Semi-Supervised Learning to Perceive Children's Affective States in a Tablet Tutor

Mansi Agarwal
Delhi Technological University,
New Delhi, India
mansiagarwal_bt2k16@dtu.ac.in

Jack Mostow
Carnegie Mellon University,
Pittsburgh, PA, USA
mostow@cs.cmu.edu

Abstract

Like good human tutors, intelligent tutoring systems should detect and respond to students' affective states. However, accuracy in detecting affective states automatically has been limited by the time and expense of manually labeling training data for supervised learning. To combat this limitation, we use semi-supervised learning to train an affective state detector on a sparsely labeled, culturally novel, authentic data set in the form of screen capture videos from a Swahili literacy and numeracy tablet tutor in Tanzania that shows the face of the child using it. We achieved 88% leave-1-child-out cross-validated accuracy in distinguishing pleasant, unpleasant, and neutral affective states, compared to only 61% for the best supervised learning method we tested. This work contributes toward using automated affect detection both off-line to improve the design of intelligent tutors, and at runtime to respond to student affect based on input from a user-facing tablet camera or webcam.

1 Introduction and Relation to Prior Work

The field of affective computing seeks to narrow the communicative gap between the naturally emotional human and the emotionally challenged computer by developing computational systems that recognize and respond to the affective states (*i.e.*, emotions) of the user (Picard 2000). In particular, considerable work has investigated the automated estimation of affective states from facial expressions (*e.g.* (Faria et al. 2017)) and other visual cues (*e.g.* (Bidwell and Fuchs 2011)). Emotional expressions are socially reactive, so users may try to mask certain unpleasant emotions (McDaniel et al. 2007).

Much of the research on affective computing has focused on making intelligent tutoring systems react to students' emotions (*e.g.*, (Woolf et al. 2009), (D'Mello and Graesser 2012), (Craig et al. 2004)). This paper likewise presents work on automated detection of children's affective states in RoboTutor (Mostow 2019), a tablet app that (like the other 4 Finalists in the Global Learning XPRIZE (XPRIZE 2015)) achieved dramatically higher learning gains in basic literacy and numeracy than a delayed-treatment group in XPRIZE's 15-month-long independent controlled study of 2700 children in 170 villages in Tanzania. RoboTutor's thousands of activities teach children who have little or no prior schooling. Our eventual goal for automated affect detection is to help RoboTutor increase children's engagement and learning gains. The work reported here is novel in several respects.

Novel population: Work on affect detection in intelligent tutors has typically focused on American high school and college students. In contrast, the work reported here is based on data from children ages 6-12 in Tanzania. This data is novel in three respects. First, few databases of facial expressions include children's faces (Egger et al. 2011) (Nojavanasghari et al. 2016). Second, even fewer include Africans (Du, Tao, and Martinez 2014). Finally, emotional expression varies from culture to culture (Matsumoto 1991), so affect detectors trained on an American population might not work for an East African population.

Authentic context: Foundational research on emotion detection has mainly focused on six "basic" emotions (happiness, sadness, surprise, disgust, anger, and fear) (Craig et al. 2008). Typically these emotions are represented by deliberate facial expressions (Kaulard et al. 2012) or elicited by experimental stimuli (Valstar and Pantic 2010). In contrast, the affective states relevant to intelligent tutors are students' normal reactions to them, namely boredom, confusion, delight, frustration, surprise, and neutral or "flow" (D'Mello, Picard, and Graesser 2007). Most facial expression data is recorded in well-controlled laboratory conditions. In contrast, this paper is based on data from authentic contexts of children using RoboTutor.

Camera-only: Even data on authentic affective states in intelligent tutors are typically collected in heavily instrumented laboratory conditions using an expensive array of devices such as pressure sensors (D'Mello and Graesser 2009) and EEG headsets (Petrantonakis and Hadjileontiadis 2009), as well as video from cameras external to the tutor itself. These input signals are informative for research but not practical outside the lab.

