Video-Based Affect Detection in Noninteractive Learning Environments

Yuxuan Chen
University of Notre Dame
384 Fitzpatrick Hall
Notre Dame, IN 46556, USA
ychen18@nd.edu

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

Sidney D'Mello
University of Notre Dame
384 Fitzpatrick Hall
Notre Dame, IN 46556, USA
sdmello@nd.edu

ABSTRACT

The current paper explores possible solutions to the problem of detecting affective states from facial expressions during text/diagram comprehension, a context devoid of interactive events that can be used to infer affect. These data present an interesting challenge for face-based affect detection because likely locations of affective facial expressions within videos of students' faces are entirely unknown. In the current study, students engaged in a text/diagram comprehension activity after which they self-reported their levels of confusion, frustration, and engagement. Data were chosen from various locations within the videos and texture-based facial features were extracted to build affect detectors. Varying amounts of data were used as well to determine an appropriate window of data to analyze for each affect detector. Detector performance was measured using Area Under the ROC Curve (AUC), where chance level is .5 and perfect classification is 1. Confusion (AUC = .637), engagement (AUC = .554), and frustration (AUC = .609) were detected at above-chance levels. Prospects for improving the method of finding likely positions of affective states are also discussed.

Keywords

Affect detection; facial expression recognition; reading

1. INTRODUCTION

Educational activities like playing educational games [9], interacting with a computerized tutor [4], and comprehending text [13] have been linked to affective experiences that potentially play important roles in the learning process. Thus, automatically detecting and responding to specific affective states can be a useful technique for improving educational software [5]. A wide variety of approaches have been used to detect students' emotions and tailor instruction to their affective needs (see [8] and [5] for reviews). Affect detection is a core challenge that needs to be addressed before affect-sensitive instructional strategies can be devised.

Affect detection during interactions with educational technologies are a widely studied problem. The two most common approaches involve the use of interaction data (e.g., clicks, response times) from log files (called sensor-free detection as reviewed in [1]) and the use of physiological/behavioral sensors, such as webcams, electrodermal sensors, posture sensors, and so on (called sensor-based affect detection as reviewed in [3]). As an illustrative example, Kai et al. [11] compared affect detectors built both interaction-based and video-based affect detectors while students played an educational game called Physics Playground [14]. Their data included affect labels corresponding to specific moments in the learning session (provided by human observers in real-time). The metric of performance was A', a close approximation of Area Under the ROC Curve (AUC), where A' = .5 is chance level and 1 is perfect classification. They were able to detect affective states at levels above chance: confusion (A' = .588 for interaction-based, .622 for face-based), engaged concentration (A' = .586 interaction, .658 face), and frustration (A' = .559 interaction, .632 face).

The aforementioned study highlights two commonalities of affect detection during learning from educational software. First, the software is typically interactive in nature, thereby providing considerable opportunities for external events (e.g., a new problem, submission of a response, system feedback, a hint) to trigger affective states. Information on these events and students' responses to these events provide valuable information to guide affect detection. Second, the data (log-files, videos, etc) used to build affect detectors is accompanied by affect labels corresponding to specific moments in a learning session. This allows label-based segmentation of the data stream and affords pinpointing the sections of the data stream for affect detection (typically windows of 10-20 seconds before the labels; e.g., [9]).

Data in some educational contexts are not well suited to creating affect detectors. For example, in self-paced reading tasks there are not necessarily many key events that are likely to trigger affective responses, unlike many educational activities where there is frequent feedback and interaction. Similarly, not all educational experiences include labeled-data that can be used to pinpoint the temporal location of affective states. For example, students might self-report their affective states *after* reading an entire passage or viewing an online lecture. This raises the additional challenge of how to segment the data stream for affect detection.

