Investigating the Generalizability of Affect Detectors from Facial Expressions

Emily TABANAO[a,b*] **, Ma. Mercedes RODRIGO**[b]
[a]*MSU-Iligan Institute of Technology, Philippines*
[b]*Ateneo de Manila University, Philippines*
*emilytabanao@g.msuiit.edu.ph

Abstract: The goal of this paper was to investigate the generalizability of affect detectors created from facial expressions. Videos of students were captured while they were playing an educational game in a natural computer laboratory setup. Trained observers annotated the learning-centered affective states which served as affect labels for training detectors. Detectors were trained using data from students in the northern part of the Philippines and were tested from data of students from the southern part of the Philippines. We discuss the results, challenges and future work of face-based affect detectors from facial expressions taken in the wild.

Keywords: Affect detection, facial expressions, Physics Playground

Affect detection refers to the automatic recognition of emotions or affective states of a person while interacting with a computer. It is a challenging problem because emotions are constructs that are not directly quantifiable. Despite the many advances in vision computing, automatic facial expression recognition systems still have many challenges in the natural environment. Human observers can easily accommodate for changes in pose, scale, illumination, occlusion, and individual differences, and other sources of variation. But in computer vision, these factors pose many challenges (Cohn, 2011). Though affect is well studied, many of the studies in the literature on emotions in facial expressions are focused on recognizing a limited set of basic emotions as defined by Ekman (happy, sad, anger, disgust, fear, surprise). However, a meta-analysis conducted by (S. K. D'Mello, 2013) found that these may not be the emotions that normally occur in a learning environment context.

In this paper we focus on face-based affect detection. Our training data is taken from a group of students studying in the Northern part of the Philippines and testing data is taken from a group of students in the Southern part of the Philippines. Data were gathered two weeks apart. In this study we tried to address the temporal and demographic generalizability of face-based affect detectors by using data gathered in different days and at different locations.

To assess whether face-based affect detectors generalize over time and demographics, the current paper attempts to answer the question: how well can a face-based affect detector perform when applied to data gathered on another week (temporal generalizability) and in a different location (demographic generalizability)?

This paper uses part of a dataset that was used previously on face-based affect detection. However, in this study we expanded it to include data gathered from another location.

We collected data while students were playing Physics Playground (PP) in the computer laboratory. Physics Playground is a two-dimensional computer game that is designed for high school students better understand physics concepts related to Newton's three laws of motion: balance, mass, conservation and transfer of momentum, gravity, and potential and kinetic energy (Shute et al., 2013).

Data were collected from three different schools in the Philippines. In the southern part (Davao City), we have 60 grade 7 students (20 male, 40 female) between ages 12 to 14. In the northern part of the country (Baguio City), we have 62 grade 10 students (32 male, 30 female) between ages 13 to 18 participated in the study. Inexpensive webcams were mounted at the top of each computer monitor. All instructions were given by the experimenters who also served as field observation coders.

Student affect and behavior was collected using the Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP), a method for recording quantitative field observations (Ocumpaugh, Baker, & Rodrigo, 2015). The affective states observed in this study were engaged concentration, confusion, frustration, boredom, happiness, delight, curious, excited, hope and anxious.

For the video data, we used Emotient FACET to extract five categories of information from raw video data input (https://imotions.com/emotient/).

For feature engineering, we synchronized the FACET and affect logs using the timestamps for alignment. Similar to the studies of (Bosch et al., 2015), we created datasets for five different window sizes (3, 6, 9, 12, and 20 seconds). The window ends at the time the affect log was observed. For each window size, we obtained the maximum, median, mean and standard deviation for each of the action units. However, when the FACET log does not register a face for at least one second in the window size, that particular BROMP data will be dropped.

Attributes that exhibited high multicollinearity were eliminated (variance inflation of factor > 5) from both the training and testing datasets. We also applied Relief-F feature selection and created models on different weights (0.3, 0.5 and 0.7) on the different time windows (Kononenko, 1994). We used the open source R statistical software in our analysis (https://www.r-project.org/).

To do supervised learning, we created two classes for each affective state. An affective state was discriminated from all other states. Imbalance in the distribution of the dataset was addressed by applying SMOTE on the training datasets only (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

A total of 10,679 affect observations were collected from both locations. Shown in Table 1 is the distribution of the different affective states. Clearly, we have a case of imbalanced datasets in the two-class modeling.

Table 1

Affect Observations in the two locations

Affect	**Training Dataset (Baguio City)**	**Testing Dataset (Davao City)**	**Total**
Boredom	219	96	315
Engagement	4,449	3,718	8,167
Confusion	413	288	701
Delight	69	96	165
Frustrated	521	323	844
Happy	186	301	487
Total	**5,857**	**4,822**	**10,679**

We have created four different sets of data for modeling, one is the result of removing the features with high multicollinearity and the three other sets were the result of applying the RELIEF-F algorithm with varying weights (weights greater than 0.3, 0.5 and 0.7). As we selected features of higher weights, we noticed a decrease in the number of remaining predictor attributes. However, when we compared the performance of these models in the test dataset, it was observed that the models from the dataset without the RELIEF-F applied to it performed better. Shown in Table 2 are the best performing models when validated using the testing datasets in this experiment.

We note several differences in our results to that of Bosch et.al. First, our detector performances are marginally lower compared to their models. Second, we also see differences in terms of the number of features and window sizes at which the detector performed best. However, we notice that Naïve Bayes tend to deliver better performance in the testing datasets even if its performance in the training datasets are way lower compared to the other classifiers.

In conclusion, we have attempted to address the generalization of face-based affect detectors across time and demographics. The marginally low detector performance confirms the challenging task of building reliable detectors from data from the wild that performs well across time and demographics.

Table 2

Details and result of classification

Affect	AUC	Classifier	No. of Features	Window Size (sec)
Boredom	0.569	Naïve Bayes	42	9
Engagement	0.563	Naïve Bayes	38	6
Confusion	0.539	Naïve Bayes	52	20
Delight	0.599	Naïve Bayes	52	20
Frustrated	0.549	Random Forest	46	12
Happy	0.657	Random Forest	42	9

Acknowledgements

We thank Miggy Andres, Jessica Sugay, Michelle Banawan, Yancy Paredes, the officials at Ateneo de Davao University High School, Nigel Bosch of University of Notre Dame, Drs. Jaclyn Ocumpaugh and Luc Paquette of Teachers College, Columbia University, and Drs. Valerie Shute, Matthew Ventura, and Matthew Small of Florida State University. This study was made possible through a grant from the Philippines' Department of Science and Technology Philippine Council for Industry, Energy and Emerging Technology Research and Development entitled "Stealth assessment of student conscientiousness, cognitive-affective states, and learning using an educational game for Physics."

References

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., & Shute, V. (2015). Temporal generalizability of face-based affect detection in noisy classroom environments. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9112*, 44-53. https://doi.org/10.1007/978-3-319-19773-9_5.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357. https://doi.org/10.1613/jair.953.

Cohn, F. D. la T. and J. F. (2011). Facial Expression Analysis. In *Visual analysis of humans* (pp. 377–409). Springer London. https://doi.org/10.1007/0-387-27257-7_12.

D'Mello, S. K. (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. Retrieved from http://dx.doi.org/10.1037/a0032674.

Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99*(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF, 171–182. https://doi.org/10.1007/3-540-57868-4_57.

Ocumpaugh, J., Baker, R. S. J. D., & Rodrigo, M. M. T. (2015). Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. Retrieved from http://www.columbia.edu/~rsb2162/BROMP.pdf.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in newton's playground. *The Journal of Educational Research*, *106*(6), 423-430.