In contrast, we use only video input recorded by Google Pixel C Android tablets running RoboTutor in the field, both indoors and outdoors. Limiting its temporal and spatial resolution served to avoid filling up tablet memory or swamping the WiFi bandwidth required to send it to our lab for analysis. This data is therefore characterized by limited resolution, variable indoor and outdoor illumination, and occlusion by children's friends and their own hands.

Multi-channel: Facial expressions are an important visual channel to convey emotions, but by no means the only one. We also use other visual features known to reflect affective states: head proximity (Stanley 2013), head orientation (Hess, Adams, and Kleck 2007), blink rate (Haq and Hasan 2016), pupil size (Partala and Surakka 2003), and eye gaze (Bidwell and Fuchs 2011).

Semi-supervised: Systems that rely on supervised machine learning require large amounts of labeled training data (*e.g.* (Michel and El Kaliouby 2003), (Reddy et al. 2018)). Labeling affective states by hand is costly and time-consuming. Therefore we employ a semi-supervised approach for training an affective state detector on a sparsely labeled dataset (Chapelle, Scholkopf, and Zien 2009). That is, we train a classifier on the manually labeled instances, ”pseudo-label” a subset of the unlabeled data using the trained classifier, retrain it on the expanded set of labeled data, repeat, and iterate.

In summary, this paper reports progress on using AI (specifically computer vision) to “improve teaching and evaluation.” From recorded screen video of children in Tanzania using RoboTutor on an Android tablet, we infer their affective states. Our longer term goal is to use this information to redesign RoboTutor off-line, and even to inform its responses in real-time. The rest of the paper is organized as follows: Section 2 describes our data set. Section 3 specifies our methodology for training the affective state detector. Section 4 reports our results. Finally, Section 5 summarizes contributions, limitations, and future work.

2 Data Set

The data for the present study come from 229 screen capture videos of approximately 30 children using RoboTutor on two tablets in Tanzania between $6/22/2016$ and $7/17/2017$. Each video typically shows one session lasting $20-30$ minutes (until the next child's turn). The videos were recorded by a free app called AZ Screen Recorder (PlayStore 2018), which displayed the front-facing camera input in a small window and included it in the screen video it recorded, as shown in Figure 1. To limit storage consumption, we configured AZ Screen Recorder to record at a temporal resolution of 48 frames per second and a spatial resolution of 1024 x 720 pixels, of which the camera window took 192 x 148 pixels.

The entire 100+ hours of video was far too large to label manually, so we selected approximately 345 short clips to label. As in previous work (Westlund, D'Mello, and Olney 2015), we excluded the first and last minute of a video so as to avoid artifacts at the start and end of each session. Based on watching a few short clips, we determined that 10-second clips were long enough to label yet short enough to be dominated by a single affective state.

We used two types of sampling to find clips to label. Based on previous research (D'Mello and Graesser 2010), we expected *neutral* to be by far the most frequent affective state. To find likely instances of less common affective states, we randomly selected seven of the videos, located unusually high or low values of various visual features, *i.e.*, local maxima or minima more than 3 standard deviations from the mean, and chose 10-second windows centered at these points. This method yielded 285 clips. To obtain an unbiased sample more representative of typical affective behavior, we randomly chose a total of 60 10-second clips from 10 other videos. We randomly intermixed the two samples and partitioned them into 16 batches, each comprising 15-20 pairs of clips.

We constructed a separate Google form for each batch, with these instructions:

1. This Google form will present a series of pairs of RoboTutor screen video clips for you to annotate.
2. The first clip of each pair contains just the zoomed-in camera input showing the kid.
3. The other clip shows the entire screen, including the camera input. White dots indicate screen touches.
4. Pick the option that fits best. If it fits poorly, or if another option fits almost as well, use the Comments field to explain why.

