Multimodal Affect Detection in the Wild: Accuracy, Availability, and Generalizability

Nigel Bosch
University of Notre Dame
384 Fitzpatrick Hall,
Notre Dame, IN 46556, USA
pbosch1@nd.edu

ABSTRACT

Affect detection is an important component of computerized learning environments that adapt the interface and materials to students' affect. This paper proposes a plan for developing and testing multimodal affect detectors that generalize across differences in data that are likely to occur in practical applications (e.g., time, demographic variables). Facial features and interaction log features are considered as modalities for affect detection in this scenario, each with their own advantages. Results are presented for completed work evaluating the accuracy of individual modality face- and interaction- based detectors, accuracy and availability of a multimodal combination of these modalities, and initial steps toward generalization of face-based detectors. Additional data collection needed for cross-culture generalization testing is also completed. Challenges and possible solutions for proposed cross-cultural generalization testing of multimodal detectors are also discussed.

Author Keywords

Affect detection; detector generalization; classroom data; in the wild.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Many techniques have been employed to improve learning in computerized education, including methods from the field of human-computer interaction. Affect sensitivity in interfaces is one such technique and has been shown to be useful for improving learning [11]. There are many ways in which an interface can leverage affect sensitivity for increasing learning. For example, bored and disengaged students may be directed to new learning tasks to help them re-engage, while students who are frequently frustrated due to excessively difficult tasks might be presented with material more appropriate to their knowledge levels. Affect detection is a key component of such affect-sensitive systems, because responding to affect requires accurate detection of affect.

The present focus is on multimodal affect detection, so unimodal detection will not be considered in detail. D'Mello and Kory [12] recently provided a review and meta-analysis of 90 multimodal affect detection systems. Their review revealed that two of the most common modalities for affect detection were facial features and audio features, which were used in 76.7% (face) and 82.2% (audio) of surveyed studies. Audiovisual affect detection is clearly the most prominent multimodal fusion approach. However, audio features are limited to interfaces that support spoken interaction.

The current paper proposes taking a somewhat different approach by employing a fusion of facial features and interaction log features, two potentially complementary modalities that have only rarely been considered together. The proposed affect detectors are designed to operate in computer-enabled classrooms, a context that is rife with noisy and missing data, thereby providing additional challenges for multimodal affect detection. Furthermore, the current paper considers challenges and solutions for cross-culture generalization of detectors, which may be mitigated in part by using interaction features. To that end we primarily consider face- and interaction- based affect detection approaches in educational contexts as related work, along with detector generalizability work.

BACKGROUND AND RELATED WORK

Affect detection from facial features and interaction log data has been studied deeply in recent years [1,12]. In this section we briefly review some of the research related to individual face- and interaction- based affect detection, followed by multimodal and generalization work.

Unimodal Affect Detection

Face-based affect detection has been studied in a variety of learning contexts in recent years. For example, Whitehill et al. [18] discriminated four levels of engagement at a fine-grained level in students as they used cognitive skills training software. Grafsgaard et al. [6] used facial features to detect a variety of affect-related variables such as perceived temporal demand (hurriedness), performance,

ICMI '15, November 09-13, 2015, Seattle, WA, USA
© 2015 ACM. ISBN 978-1-4503-3912-4/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2818346.2823316

and frustration. A variety of computer vision methods have been used for affect detection, such as texture, shape, and motion features extracted from faces. A full review of methods is beyond the scope of this paper, but recent review articles cover it in detail (e.g., [12]).

Interaction-based affect detection has been increasingly studied over the last decade [1,14]. Unlike physical sensor-based detectors, which rely upon the physical reactions of the student, these detectors infer affective states from students' interactions with computerized learning systems. Their unobtrusive and cost-efficient nature also makes it feasible to apply interaction-based detectors at scale, leading to a growing field of research regarding discovery with models.

Multimodal Affect Detection

As noted above, there have been a large number of studies that have considered face- and interaction- features independently. However, multimodal combinations of these features are few and far between. These studies are reviewed below, with a special emphasis on affect detection in learning environments – the context of the present work.

In one of the first such studies, Kapoor et al. [9] used multimodal techniques with face- and posture- based features collected from naturalistic data to create a detector of student interest. Facial features such as detected head nods, shakes, and smiles were combined with posture features gathered from a pressure-sensitive chair and interaction log features from the learning environment. They classified interest/disinterest with 87% accuracy (chance being 52%).

More recently, Grafsgaard et al. [7] developed multimodal affect detectors that utilized both facial features and student interaction features (as well as tutor-student dialog) from a dialog-based computerized learning environment designed to teach Java computer programming. They predicted engagement, frustration, and learning (all self-reported by the students) using linear regression with leave-one-student-out cross validation. A combination of facial features and interaction features predicted engagement (R^2 = .112) and frustration (R^2 = .134) more accurately than unimodal features (best R^2 = .048 and -.010 respectively for engagement and frustration). Adding tutor-student dialogue features improved results even further, to R^2 = .282 and R^2 = .520 respectively. This shows the potential for a fusion of features to outperform individual modalities. However, affect detection was done at a course-grained level across an entire learning session, which limits its applicability to drive real-time interventions.

