Temporal Generalizability of Face-Based Affect Detection in Noisy Classroom Environments

Nigel Bosch[1], Sidney D'Mello[1,2], Ryan Baker[3], Jaclyn Ocumpaugh[3], Valerie Shute[4]

Departments of Computer Science[1] and Psychology[2], University of Notre Dame
Notre Dame, IN 46556, USA
Department of Human Development[3], Teachers College, Columbia University
New York, NY 10027, USA
Department of Educational Psychology and Learning Systems[4], Florida State University
Tallahassee, FL 32306-4453, USA
pbosch1@nd.edu, sdmello@nd.edu, baker2@exchange.tc.columbia.edu, jocumpaugh@wpi.edu, vshute@fsu.edu

Abstract. The goal of this paper was to explore the possibility of generalizing face-based affect detectors across multiple days, a problem which plagues physiological-based affect detection. Videos of students playing an educational physics game were collected in a noisy computer-enabled classroom environment where students conversed with each other, moved around, and gestured. Trained observers provided real-time annotations of learning-centered affective states (e.g., boredom, confusion) as well as off-task behavior. Detectors were trained using data from one day and tested on data from different students on another day. These cross-day detectors demonstrated above chance classification accuracy with average Area Under the ROC Curve (AUC, .500 is chance level) of .658, which was similar to within-day (training and testing on data collected on the same day) AUC of .667. This work demonstrates the feasibility of generalizing face-based affect detectors across time in an ecologically valid computer-enabled classroom environment.

1 Introduction

Students experience various affective states that influence learning in striking ways [1, 2]. For example, boredom has been shown to be negatively related to learning in multiple computerized learning environments [3, 4], while engagement has been shown to be positively associated with learning [4]. Affect has also been shown to influence learning by modulating cognitive and motivational processes in striking ways (see [5] for a review). Given the importance of affect to learning, researchers have been creating computerized learning environments that automatically detect and respond to students' affective states [6]. For example, one experiment comparing an affect-sensitive intelligent tutoring system (ITS) to the same ITS without affect sensitivity found that learners with low prior knowledge learned significantly more (d = .713) from the affect-sensitive version compared to the plain version of the ITS [7]. Despite the success of such affect-sensitive learning environments, work remains to

adfa, p. 1, 2011.

be done to enable affect-sensitivity to function in contexts outside of a laboratory study, such as a noisy classroom. One key challenge involves the development of accurate affect detectors that can function in computer-enabled classrooms. Some work is being done in this area ([4, 8]).

Another important issue is temporal generalizability (generalization over time), which seeks to ascertain whether a detector trained on data from one day will still work well when classifying data from a different day. Such a detector might not work well because the features it uses may be influenced by factors specific to a day. Physiology-based sensors have been shown to suffer from such differences [9]. For example, a student's average skin conductance (a physiological feature) may change from one day to the next. Thus, detectors created using skin conductance data from one day may not work well on a different day unless specific measures are taken to compensate for the day-to-day differences.

In this paper we focus on face-based affect detection. Potential causes of day differences in face-based affect detection include lighting in the classroom (which can change how well computer vision algorithms detect facial features), students' mood (which might alter their affect and facial expressions), number of students (which might influence how distracted students are by their friends and how much they converse with each other), and perhaps other factors. However, whether day-to-day differences impact face-based affect detection accuracy (as they do physiology-based detectors) is currently unknown, and is the central focus the current paper.

Related Work. Many different modalities (e.g., facial features, physiology, audio) have been proposed and evaluated for affect detection [10]. To keep the scope manageable, we focus on affect detection efforts in educational contexts and those testing differences across days.

Interaction data from log-files has shown promise for building affect detectors that generalize across time. Pardos et al. [4] used interaction data collected over the span of a few days in 2010 to build affect detectors. These detectors were then applied to a separate, previously collected dataset from two school years (Fall 2004-Spring 2006). The detectors' predictions were correlated with students' scores on a standardized test. Several of these correlations demonstrated the consistency of detectors across two school years. Predicted boredom (r = -.119 for year 1, r = -.280 for year 2), confusion (r = -.165, r = -.089), and gaming the system (r = -.431, r = -.301) negatively correlated with test score in both school years, while engaged concentration (r = .449, r = .258) positively correlated in both years. However, they did not directly test cross-year generalization by building detectors on one year of data and testing on the other.