This protocol first elicited judges' perception of the child's affective state based solely on the video clip of the child, without any additional context, and then based on the video showing the same time interval but in the context of the entire RoboTutor screen. Thus changes in label from the first clip to the second clip could reveal the influence of context on the judge's perception of the affective state.

There were two questions for each clip.

1. Is the student paying attention? *(Yes, No, Can't tell)*
2. Which of the following best describes the kid's state? *(Boredom, Confusion, Delight, Frustration, Neutral, Surprise, or I can't tell.)*

We expected the first question to be easy for both humans and computer to answer based simply on gaze, *i.e.* whether the child was looking at the screen. The second question required finer-grained distinctions, and in fact proved much harder. Figure 2 shows the six affective states manifested while the children were using RoboTutor.

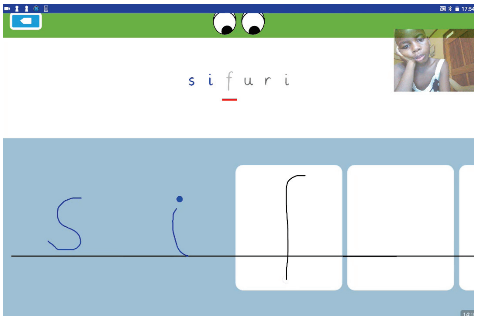
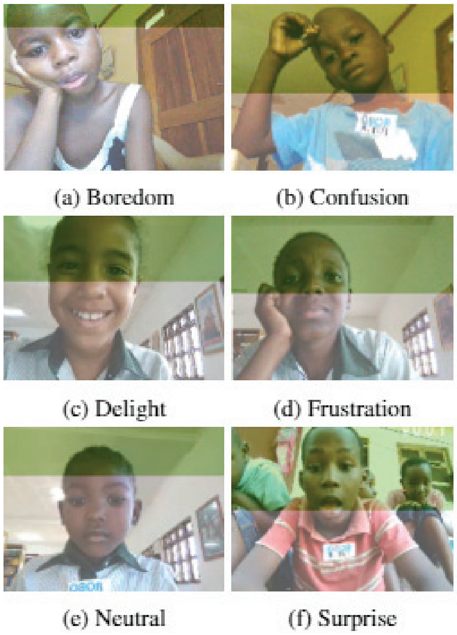


Figure 1: RoboTutor interface with camera window.


(a) Boredom (b) Confusion

(c) Delight (d) Frustration

(e) Neutral (f) Surprise

Figure 2: The six affective states

To label our data, we recruited a Kenyan professor with PhDs in English and International Education, a Tanzanian with a PhD in Instructional Technology, a Tanzanian doctoral student in Linguistics, and two American undergraduate Psychology majors, all of whom were familiar with RoboTutor. It was important to include East African judges not only because they understood the Swahili spoken by RoboTutor and its users, but also because perception of affective states is known to be culture-dependent (Matsumoto 1991). All the judges classified the clips independently.

2.1 Inter-rater reliability

As an intuitive measure of consistency in labeling, we computed the percentage agreement between the judges. To measure the degree to which it exceeded the amount of agreement expected by chance, we computed Cohen's Kappa κ.

To quantify the influence of cultural differences on annotation, we compared pairwise agreement between judges from similar backgrounds (East Africa or USA) versus agreement between judges from different backgrounds. The East African judges agreed 61% of the time, with κ of 0.58, compared to 55% and κ of 0.47 for the American judges, who averaged only 50% agreement with East African judges, with κ of 0.41. That is, judges agreed more often with judges from the same culture than with judges from another culture. Accordingly, we used only the East African judges' labels to train and test the classifier.

Judges agreed on some distinctions more than on others. In particular, they had trouble distinguishing frustration from confusion. One clue to the reason comes from the labeling protocol. The frequency with which judges changed their initial labels, which were based just on the camera input, reflects the extent to which they inferred affective states based at least in part on children's interactions with RoboTutor. Figure 3 shows the transition frequency from label i to label j, represented graphically by the arrow from i to j. The number of label changes was highest for frustration and confusion.

