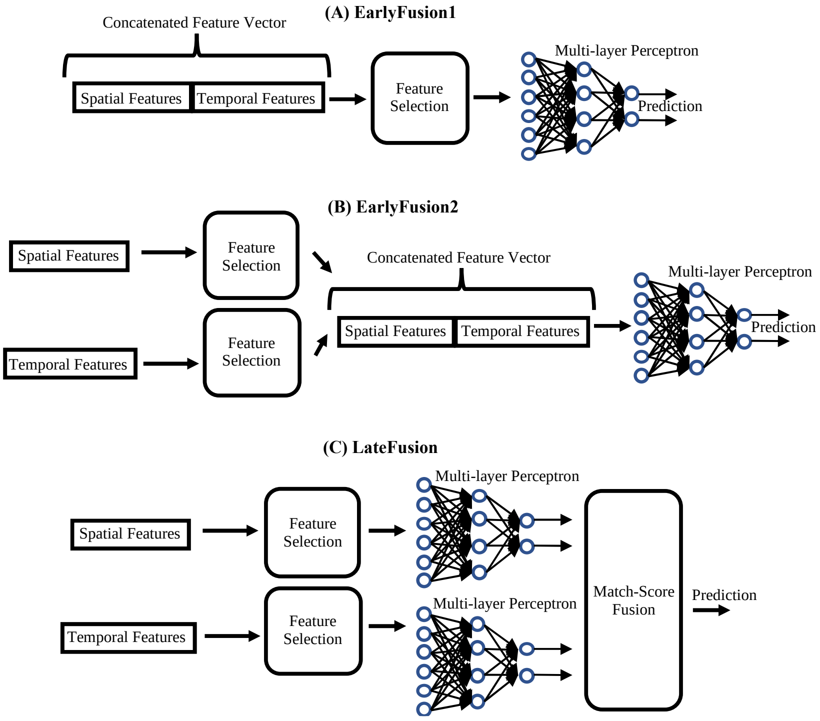

Fig. 2. Three data fusion techniques for integrating spatial and temporal posture-based frustration detection methods.

on the top-selected features from each modality (Fig. 2B). LateFusion involves training a model on each modality separately and integrating the results of each classifier to produce a single prediction (Fig. 2C). This prediction can be determined using several different methods, such as majority voting, averaging, or weighting [25]. In this work, we compare the results of late fusion using match-score fusion [26] and the highest confidence level of the late-fusion output.

5 Results

Frustration detectors were trained using 10-fold student-level cross validation. Data splits were maintained across all modeling approaches to ensure fair comparisons. To ensure adequate training coverage for both target classes (i.e., *frustrated* and *not-frustrated*), the training data was oversampled using cloning of minority class instances. Feature selection and early fusion techniques were implemented in RapidMiner 9.0 [23]. RapidMiner does not support decision-level fusion, as required by our LateFusion method. Therefore, feature selection and deep neural network models were created using RapidMiner, the raw outputs of the models were recorded, and then decision-level fusion was performed outside of RapidMiner using Python.

We observed that z-score feature normalization has a sizable impact on the predictive accuracy of posture-based frustration detectors. As a baseline, we reproduced a machine learning pipeline for training SVM-based frustration detectors using spatial posture data, which had been previously reported in [3], and we investigated how the resulting models compared to an equivalent machine learning pipeline with z-score feature normalization added. Evaluation metrics included Cohen's kappa [27], area under the curve (AUC), total accuracy, and F1 score. Results are shown in Table 1.

Table 1. Effect of z-score normalization on sensor-based frustration detection using spatial posture features.

Classifier	Kappa	AUC	Accuracy	F1 Score
SVM	0.056	0.600	0.687	0.113
SVM (Normalized)	0.190	0.500	0.737	0.249

Based on these results, z-score normalization was used for the remainder of the analyses reported in this section. Next, we replaced the SVM classifier with the deep neural network described in Section 4.5. Results from comparing the deep neural network-based frustration detector with the SVM-based detector are shown in Table 2. The deep neural network model did not show significant improvement regarding Kappa and F1 score, and even displayed a decrease in the raw accuracy compared to the SVM model. However, there was substantial improvement in the AUC measurement. Slight increases in Kappa and F1 score, as well as AUC score indicated that the neural network had the potential to capture complex patterns in the training data possibly not detected by the SVM.