The present paper involves affect detection in the context of a noninteractive, but everyday learning task, involving about mechanical reasoning from illustrated texts [7]. Students were presented with a complete text passage with an associated diagram for two minutes of study. Students self-reported their affective states after each a two minute study session, rather than any specific moment in the session. This data raised many challenges. First, interaction data was non-existent as there are no page turns or other navigation features that can be used to gain information about student behaviors. Due to the lack of interaction information, we use facial features extracted from videos of students' faces to detect affective states as they processed the

text/diagram. Second, without predictable events in the task that could trigger affective states and without affect labels during the study session, the position within a video where facial expressions of affective states are likely to occur is unknown. Rather than analyzing the entire video, knowing the location of affective states is important because the duration of affective experiences can be short and the facial expressions associated with affective states can be even shorter [2,6]. To address this problem we explore affect detection using different data window sizes and window positions within face videos to determine where displays of affect tend to occur and how long they last.

We also studied the role of learning goals on affect detection performance. Specifically, students studied the illustrated texts under two different instructional conditions. The first was to simply learn about a mechanical device (general instructions). This was followed by a focused goal that either directed students to review key components of the device or to pinpoint a particular problem with the device (specific instructions). We expect differences in affect detection results between the two types of instructions because they are expected to engender different levels of processing. Thus, we also build separate detectors for the two types of instructional goals to determine if there was a notable difference in detection performance.

Our main approach consisted of applying machine learning techniques to build detectors of confusion, engagement, and frustration with features extracted from facial videos using CERT [12], which is a well validated computer vision tool for extracting texture-based facial features. Detection results with different window sizes and positions show both the potential and the difficulty of detecting affective states from face videos when little is known about when displays of affect might likely occur. The data in this study come from studying instructional texts with illustration, and as such is representative of potential real-world education scenarios. Thus, determining how to detect affective states in this context is important for improving computerized education systems.

2. METHOD

Data Collection. Data were collected from 88 college students at a large public university in the mid-South. These students were asked to study illustrated texts about four everyday devices: an electric bell, a toaster, a car temperature gauge, and a cylinder lock. The illustrated texts were taken from Macaulay's book, *The Way Things Work* (1988). Each of the general and specific study instructions lasted for two minutes. Videos of the students' faces were recorded with webcams cameras mounted on the computer monitors. Upon completion of each two minute study session, students rated their levels of engagement, confusion, and frustration on scales of 1 (very little) to 6 (very much). Students studied all four devices with device order counterbalanced across students, thereby resulting in 704 videos (88 students × 4 devices × 2 study goals per device).

Three students' videos were discarded due to recording errors, which resulted in 680 usable videos. These videos were then analyzed using CERT, which computed the likelihoods of occurrence for facial action units (AUs) in every video frame. Large outliers in AU likelihoods were found in the last two seconds of most videos, which are probably the result of students posture shifts in response to the end of the session. The last 2 seconds were removed to compensate for these anomalies, so each video was then exactly 1 minute 58 seconds long.

Feature Engineering. CERT was able to detect 19 different AUs as well as unilateral (one side of the face only) AUs, head orientation, and nose position. From the CERT data, windows of eight different sizes (2, 3, 6, 9, 12, 15, 20, and 30 seconds) were generated. For each size, windows were drawn from the beginning, the middle, and the end of each video. If the window came from the beginning or the end of the video, the margin from the beginning or the end was equal to the length of the window. Figure 1 illustrates examples of windows created in this manner.

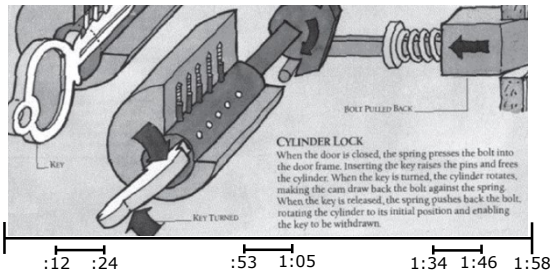

Figure 1. Positions of 12-second windows during the task.