Affect Detector Generalization

The unimodal and multimodal detectors previously discussed have been demonstrated in a variety of contexts and are often cross validated at the student level, demonstrating generalization to new students within the same population. In this section selected works that demonstrate generalizability across time or demographic variables are also discussed.

Prior work has demonstrated cross-corpus generalization of affect detection in some contexts, especially using audio features (e.g., [15,17]). Audio, though useful for affect detection, is not considered in the current research. The generalizability of speech-based affect detection across languages is an important but large goal that is beyond the scope of the current project.

Interaction data from log-files have shown promise for building affect detectors that generalize across time. Pardos et al. [14] 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, p < .01 for year 1; r = -.280, p < .01 for year 2), confusion (r = -.165, p < .01; r = -.089, p < .05), and gaming the system (r = -.431, p < .01; r = -.301, p < .01) negatively correlated with test score in both school years, while engaged concentration (r = .449, p < .01; r = .258, p < .01) 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.

The previously discussed face-based engagement detection work by Whitehill et al. (see above) also investigated generalization of engagement detection across ethnicity in a small separate sample. Their training set consisted of 26 black students from a Historically Black College/University (HBCU) while the generalization testing set consisted of 8 Caucasian-Americans and Asian-Americans. Their best result for cross-ethnicity testing was AUC = .691, versus .729 for training/testing on the same sample. Thus they demonstrated above-chance detection accuracy across ethnicity, with a slight reduction in accuracy. However, their testing set was small, only considered engagement, and tested only one direction of generalization (Caucasian and Asian → Black).

PROPOSED RESEARCH

The literature review revealed a variety of affect detection methods and some work that has been done to investigate issues of generalization for affect detectors. In this doctoral consortium paper we propose a series of studies and contributions intended to develop answers to some remaining key research questions related to multimodal methods for cross-population affect detection in the wild. Each of these questions is partially addressed in the current paper, though much work remains to be done.

First, there are questions of the feasibility of affect detection in the wild, specifically for educational technology in a computer-enabled classroom. Factors such as conversation, movement, posture, and lighting occur in classroom environments, which cannot be as easily controlled as laboratory contexts. These factors have undesirable effects on affect detection (e.g., missing or noisy data for face-based affect detection). In the proposed research we consider facial features and interaction log data (e.g., clicks, response times) as two modalities for affect detection in this context, each with their own advantages.

Second, it is important to explore the potential benefits of multimodal techniques for improving affect detection in a computer-enabled classroom context. Facial features capture expressions of affective states at brief time scales compared to interaction features, while interaction data may be more indicative of affective states with less active facial expressions, such as boredom. Additionally, facial feature detection is highly influenced by factors such as lighting and movement, which may also cause missing data. Interaction log data on the other hand is available as long as students continue to interact with the educational interface, and may thus complement face-based affect detection in situations where facial features cannot be extracted.

Third, there are questions of generalizability for affect detection methods in computer-enabled classroom contexts. Student-independent cross-validation techniques have become commonplace in affect detection research, but there are aspects of generalization not specifically tested by student-level generalization that may be important for applying affect detection techniques in an educational interface. Additional dimensions of generalization include time (applying detectors at a different time than when they were trained), gender, ethnicity, and culture. A multimodal approach to affect detection in the wild should be explicitly tested across these dimensions if such an approach can be applied in a computer-enabled classroom context to students in a new population.

PROPOSED METHOD

We collected a dataset and employed machine learning techniques that have served to answer some (but not all) of the proposed research questions.

Current Dataset

Students played an educational physics game (see [16] for information about the game) in their school's computer-enabled classroom, while videos of their faces were recorded and their interaction (mouse and keyboard) behaviors were logged. The sample consisted of 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. There were about 20 students per class period (four periods) on four different days (55 minutes per period). For affect detection we considered data from the second and third days (roughly two hours total) when students were only playing the game and not being tested.

Students' affective states were "live" annotated during their interactions with Physics Playground using the Baker-Rodrigo Observation Method Protocol (BROMP) field observation system (see [2] for details). These observations served as ground truth labels for affect detection. Affective states of interest were boredom, confusion, delight, engaged concentration, and frustration. This list of states was selected based on previous research [10] and from observing students during the first day of data collection (these data were not used when creating detectors).

Affect Detection Method

We employed supervised machine learning to build affect detectors trained and tested on the dataset collected. Specifically, we have developed methods for face- and interaction- based detectors, multimodal fusions of these modalities, and cross-validation methods for testing some aspects of generalization.