To avoid training our classifier on distinctions with low inter-rater reliability, we combined hard-to-distinguish states, thereby reducing the original set of 6 affective states to just 3 classes, namely *pleasant* (delight and surprise), *unpleasant* (boredom, confusion, and frustration), and *neutral* (flow). Inter-rater reliability was higher for this reduced set, with 67% agreement and κ of 0.63 on the cropped clips. Reliability was higher on the uncropped clips thanks to the additional context they provided, with 73% agreement and κ of 0.65. We used the cropped clips where both the judges agreed. These 231 "consensus" clips were distributed more equally among the 3 classes than among the original 6 affective states: 42 clips were labelled as *pleasant*, 91 as *unpleasant*, and 98 as *neutral*.

3 Approach

Our approach has 4 steps:

1. Extract features from the videos.
2. Aggregate each feature over the 10-second duration of a video clip into a single value.
3. Use semi-supervised learning to train a classifier on labeled and unlabeled data.
4. Use the trained classifier to predict the affective state of a child in a video clip.

We now describe each step in more detail.

3.1 Feature extraction

We started by extracting the camera input, which AZ Screen Recorder displayed over RoboTutor in a translucent window as shown in Figure 1. This window overlapped with the green banner at the top of the screen. Fortunately, this overlap did not prevent us from detecting faces and extracting useful information.

As Figure 4 shows, this information consisted of visual features relevant to affective state, namely head proximity,

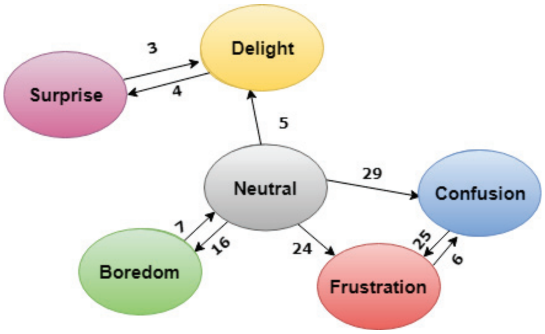


Figure 3: Number of label changes from cropped to uncropped video clip

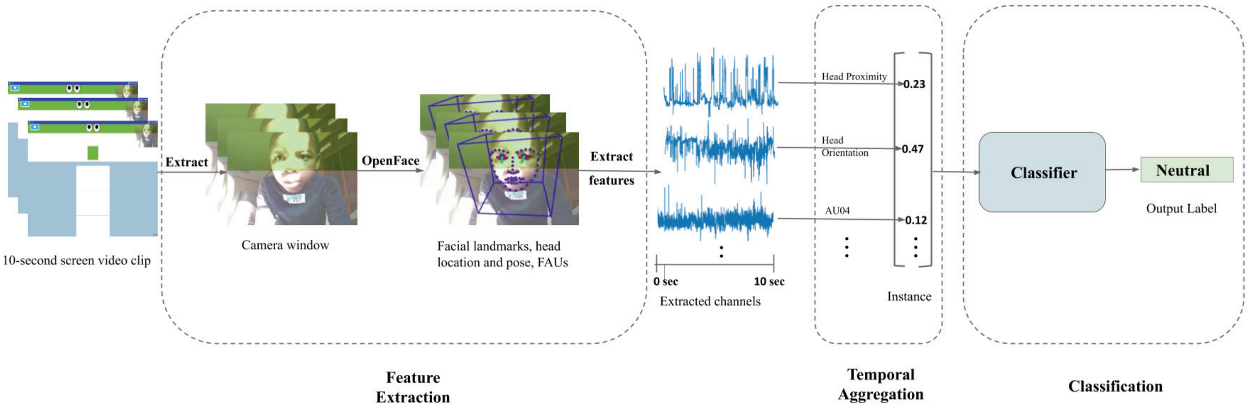


Figure 4: System Architecture.

head orientation, facial action units, blink rate, pupil size, and eye gaze. Each of these features provides a different channel of visual information. To help extract these features from video input, we used OpenFace (Baltrušaitis, Robinson, and Morency 2016), an open-source facial behavior analysis toolkit trained on a large collection of facial data sets, both static images and videos, diverse in age, gender, and ethnicity.