The AU data of the windows were standardized within each student. This was followed by feature generation, in which the median, maximum, and standard deviation of the frame-level AU likelihoods were computed within each window and used as features. Some windows had less than one second of valid data, largely because the camera could not capture the student's face when they moved too much, leaned outside the camera's field of view, or when the face was occluded due to gestures. These windows were removed from the dataset, as we assumed that an affective facial expression would usually be longer than one second. Features exhibiting high multicollinearity (variance inflation factor > 5) were removed.

Supervised Classification. The features obtained above were used to construct classification models using the Waikato Environment for Knowledge Analysis (WEKA), a machine learning tool. Fifteen different classifiers were tested in order to identify the best performing classifier.

The classification task comprised of binary high vs. low affect ratings for confusion, frustration, and boredom. The medians of the engagement, confusion, and frustration ratings on the 1-6 scale were 4, 2, and 1, respectively. We used a median split to discretize the affect ratings into "low" and "high", discarding the median instances except in the case of frustration where the median was 1. For frustration 1 was used as the "low" label.

For model validation, leave several out student-level cross-validation was applied. The training data were randomly chosen from two thirds of the students. RELIEF-F feature ranking was used to select the most diagnostic features on the training data only. The data of the remaining students were used to test the generalizability of the classifiers. Each model was trained and tested for 150 iterations with random students selected for training and testing each iteration to reduce random sampling error. Fifteen different classifiers were applied to help determine which among the eight window sizes tended to work best.

3. RESULTS

The best classification models that merged videos recorded during both general and specific study instructions are listed in Table 1. The AUCs for confusion and frustration were well above chance, whereas the AUC for engagement was only slightly higher than chance level.

It should be noted that there were fewer than 680 instances (the total number of usable videos) for these classification models. This was largely because instances that captured less than a

Table 1. Overview of results when general and specific instructional videos were combined.

Affective State	Classifier	AUC	Accuracy	No. Instances	No. Features	Window Size
Confusion	Updateable Naïve Bayes	0.637	62%	352	65	9 seconds
Engagement	AdaBoostM1	0.554	55%	403	49	20 seconds
Frustration	AdaBoostM1	0.609	64%	356	39	6 seconds

second of data were eliminated and the median splits that were performed to ascertain "low" and "high" values resulted in the loss of instances with affect ratings at the median.

General vs. Specific Study Instructions. The best AUCs for each video type are in Table 2. We note that for engagement, AUCs for individual general-instruction and specific-instruction models were higher than when the videos were combined. However, for confusion and frustration, it seems that the best AUCs are mostly equivalent across both individual videos and combined videos.

Table 2. Comparison of classification performance (AUC) for models using only explanation, only review, or both types of data.

Affective State	General	Specific	Both
Confusion	0.664	0.606	0.637
Engagement	0.610	0.580	0.554
Frustration	0.600	0.620	0.609

Window Position. The best AUCs (for combined models) with respect to the three window positions (i.e., beginning, middle, and end) are shown in Figure 2. Clear patterns stand out for confusion and frustration. The windows taken from the beginning of the videos seem to be more effective for confusion than those taken from the middle or the end of the videos, whereas the windows drawn from the end of the videos may best capture frustration. There is no clear pattern for engagement.

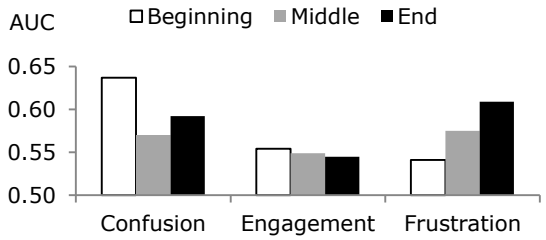

Figure 2. AUC of models using data from different positions within videos.