Details of face-based detection methods (classifiers, features, etc.) are available in a recent publication [2]. In brief, facial features were extracted using FACET computer vision software for each frame of video, and aggregated across a window of time leading up to the BROMP-coded affect label. Gross body movement features were extracted from the videos using a previously validated method. A variety of classification models were evaluated by training and testing using student-independent cross-validation.

Interaction-based affect detectors were built using similar methods [8]. Features extracted from the interaction log data consisted of features specific to the learning environment, such as the number of recently occurring events in the game, as well as more general features such as variability of mouse clicks throughout time. These features were also aggregated across time leading up to the affect labels, and student-independent detectors were built.

Generalization across time (to new days and new times of day) was tested for face-based detectors (and will be for interaction-based detectors) by developing cross-validation methods that preserved student independence while training a detector on one day of data and testing it on data from a different day [3]. Similar techniques were employed for testing across time of day, gender, and ethnicity within the sample of students collected (publication currently in review). In the future a similar method will be used to test cross-cultural generalization, by training a detector on data from students in one country and testing on students from a different country.

CURRENT RESULTS

We have completed work toward answering some, but not all of the broad questions posed in the previous section. We collected a dataset in a computer-enabled classroom environment, which was then used to develop interaction- and face- based affect detectors, analyze the benefits of multimodal techniques in this context, and tested generalization of affect detectors across some of the proposed generalization dimensions. In this section we very briefly review the results obtained thus far (Table 1).

Table 1. Summary of Current Results.

	Mean AUC	Availability	Generalization Effect
Face-based	.687	65%	-2.19%
Interaction-based	.608	94%	
Multimodal	.637	98%	

Face-based Affect Detection

Area Under the ROC Curve (AUC) was used to measure classification accuracy, where chance level is .500 and perfect accuracy is 1.00. Boredom (AUC = .610), confusion (.649), delight (.867), engagement (.679), and frustration (.631) were all detected with better than chance-level accuracy [2].

Interaction-based Affect Detection

Interaction-based affect detectors were not as accurate as face-based detectors on average (.608 vs. .687), though still above chance level. Interaction-based boredom detection was slightly more accurate than face-based detection (AUC = .629), while confusion (.588), delight (.679), engagement (.586), and frustration (.559) were not as accurate [8].

Multimodal Fusion

We explored several methods for creating multimodal affect detectors with the face and interaction data, using late fusion. First, we created models using only instances where both face and interaction data were available, to provide a direct comparison of face, interaction, and multimodal techniques. Face-based affect detection was more accurate than interaction-based affect detection on average (AUC = .667 vs. .574), and multimodal fusion performed at least as well (AUC = .671) as face-based detection.

The benefit of multimodal fusion in this context was more apparent in subsequent analyses of detector availability and accuracy. Interaction log data were available in 94% of instances, while facial features were available in just 65% of instances. By training an additional classifier on the outputs of detectors from the individual modalities, we created affect detectors with average AUC = .637, close to face-based detectors (.687), notably better than interaction-based detectors (.608), and with 98% availability.

Detector Generalization

Thus far we have investigated aspects of generalization for face-based affect detection only. Average accuracy was not severely diminished by testing generalization across different dimensions. Accuracy was reduced by 1.89% when cross-validating across time of day, 1.51% across days, and 1.81% across gender, and 3.53% when testing across ethnicities.

REMAINING WORK

Results thus far suggested that multimodal affect detection was possible in a noisy computer-enabled classroom context. Face-based affect detection also appeared to generalize well across time and demographic variables tested. The remaining work will focus on generalizability aspects of multimodal affect detectors. Toward this end we have collected data in another country to enable generalization testing across populations with differing cultures. In this section we identify key cross-culture generalization challenges and propose potential solutions.

Perhaps the most well-studied challenge in cross-culture generalization is the variation in facial expression of affect between cultures. Classic studies of this problem have shown that people of one culture can recognize affect from people of another culture at above chance levels [4]. However, there is also research demonstrating a within-group advantage, i.e. people are better at recognizing the facial expressions of people from their own culture [5]. This may also indicate a within-group advantage for automatic affect detectors built with. We plan to expand face-based feature extraction methods (texture-based and motion-based features) with geometric shape-based features. We will divide features into groups based on categories, such as representation (e.g., shape-based) and area of the face (e.g., mouth), building separate models for these categories to determine which features may be most generalizable across cultures. Additionally, the interaction-based predictions may be less influenced by cultural norms, which could prove to be an added multimodal advantage.

Observing and labeling affecting states poses a related challenge when collecting data from multiple cultures. The observers should be members of the students' culture to exploit in-group recognition advantage. However, this creates a challenge because observers will differ between datasets, and may have slightly different interpretations of the labels (as they should if there are cultural differences). We will examine frequently misclassified cross-cultural instances from each affective state to determine if there are salient differences that could not only be adapted for, but also shed light on cross-cultural expressions of affect.