Physiology has been used for affect detection with channels such as skin conductance and heart rate [8, 11]. However, multiple studies have observed degraded affect detection performance using physiological data when models trained using data from one day are tested using data from another day [9, 11]. In a classic study, Picard et al. [11] found that physiological data were more tightly clustered by affective state within data from the same day than were data from another day. The data distribution parameters become less reliable due to changing factors like mood and attention, and the decision boundaries for classifiers became less effective for discriminating instances of affect from a later day.

Cameras are a ubiquitous part of modern computers, from tablets to laptops to webcams, so face-based affect detection is an attractive option compared to modalities that require special equipment like skin conductance sensors or heart rate monitors. A variety of approaches have been used for face-based affect detection [10, 12]. Frustration [13, 14], engagement [13, 15], confusion [16], and other learning-centered affective states can be detected using facial features. However, many of these and other studies have taken place in a lab environment where data were collected one or two students at a time over the course of a few months. The conditions in lab-based studies are typically tightly controlled in an effort to reduce outside influences on the outcomes of studies. Changes like lighting and mood, and differences in classes (e.g., number of students, teaching strategies) that could influence affect detection may not be salient, unlike in a more ecological data collection setting such as a computer-enabled classroom. Thus, temporal generalizability of face-based affect detectors is currently an open question with important practical implications.

Current Study. To assess the ability of face-based affect detection to generalize over time in the ecologically valid setting of a computer-enabled classroom, the current paper attempts to answer three novel questions: 1) how does performance change when affect detectors are trained on one day and tested on another? (*cross-day*) compared to training and testing on data collected on the same day (*within-day*); 2) how do model and data parameters differ between the best-performing models built using data from different days? and 3) how much do different cross-day models rely on the same features for affect detection? The novelty of our contribution is that, to the best of our knowledge, this is the first paper to attempt to explore the possibility of face-based affect detection generalizing across time.

This paper uses a dataset that has been used for previous work on face-based affect detection [17]. Face videos were recorded in a noisy computer-enabled classroom environment where up to thirty students at a time played an educational physics game. Students talked to others and themselves, gestured, and occasionally left for bathroom breaks. Factors such as number of students in a class, lighting, and time of day varied as well. A subset of learning-centered affective states (boredom, confusion, delight, engagement, frustration, and off-task behavior) were detected using in a student-independent fashion to ensure generalization to new students. Detection was successful with average Area Under the ROC Curve (AUC) of .709. Data were collected over two days (with a 3-day interval). However, data from two days were pooled together for building detectors, and thus there was no evidence of generalization across days. In the current study we compare results of generalization against the baseline standards established in this previous work by training detectors on data collected on one day and testing them on data collected on a second day [17].

2 Method

A more thorough treatment of the data collection procedure and the model building method used in this study can be obtained by examining [17]. In this paper, we focus on only the most important aspects and those closely related to the goals of this paper.

2.1 Data

The data were collected from 137 8th and 9th grade students (57 male, 80 female) who were enrolled in a public school in a medium-sized city in the Southeastern U.S. Students were tested in groups of about 20 students per class period (55 minutes per period). The study took place in one of the school's computer-enabled classrooms, which was equipped with about 30 desktop computers. Inexpensive webcams ($30) were affixed at the top of the monitor on each computer.

Students played Physics Playground [18], a two-dimensional game that requires the player to apply principles of Newtonian physics in an attempt to guide a green ball to a red balloon (key goal) in many challenging configurations. Fig. 1 illustrates a configuration requiring a ramp drawn to build up speed before launching up to red balloon. Everything in the game obeys the basic laws of physics relating to gravity and Newton's three laws of motion. Students' affective states and on-task vs. off-task behaviors were coded using the Baker-Rodrigo Observation Method Protocol (BROMP) field observation system [19]. The affective states of interest were boredom, confusion, delight, engaged concentration, and frustration. This list of states was selected based on previous reviews on affect during learning with technology [1] and from observing students during pilot data collection (these data were not used in the current models). Delight was only added to the list of affective states on the second day, so cross-day generalization testing could not be performed for delight. In addition to affect, students were coded as being: *on task* when working on their own computer, *on-task conversation* when conversing with other students about the task, and *off task* in other situations (e.g., task-unrelated conversation, watching other students without conversation, using a cellphone).