Next, temporal posture features were computed from the Kinect data, normalized, and used as input for the three data fusion methods. For the LateFusion model, two

Table 2. Comparison of SVM and deep neural network models for spatial posture-based frustration detection under 10-fold student level cross validation.

Classifier	Kappa	AUC	Accuracy	F1 Score
SVM	0.190	0.500	0.737	0.249
Deep Neural Network	0.192	0.808	0.685	0.254

different selection schemes were tested. The first selection scheme used the model prediction with the highest confidence level. The second selection scheme took the average of all confidence levels for a predicted class and used the highest average, similar to match-score fusion [26]. However, detector accuracy did not change when the two selection methods were interchanged. This may be due to the high confidence levels of the classifiers for this particular data set, as well as the relatively small amount of test data available.

Results from a comparison of early- and late-fusion methods combining spatial and temporal posture data are shown in Table 3 alongside results from the deep neural network trained with spatial posture data only as a baseline. Best results for each evaluation metric are shown in bold. It is apparent that the addition of temporal feature information improved the quality of frustration detection, particularly for the LateFusion model. Due to the high proportion of non-frustration observations versus frustration observations in the test data, additional emphasis is placed on the Cohen's kappa metric, as it accounts for the potential of obtaining true-positives by chance.

Table 2. Results of early fusion and late fusion on posture and temporal feature data.

Classifier	Kappa	AUC	Accuracy	F1 Score
Baseline Network	0.192	0.808	0.685	0.254
EarlyFusion1 Network	0.178	0.780	0.845	0.213
EarlyFusion2 Network	0.281	**0.854**	0.900	0.321
LateFusion Network	**0.355**	0.809	**0.906**	**0.396**

EarlyFusion2 outperformed EarlyFusion1 across all evaluation metrics. This may be attributable to the dimensionality of the datasets used to train the respective models. Because feature selection operated on a single data stream for EarlyFusion1, the main difference between this model and the baseline deep neural network was the set of candidate features subjected to SVM-based feature selection, as EarlyFusion1 concatenated temporal velocity features with spatial posture features prior to feature selection. The temporal posture features added 48 additional attributes to the existing 73 spatial posture features, but feature selection only returned up to 10 features in each scenario. Alternatively, in EarlyFusion2, two separate feature selection processes are employed in parallel, yielding a maximum of 20 features as input to the neural network. This increase in number of attributes is a possible explanation for the improved accuracy of EarlyFusion2 over EarlyFusion1.

Late fusion offers a different approach due to its capacity to "correct" a single model's prediction during circumstances where the model's confidence level is relatively low. Upon closer examination, several instances were observed when the spatial posture-based detector made an incorrect prediction with a low confidence level, and the temporal posture-based model made a correct prediction with a high confidence level, and the latter was chosen as the representative prediction during match-score fusion. Several instances of the inverse scenario—the spatial posture-based model corrected a prediction by the temporal posture-based model—were also observed. This interaction contributed to the increased accuracy of LateFusion frustration detection over baseline SVM and deep neural network models, as well as the early fusion methods.

6 Conclusion

Detection of learner frustration is critical to the creation of affective-sensitive learning technologies. However, devising sensor-based run-time models of learner frustration using posture data poses significant challenges. We have introduced a data-driven framework that combines deep neural network-based data fusion and spatiotemporal representations of posture data to improve run-time models of frustration detection. Posture features were distilled using sensor data collected from participants engaging with a game-based learning environment for emergency medical training. We found that late fusion methods combining deep neural network-based frustration detectors trained with spatial and temporal posture feature data outperform several baseline techniques, including early fusion-based models and spatial posture-based models.

The results suggest several promising directions for future research. First, it will be important to investigate whether posture-based temporal data fusion techniques are transferable to other learning emotions (e.g., boredom, confusion, engaged concentration, surprise) as well as other learning environments. A key promise of sensor-based affect detection is the potential for creating computational models of learner affect that generalize across different educational subjects and settings. Second, alternate deep neural network architectures should be investigated, particularly those that are explicitly designed for modeling sequential data, such as recurrent neural networks, to better capture the temporal dynamics of affect as expressed through posture. Recent work has shown that recurrent neural network architectures, such as long short-term memory networks, yield significant improvements to sensor-free affect detection, but it remains to be seen how these methods are best utilized in sensor-based models of affect. Finally, there is significant promise in integrating posture-based temporal data fusion techniques for affect detection into run-time learning environments, enabling delivery of dynamic interventions designed to support student engagement and foster improved learning.