Window Size. Figure 3 shows the best AUCs as a function of window size for the combined models. The window position was held constant as the best window position for each affective state as noted in Figure 2. Confusion and frustration again show interesting patterns. AUC peaks at a certain window size where classification is much more successful than the surrounding window sizes. The peaks for the AUCs of confusion and frustration both occur when the window size is relatively small (9 seconds for frustration and 6 seconds for confusion). Conversely the window size seems to have no notable relationship with AUC for engagement.

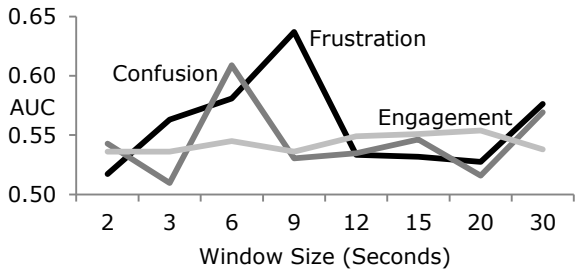

Figure 3. AUC of models as window size varies.

4. DISCUSSION

The novelty of the contributions in this paper stems from the differences between data in this study and previous affect detection work. Facial expressions of affect are often related to events in an interface (e.g., feedback, new problems), but the present study tracked affect in a noninteractive study activity – comprehension from illustrated texts. Affect labels used for detection in this study were given as retrospective judgments covering an entire 2-minute study period, so they do not provide any information about the appropriate position in the video to search for facial expressions. Thus the position of potential facial expressions in the face videos is entirely unknown. Unlike related studies with affect labels not tied to specific moments in a learning session (e.g., [10]), the current research used a subset of data from the session rather than considering all data in the session. This approach was chosen to better capture the brief nature of affective facial expressions. In the remainder of the section we discuss our main findings and highlight limitations and avenues for future work.

Main Findings. The results above show that confusion and frustration ratings of the students can be detected with greater accuracy than the engagement rating, but that detection was successful above chance for all three affective states despite the difficulty of identifying a brief affective facial expression within the videos. However, if we split the general-instruction videos from the specific-instruction videos, the engagement rating may be better modeled, especially for the general videos. For confusion, a 9-second window at the beginning of the video works best for classification; and for frustration, a 6-second window at the end of the video has the best performance. There were no clear patterns with respect to window position or window size for engagement.

The results suggest that when given a video with the occurrences of different affects unknown, affect ratings for confusion, frustration, and potentially engagement can still be well modeled. Smaller window sizes such as 6 or 9 seconds can be a good start to find such best models for confusion and frustration, which parallels the results in previous research [2]. Also, clips taken from the beginning of the video may yield good models for confusion, and those taken from the end of the video may work well for frustration. This seems to suggest that students' facial expressions at the beginning of the 2-minute study session can

potentially indicate how confused they think they are in the end, and that their facial features at the end of a session may provide evidence as to how frustrated they rate themselves to be. It seems that when students confront a specific task, their first impression or assessment of the difficulties and intricacies of the task can last until the end of the task. As they try to understand new concepts or to tackle problems, they experience the details of the task that they might not have known before. This may be why at the end of the task, whether they completely absorb the concepts or solve the problems or not, they may still feel frustrated and challenged and such emotions can be detected by analyzing facial expressions.

The reasons why engagement detection is a difficult task in this context may be due to differences in facial expressions of engagement between the general and specific study periods. It is possible that people's definitions of engagement may be linked to the particular tasks they are working on. When videos associated with the general and specific study periods were separated, engagement can be modeled more successfully compared to when the videos were combined. This is intuitively plausible as the general and specific study sessions may be essentially different tasks, the former requiring students to intake new concepts and the latter challenging students to focus on specific aspects based on the concepts they have learned. So students may have diverse judgments on how engaged they are based on the type of tasks they have in hand, and that may explain why engagement seems to be better modeled when general and specific videos were analyzed independently.