We now describe how we computed and used each feature.

Head proximity: Research on body language has shown that leaning forward indicates an increase in interest and leaning backward shows disinterest (Stanley 2013). This finding motivated us to measure head distance to the camera. Openface gives the location of the head in millimeter coordinates as (H_x, H_y, H_z) in a 3-dimensional reference frame with the camera at the origin, where the X axis is horizontal, the Y axis is vertical, and the camera is pointed along the Z axis. We computed the Euclidean distance of the head from the camera as shown in Equation 1.

$$H_d = \sqrt{(H_x^2 + H_y^2 + H_z^2)} \tag{1}$$

Head orientation: OpenFace computes pitch (R_x), yaw (R_y), and roll (R_z) of the head rotation relative to the location and orientation of the camera. The rotation is in radians around the X, Y, and Z axes. When restless, humans tend to be more fidgety and hence move their heads unconsciously. Therefore, head orientation is the overall angle of the head from the baseline and reflects affective state. Equation 2 specifies head orientation as a function of pitch, yaw, and roll.

$$H_o = R_x * R_y * R_z \tag{2}$$

Facial action units: Prior work on affective state recognizers has focused on Facial Action Units (FAUs) that were most diagnostic of the learning-centered emotions. Following (McDaniel et al. 2007), we employ AU04, AU07, AU12, AU25, AU26, and AU45. For every FAU, OpenFace outputs a classification and regression value. We used the regression value, which characterizes the intensity of the FAU's presence as absent (value $= 0$), low (< 0.2), medium ($0.2 - 0.7$), or high (> 0.7).

Blink rate: Researchers have found that when nervous or troubled, humans' blink rate increases (Haq and Hasan 2016). Using the eye coordinates obtained from OpenFace, we calculated an eye-aspect ratio (Haq and Hasan 2016) for each eye and used the average value of both eyes. If the eye aspect ratio was below a threshold θ for t frames, we considered it to be a blink. We tried different values for these two thresholds. $\theta = 0.4$ and $t = 4$ gave the best accuracy on a sample of 10 video clips. Equation 3 formally defines the eye aspect ratio (E_r) as:

$$E_r = \frac{\|(h - b)\| + \|(f - d)\|}{2 * \|(e - a)\|} \tag{3}$$

where a, b, d, e, f, and h are eye landmark coordinates obtained from OpenFace (Fig. 5).

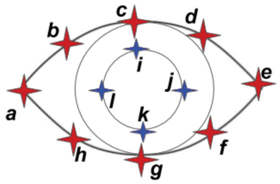


Figure 5: Eye coordinates obtained from OpenFace

Pupil size: Pupil size reflects whether a person is aroused and alert, or bored and fatigued (Kret 2018), so it is a useful indicator of affective state. Using the eye coordinates computed by OpenFace, we determined the ratio of the area of the pupil to the area of the eye using Equation 4. The ratio helped us deal with situations where the child was too close to the screen, leading to a large pupil size, without any role of affect. We used the average of this ratio for both eyes as an input to our classifier. We used Equation 4 to compute this ratio (P_r) as follows:

$$P_r = \frac{\|(l - j)\| * \|(k - i)\|}{\|(e - a)\| * \|(g - c)\|} \tag{4}$$

where *a*, *c*, *e*, *g*, *i*, *j*, *k*, and *l* are eye landmark coordinates obtained from OpenFace (Fig. 5).