Acknowledgements. We wish to thank Dr. Jeanine DeFalco and Dr. Benjamin Goldberg at the U.S. Army Combat Capabilities Development Command – Simulation and Training Technology Center (CCDC-STTC), Dr. Mike Matthews and COL James Ness at the United States 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. D’Mello, S.: A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. Journal of Educational Psychology. 105, 1082–1099 (2013).
2. Grafsgaard, J.F., Wiggins, J.B., Vail, A.K., Boyer, K.E., Wiebe, E.N., Lester, J.C.: 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).
3. DeFalco, J.A., Rowe, J.P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B.W., Baker, R.S., Lester, J.C.: Detecting and addressing frustration in a serious game for military training. International Journal of Artificial Intelligence in Education. 28, 152–193 (2018).
4. Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.C.: Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue. In: Proceedings of the Seventh International Conference on Educational Data Mining. pp. 122–129. International Educational Data Mining Society, London, UK (2014).
5. Harley, J.M., Bouchet, F., Azevedo, R.: Aligning and comparing data on emotions experienced during learning with MetaTutor. AIED. 7926, 61–70 (2013).
6. Pardos, Z., Baker, R., Pedro, M.S., Gowda, S.M., Gowda, S.M.: 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, 107–128 (2014).
7. D’Mello, S., Graesser, A.: The half-life of cognitive-affective states during complex learning. Cognition and Emotion. 25, 1299–1308 (2011).
8. Cooper, D.G., Arroyo, I., Woolf, B.P.: Actionable affective processing for automatic tutor interventions. In: New perspectives on affect and learning technologies. pp. 127–140. Springer, New York, NY (2011).
9. Botelho, A.F., Baker, R.S., Heffernan, N.T.: 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).
10. Jiang, Y., Bosch, N., Baker, R.S., Paquette, L., Ocumpaugh, J., Andres, J.M.A.L., Moore, A.L., 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).
11. Bosch, N., D’mello, S.K., Ocumpaugh, J., Baker, R.S., Shute, V.: Using Video to Automatically Detect Learner Affect in Computer-Enabled Classrooms. ACM Transactions on Interactive Intelligent Systems. 6, 1–26 (2016).
12. Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K., Christopherson, R.: Emotion sensors go to school. In: Artificial Intelligence In Education. pp. 17–24 (2009).
13. Bosch, N., Mello, S.K.D., Dame, N., Dame, N., Baker, R.S., Shute, V., Ventura, M., Wang, L., Zhao, W.: Detecting student emotions in computer-enabled classrooms. Proceedings of the 25th International Joint Conference on Artificial

Intelligence. 4125–4129 (2016).

14. Henderson, N., Aygun, R.: Human Action Classification Using Temporal Slicing for Deep Convolutional Neural Networks. In: 2017 IEEE International Symposium on Multimedia (2017).

15. Yang, J., Wang, K.: Deep Recurrent Multi-instance Learning with Spatio-temporal Features for Engagement Intensity Prediction. In: Proceedings of the 2018 on International Conference on Multimodal Interaction. pp. 594–598. ACM (2018).

16. Ocumpaugh, J., Baker, R.S., Rodrigo, M.T.: Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. (2015).

17. Grafsgaard, J., Boyer, K., Wiebe, E., Lester, J.: Analyzing posture and affect in task-oriented tutoring. In: FLAIRS Conference. pp. 438–443 (2012).

18. Patwardhan, A., Knapp, G.: Multimodal affect recognition using Kinect. arXiv preprint arXiv:1607.02652. (2016).

19. Sottilare, R.A., Baker, R.S., Graesser, A.C., Lester, J.C.: 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, 139–151 (2018).

20. Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P.W., 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).

21. Platt, J.C.: Sequential minimal optimization: A fast algorithm for training support vector machines. 1–21 (1998).

22. Zeiler, M.D.: ADADELTA: An adaptive learning rate method. (2012).

23. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M.: Yale: Rapid prototyping for complex data mining tasks. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 935–940 (2006).

24. Soleymani, M., Pantic, M., Pun, T.: Multimodal emotion recognition in response to videos. IEEE transactions on affective computing. 3, 211–223 (2012).

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

26. Rahman, W., Gavrilova, M.L.: Emerging EEG and Kinect face fusion for biometric identification. In: Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1–8. IEEE (2017).

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