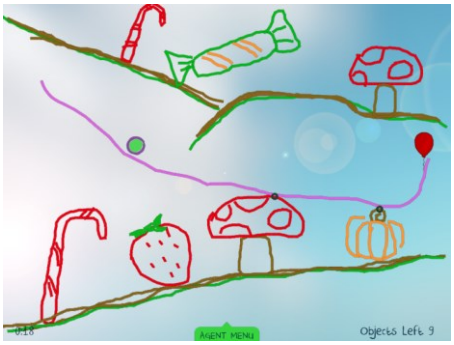

Fig. 1. Ramp solution in Physics Playground.

We consider data from two days of game-play (Day 1 and Day 2). Day 2 data were collected three days after Day 1. We obtained 1,767 observations of affective states and 1,899 observations of on-task/off-task behavior across these two days. Engaged concentration occurred most frequently (73.0% Day 1, 78.6% Day 2), followed by frustration (16.4%, 11.3%), boredom (6.7%, 2.7%), and confusion (3.9%, 1.2%), after removing instances of delight. Despite slightly increased off-task behavior on Day 2 (5.3% vs. 4.3% on Day 1), there appeared to be a general trend toward more engagement and less of the other affective states on Day 2.

2.2 Building the Affect Detectors

Machine learning methods, used to build the affect detectors, proceeded in three phases: feature engineering, supervised classification, and detector validation.

Feature Engineering. The Computer Expression Recognition Toolbox (CERT) [20] is a computer vision tool used to automatically detect the likelihood of 19 different action units (AUs, facial muscle movements; [21]) in any given frame of a video stream. Estimates of head pose and head position information are given by CERT as well. CERT has been tested with databases of both posed facial expressions and spontaneous facial expressions, achieving accuracy of 90.1% and 79.9%, respectively, when discriminating between instances of the AU present vs. absent [20].

We used FACET SDK (no longer available as standalone software), a commercialized version of the CERT computer vision software, for facial feature extraction. Features were created by computing the median, and standard deviation for the frame-level likelihood values of AUs and head position obtained from FACET in a window of time leading up to each observation. For example, we created two features from the AU4 channel (brow lower) by taking the median, and standard deviation of AU4 likelihoods within a six second window leading up to an affect observation. Window sizes of 3, 6, 9, and 12 seconds were explored. We also used two features (median and standard deviation) computed from gross body movement in the videos. Body movement was calculated as the proportion of pixels in each video frame that differed from a continuously updated estimate of the background image generated from the four previous frames using a motion silhouette algorithm [22].

Poor lighting, extreme head pose or position, occlusions from hand-to-face gestures, and rapid movements can all cause face registration errors; these issues were not uncommon due to the game-like nature of the software and the active behaviors of the students in this study. A third (34%) of the instances were discarded because FACET was not able to register the face for at least one second (13 frames) during an observation (a common problem in face-based affect detection), and thus the presence of AUs could not be estimated.

Tolerance analysis was used to eliminate features with high multicollinearity (variance inflation factor > 5) [23]. RELIEF-F feature selection [24] was used to obtain a sparser, more diagnostic set of features for classification. Feature selection was performed using 10 iterations of leave-33%-of-students-out nested cross-validation within the training data only.

Supervised Learning. A two-class approach was used for each affective state, where that affective state was discriminated from all others. For example, engaged was discriminated from all frustrated, bored, delighted, and confused instances combined (referred to as "all other"). Behaviors were similarly grouped into two classes: off-task and both on-task and on-task conversation. Weka, a popular machine learning tool, was used to train supervised classifiers [25]. Bayes net, updateable naïve Bayes, classification via clustering, and logistic regression classifiers were chosen based on our prior results [17]. Synthetic oversampling (with SMOTE; [26]) was used to equalize class sizes on the training data only. The distributions in the testing data were not changed, to preserve the validity of the results.