Limitations and Future Work. The results were promising, but there are a few limitations to this research. First, the number of videos was rather low and around 30% of the windows had to be discarded due to difficulties in registering the face (mostly due to hand-over face gestures). Also, the videos for the research were only 2 minutes long. If the window size is 30 seconds, trimming off the beginning and end 30 seconds from a video indicates that we only have one minute left for the video and the segments taken from this video can be overlapping, which is not ideal. Further research with a greater number of longer videos would allow a more thorough search of window positions and window sizes.

In addition, we adopted a rather simplistic approach of searching the start, middle, and end of each video to identify diagnostic affect expressions. In future work, we will delve more deeply into the data we already have using different methods of searching for positions in the videos where affective facial expressions occur. For example, we may utilize the 9-second window size to perform a random sampling across all videos, taking segments from random positions within each video to offer more insight into how facial expressions can be leveraged for affect detection. It may also be possible to develop techniques for finding the optimal window position on a per-video basis, for example by searching for peaks or valleys in calculated features, and using windows of data specific to each video.

Concluding Remarks. In summary, this paper introduces a potential method to detect students' affective states in non-interactive instructional contexts when the locations and durations of affective facial expressions are unknown. Much work remains to be done to improve these techniques, but our results show that detecting affective states with these challenging data is certainly possible and highlight the importance of correctly identifying the position and length of windows of data within each video.

5. ACKNOWLEDGEMENTS

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 are those of the authors and do not necessarily reflect the views of the NSF or the Bill & Melinda Gates Foundation.

6. REFERENCES

1. Baker, R. and Ocumpaugh, J. Interaction-Based Affect Detection in Educational Software. In R. Calvo, S. D'Mello, J. Gratch and A. Kappas, eds., *The Oxford Handbook of Affective Computing*. New York: Oxford University Press, 2015, 233–245.
2. Bosch, N., D'Mello, S., Baker, R., et al. Automatic Detection of Learning-Centered Affective States in the Wild. *Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015)*, ACM (In Press).
3. Calvo, R.A. and D'Mello, S. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing 1*, 1 (2010), 18–37.
4. Craig, S., Graesser, A., Sullins, J., and Gholson, B. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media 29*, 3 (2004), 241–250.
5. D'Mello, S., Blanchard, N., Baker, R., Ocumpaugh, J., and Brawner, K. I feel your pain: A selective review of affect-sensitive instructional strategies. In R. Sottilare, A. Graesser, X. Hu and B. Goldberg, eds., *Design Recommendations for Intelligent Tutoring Systems - Volume 2: Instructional Management*. 2014, 35–48.
6. D'Mello, S. and Graesser, A. The half-life of cognitive-affective states during complex learning. *Cognition & Emotion 25*, 7 (2011), 1299–1308.
7. D'Mello, S. and Graesser, A. Inducing and tracking confusion and cognitive disequilibrium with breakdown scenarios. *Acta Psychologica*, 151 (2014), 106–116.
8. D'Mello, S. and 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*, ACM (2012), 31–38.
9. Graesser, A., Chipman, P., Leeming, F., and Biedenbach, S. Deep learning and emotion in serious games. *Serious games: Mechanisms and effects*, (2009), 83–102.
10. Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., and Lester, J.C. Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. *Proceedings of the 6th International Conference on Educational Data Mining*, (2013).
11. Kai, S., Paquette, L., Baker, R., et al. Comparison of Face-based and Interaction-based Affect Detectors in Physics Playground. (in review).
12. Littlewort, G., Whitehill, J., Wu, T., et al. The computer expression recognition toolbox (CERT). *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, (2011), 298–305.
13. Mills, C., Bosch, N., Graesser, A., and D'Mello, S. To Quit or Not to Quit: Predicting Future Behavioral Disengagement from Reading Patterns. *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)*, Switzerland: Springer International Publishing (2014), 19–28.
14. Shute, V.J., Ventura, M., and Kim, Y.J. Assessment and learning of qualitative physics in Newton's Playground. *The Journal of Educational Research 106*, 6 (2013), 423–430.