Another issue related to the affect labels is the fact that prior proportions of affect could be quite different between datasets collected in different cultures. Differing data distributions between training and testing sets can cause problems for machine learning methods. For example, detectors built on one dataset might be too biased toward or away from some affective states when applied to another dataset. We propose applying methods from transfer learning [13], designed to solve distribution-related issues, in order to adapt models from one culture to the other.

Future improvements to affect detection will also include information about affect developing through time. For example, long short-term memory (LSTM) recurrent neural networks have shown improved affect classification performance compared to traditional methods [19]. LSTM is capable of capturing patterns that evolve over long periods of time, which is relevant to affect detection.

CONCLUSION

Applying affect detection methods to new data collected in a different country will serve as a thorough test of multimodal face- and interaction- based affect detectors. If detectors trained on data collected in year 2013 in the Southeastern U.S. successfully detect affect at above-chance levels in data collected two or more years later in a different country, these detectors can indeed by applied to computer-enabled classroom contexts with confidence that detectors are robust to potential sources of systematic bias. The work already completed as well as the results of cross-cultural generalization testing will serve as the central theme of a PhD dissertation for the author.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Sidney D'Mello, for his guidance on this research. This research was supported by the National Science Foundation (NSF) (DRL 1235958) and the Bill & Melinda Gates Foundation. Any opinions, findings and conclusions, or recommendations expressed in this paper do not necessarily reflect the views of the NSF or the Bill & Melinda Gates Foundation.

REFERENCES

1. Ryan Baker and Jaclyn Ocumpaugh. 2015. Interaction-based affect detection in educational software. In *The Oxford Handbook of Affective Computing*, Rafael Calvo, Sidney D'Mello, J. Gratch and A. Kappas (eds.). New York: Oxford University Press, 233–245.
2. Nigel Bosch, Sidney D'Mello, Ryan Baker, et al. 2015. Automatic detection of learning-centered affective states in the wild. *Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015)*, New York, NY: ACM, 379–388.
3. Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, and Valerie J. Shute. 2015. Temporal generalizability of face-based affect detection in noisy classroom environments. *Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015)*, Berlin Heidelberg: Springer-Verlag, 44–53.
4. Paul Ekman. 1994. Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychological bulletin* 115, 2: 268–287.
5. Hillary Anger Elfenbein and Nalini Ambady. 2002. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin* 128, 2: 203.
6. Joseph F. Grafsgaard, Joseph B. Wiggins, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. 2013. Automatically recognizing facial expression: Predicting engagement and frustration. *Proceedings of the 6th International Conference on Educational Data Mining*.
7. Joseph F. Grafsgaard, Joseph B. Wiggins, Alexandria Katarina Vail, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. 2014. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. *Proceedings of the 16th International Conference on Multimodal Interaction*, ACM, 42–49.
8. Shiming Kai, Luc Paquette, Ryan Baker, et al. 2015. Comparison of face-based and interaction-based affect detectors in physics playground. *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*, International Educational Data Mining Society, 77–84.
9. Ashish Kapoor and Rosalind W. Picard. 2005. Multimodal affect recognition in learning environments. *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ACM, 677–682.
10. Sidney D'Mello. 2013. A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology* 105, 4: 1082–1099.
11. Sidney D'Mello, Nathan Blanchard, Ryan Baker, Jaclyn Ocumpaugh, and Keith Brawner. 2014. I feel your pain: A selective review of affect-sensitive instructional strategies. In *Design Recommendations for Intelligent Tutoring Systems - Volume 2: Instructional Management*, Robert Sottilare, Art Graesser, Xiangen Hu and Benjamin Goldberg (eds.). 35–48.
12. Sidney D'Mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys* 47, 3: 43:1–43:36.
13. Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10: 1345–1359.
14. Zachary A. Pardos, Ryan S. J. D. Baker, Maria O. C. Z. San Pedro, Sujith M. Gowda, and Supreeth M. Gowda. 2013. 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*, ACM, 117–124.
15. B. Schuller, B. Vlasenko, F. Eyben, et al. 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing* 1, 2: 119–131.
16. Valerie Shute and Matthew Ventura. 2013. *Measuring and supporting learning in games: Stealth assessment*. The MIT Press, Cambridge, MA.
17. Marie Tahon, Agnes Delaborde, and Laurence Devillers. 2011. Real-life emotion detection from speech in human-robot interaction: Experiments across diverse corpora with child and adult voices. *INTERSPEECH*, 3121–3124.
18. J. Whitehill, Z. Serpell, Yi-Ching Lin, A Foster, and J.R. Movellan. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* 5, 1: 86–98.
19. Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. 2013. LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing* 31, 2: 153–163.