Model Validation to Test Generalization. Testing the generalization of models across days was performed with a nested cross validation approach to ensure generalization to new students and new days. First, data from one day were chosen as training data. Then, 67% of students were randomly selected from that day and their data

were used to build a model using a repeated random sub-sampling approach with 150 iterations for model selection and evaluation. This model was tested using data from the remaining 33% of students in the same day (same-day generalization: e.g., train on Day 1, test on Day 1) or on the opposite day (cross-day generalization: e.g., train on Day 1, test Day 2). Student-level independence was thus ensured for both testing on the same day and on a different day. Fig. 2 illustrates the validation process.

The process of randomly selecting students for training and testing was repeated 150 times for each model (train-test: Day 1-Day 1; Day 1-Day 2; Day 2-Day 1; Day 2-Day 2) and the results were averaged across iterations.

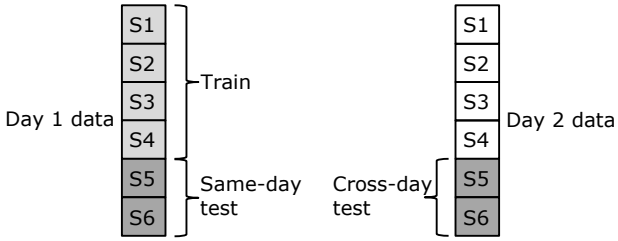

Fig. 2. Example of same-day and cross-day testing with student-level independence.

3 Results and Discussion

Results are organized with respect to the research questions listed in the Introduction.

Research Question 1: Cross-Day Generalization. To explore the performance of models trained on one day and tested on another, we compared performance against the model trained and tested on data from the same day. The same-day results were averaged across both of the same-day models (e.g., train on Day 1, test on Day 1; train on Day 2, test on Day 2). Likewise, the cross-day results were obtained by averaging both cross-day models (train Day 1, test Day 2; train Day 2, test Day 1) for each affective state. Table 1 contains the results for each affective state. Previous work with both days of data combined is also provided as a reference point [17].

Table 1. Results of cross-day compared to same-day detection.

Classification	Cross-day AUC	Same-day AUC	Change (Cross-day - Same-day)	Combined-days AUC	Change (Cross-day - Combined-days)
Boredom	.577	.574	0.23%	.610	-3.34%
Confusion	.639	.665	-2.61%	.649	-1.02%
Engagement	.662	.679	-1.72%	.679	-1.70%
Frustration	.631	.643	-1.21%	.631	0.02%
Off Task	.781	.774	0.72%	.816	-3.48%
Mean	**.658**	**.667**	**-0.92%**	**.677**	**-1.90%**

Note. Cross-day change is percentage of change in AUC, which is bounded on [0, 1].

The key result is that the cross-day models performed with similar accuracy (average AUC = .658) to same-day models (average AUC = .667). Though there was a slight decrease overall in classification accuracy (< 1%), no detector suffered notably. The largest drop (2.61% drop in AUC) occurred for confusion.

Compared to previous work with combined-days models, same-day models had 0.99% lower performance and cross-day models had 1.90% lower performance. Decreased performance may be attributable to the fact that the combined-days models have the advantage of twice as much training data. With more data, cross-day model performance might improve and approach the combined-days models' performance.

Research Question 2: Comparison of Model Parameters. In addition to number of instances, we compared other data and model parameters to illustrate potential differences between the individual day models. Window size (3, 6, 9, or 12 seconds), feature selection (yes or no), and classifier were compared. Table 2 shows the differences in the best-performing Day 1 and Day 2 models.

Table 2. Differences between Day 1 and Day 2 models.

Classification	Window Size		Feature Selection		Classifier	
	Day 1	Day 2	Day 1	Day 2	Day 1	Day 2
Boredom	12 sec	12 sec	Yes	Yes	CVC	CVC
Confusion	12 sec	12 sec	Yes	Yes	**BN**	**UNB**
Engagement	9 sec	9 sec	No	No	BN	BN
Frustration	**9 sec**	**6 sec**	No	No	BN	BN
Off Task	12 sec	12 sec	Yes	Yes	UNB	UNB

Note. Bold indicates parameters that differed between models. Classifiers listed were Classification Via Regression (CVC), Bayes Net (BN), and Updateable Naïve Bayes (UNB).

Of the 15 parameters for each individual day model, only two differed (denoted in bold in Table 2). The close similarities between the data and model parameters of the best-performing detectors suggest that the particular day does not make a notable difference in parameters, and thus the chosen parameters are not the result of overfitting models to day-specific attributes.

Research Question 3: Comparison of Feature Selection Rankings. We examined correlations of feature rankings between the Day 1 models and Day 2 models to determine how similar the set of selected features might be on different days. Feature rankings were recorded during each iteration of RELIEF-F feature selection [24] in each model of the three that used feature selection (boredom, confusion, and off-task). RELIEF-F uses L1 distance to rank a feature based on within-class distance vs. between-class distance, so rankings are subject to substantial variation unless data are tightly clustered within that feature. Thus, correlations were expected to be modest provide a measure of how similar the data clustering is within features between days.

The correlation between Day 1 and Day 2 feature rankings was lower for the affective states than for off-task behavior. Confusion was correlated least (r = .248), fol-

lowed by boredom (r = .270). However, feature rankings for both affective state models were correlated between days in the positive direction, demonstrating that at least some of the same clustering present in the features was detected by RELIEF-F in both days. Off-task behavior was correlated more highly and in the expected direction between days (r = .402).

We also examined the correlations between each feature rankings from each individual day and the combined-days models. Both affective states' feature rankings correlated with the combined-days models (r = .595 for boredom, r = .654 for confusion). Off-task feature rankings were also correlated (r = .613). The comparatively large magnitude of these correlations was not surprising given that half of the data in the combined-days models matches the data in each individual day model.

4 General Discussion

Creating affect detectors that generalize across time is important if intelligent education environments are going to be useful for learning across multiple sessions. We tested if affect detection models built using facial features and machine learning techniques could be generalized across days with reasonable accuracy for several affective states (boredom, confusion, engagement, and frustration) that are important to learning, as well as off-task behavior.

Main Findings. We built student-independent cross-day affect detection models and compared them to same-day and combined-days models. Cross-day detection was successful using training data from one day and testing data from another day with AUC = .658. Compared to performance of the same-day models (AUC = .667) and combined-days models (AUC = .677) for these affective states, the cross-day generalization models show similar performance, though marginally lower for some affective states. We also found some similarity (and difference) in the feature selection rankings between days of data (average r = .249 for affective states, r = .402 for off-task behavior). Cross-day models could still successfully classify data from a different day at levels well above chance despite differences in feature selection rankings.

Limitations and Future Work. This study is not without its limitations, particularly with regards to the breadth of data used. Though we collected data from multiple class periods and two days, all data were collected in the same computer-enabled classroom and learning environment. Lighting conditions and the students who participated varied somewhat between days, but more variation (such as could be obtained from different learning environments at multiple schools) might make the task even more difficult and produce new insights on generalization of face-based affect detection to new contexts. Similarly, the amount of time represented in this study (two different days) is enough to explore the first steps of cross-day generalization, but not enough to explore larger differences such as cross-seasonal generalization (i.e., train models in fall test in spring). Future work will address these issues by expanding data collection to encompass more geographical areas and extended periods of time.

Concluding Remarks. We took the first steps in studying the temporal generalizability of face-based affect detectors in classroom contexts. With affect detectors that

generalize well across time and work in noisy school environments, affect-sensitive computerized education environments can respond to the affective needs of students with confidence that detections are not simply the result of factors specific to a particular day. The next step is to study temporal generalization across extended time frames, such as months or years, so that seasonal differences can be better understood.

Acknowledgment. This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958) and the Bill & Melinda Gates Foundation. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

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. Schutz, P., Pekrun, R. eds: Emotion in Education. Academic Press, San Diego, CA (2007).
3. Bosch, N., D'Mello, S., Mills, C.: What Emotions Do Novices Experience during Their First Computer Programming Learning Session? In: Lane, H.C., Yacef, K., Mostow, J., and Pavlik, P. (eds.) Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED 2013). pp. 11–20. Springer-Verlag: Berlin Heidelberg (2013).
4. Pardos, Z.A., Baker, R.S.J.D., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M.: Affective States and State Tests: Investigating How Affect Throughout the School Year Predicts End of Year Learning Outcomes. Proceedings of the Third International Conference on Learning Analytics and Knowledge. pp. 117–124. ACM, New York, NY, USA (2013).
5. Fiedler, K., Beier, S.: Affect and cognitive processes in educational contexts. International handbook of emotions in education. 36–56 (2014).
6. D'Mello, S., Blanchard, N., Baker, R., Ocumpaugh, J., Brawner, K.: I feel your pain: A selective review of affect-sensitive instructional strategies. In: Sottilare, R., Graesser, A., Hu, X., and Goldberg, B. (eds.) Design Recommendations for Intelligent Tutoring Systems - Volume 2: Instructional Management. pp. 35–48 (2014).
7. D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Perkins, L., Graesser, A.: A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In: Aleven, V., Kay, J., and Mostow, J. (eds.) Intelligent Tutoring Systems. pp. 245–254. Springer, Berlin Heidelberg (2010).
8. Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K., Christopherson, R.: Emotion sensors go to school. AIED. pp. 17–24 (2009).
9. AlZoubi, O., Hussain, M.S., D'Mello, S., Calvo, R.A.: Affective modeling from multichannel physiology: analysis of day differences. Proceedings of the 4th International Conference on Affective Computing an Intelligent Interaction (ACII 2011). Springer-Verlag: Berlin Heidelberg (2011).
10. D'Mello, S., Kory, J.: Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. Proceedings of the 14th ACM international conference on Multimodal interaction. pp. 31–38. ACM, New York, NY, USA (2012).

11. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence. 23, 1175–1191 (2001).
12. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence. 31, 39–58 (2009).
13. Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.C.: Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. Proceedings of the 6th International Conference on Educational Data Mining (2013).
14. Kapoor, A., Burleson, W., Picard, R.W.: Automatic prediction of frustration. International Journal of Human-Computer Studies. 65, 724–736 (2007).
15. Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing. 5, 86–98 (2014).
16. Bosch, N., Chen, Y., D'Mello, S.: It's written on your face: detecting affective states from facial expressions while learning computer programming. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., and Panourgia, K. (eds.) Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014). pp. 39–44. Switzerland: Springer International Publishing (2014).
17. Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V.J., Ventura, M., Wang, L., Zhao, W.: Automatic Detection of Learning-Centered Affective States in the Wild. Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015). ACM, New York, NY, USA (In Press).
18. Shute, V.J., Ventura, M., Kim, Y.J.: Assessment and learning of qualitative physics in Newton's Playground. The Journal of Educational Research. 106, 423–430 (2013).
19. Ocumpaugh, J., Baker, R., Rodrigo, M.M.T.: Baker-Rodrigo observation method protocol (BROMP) 1.0. Training manual version 1.0. Technical Report. New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences (2012).
20. Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M.: The computer expression recognition toolbox (CERT). 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011). pp. 298–305 (2011).
21. Ekman, P., Friesen, W.V.: Facial action coding system. Consulting Psychologist Press. Palo Alto, CA (1978).
22. Kory, J., D'Mello, S., Olney, A.: Motion Tracker: Cost-effective, non-intrusive, fully-automated monitoring of bodily movements using motion silhouettes. Presented at the (in review).
23. Allison, P.D.: Multiple regression: A primer. Pine Forge Press (1999).
24. Kononenko, I.: Estimating attributes: Analysis and extensions of RELIEF. In: Bergadano, F. and Raedt, L.D. (eds.) Machine Learning: ECML-94. pp. 171–182. Springer, Berlin Heidelberg (1994).
25. Holmes, G., Donkin, A., Witten, I.H.: WEKA: a machine learning workbench. Proceedings of the Second Australian and New Zealand Conference on Intelligent Information Systems. pp. 357–361 (1994).
26. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 16, 321–357 (2011).