Eye gaze: Eye gaze has been used by many researchers to detect alertness, attentiveness, and awareness (Bidwell and Fuchs 2011). OpenFace outputs the gaze direction averaged across the two eyes as *(gaze_angle_x, gaze_angle_y)* in radians. Looking from right to left changes *gaze_angle_x* from negative to positive; looking from up to down changes *gaze_angle_y* from negative to positive. Looking straight ahead is represented as zero for both *gaze_angle_x* and *gaze_angle_y*.

3.2 Temporal aggregation

To avoid the complications and computational cost of time series analysis (Ceballos and Sorrosal 2002), we reduced each feature of a clip to a single summary value. To aggregate discrete features, we simply counted the number of occurrences in the video clip and divided by its duration to obtain a rate, *e.g.* blinks per second. To aggregate continuous features, we experimented with several functions:

- To summarize the feature, we computed its mean over the 10-second window. However, this function can be distorted by outliers.
- To combat distortions caused by noise, especially outliers, we computed its median over the window. However, this function fails to capture significant events shorter than half the duration of the clip.
- To measure the spike caused by the main event in the clip, we computed the maximum value of the feature. However, this function can be distorted by outliers.
- To accentuate spikes while reducing distortion by outliers, we computed the root mean squared value. However, this function is invariant to scrambling the order.
- To select clips based on extreme values of features, i.e. local minima and maxima at least 3 or more standard deviations from the mean, we chose the 10-second clip centered around each extreme value. To focus on its central region, we weighted the mean and root mean squared by dividing the value at each point in the clip by its distance *t* from the midpoint of the clip (plus an offset *o* to prevent division by zero).

We then normalized each aggregate feature value *v* to the interval $[0, 1]$ as $(v - min)/(max - min)$, where max and min are the largest and smallest aggregate values of the feature over the entire set of 10-second clips. We found that proximity-weighted root mean squared achieved the highest cross-validated accuracy when used in a classifier trained as we now describe.

3.3 Semi-supervised learning

Semi-supervised learning (Chapelle, Scholkopf, and Zien 2009) trains a classifier by using unlabeled data to augment sparse labeled data in order to achieve higher classification accuracy.

Algorithm 1 Semi-Supervised Learning

```
Initialize the training set T to the set of labeled instances.
while unlabelled instances remain do
    Train a classifier on T.
    Select an unlabelled instance.
    Run the classifier on it.
    if classifier predicts class C with confidence $\tau$ then
        Add the instance to the training set T with pseudo-label C.
    end
end
```

One can select an unlabelled instance at random, or choose the one closest to a labeled instance. For efficiency, we chose the 10 unlabeled instances closest to any of the labeled instances. In practice the method always exhausted all the unlabelled instances. In theory it could reach a state where it couldn't classify any of them with confidence τ, in which case it should terminate.

We considered several popular classifier learning methods. We chose Random Forest because it performed best on our data (see Table 1). We computed the confidence of a prediction as the percentage of trees in the forest that predicted class C. We set our confidence threshold τ to 0.9.

4 Results

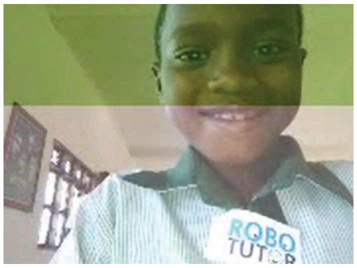


Figure 6: Pleasant Instance

To illustrate how our detector works in practice by using the features defined in Section 3.1, Figure 6 shows an instance classified as a pleasant affective state. After struggling to write the preceding number *0*, the child wrote the number *1* correctly on the first try, and smiled when RoboTutor responded *Mzuri!* ("good" in Swahili). This clip had negligible deflection from typical head orientation. Eye gaze and blink rate were normal, *i.e.*, within one standard deviation of their respective means. However, the child was very close to the screen, *i.e.*, head distance was less than its mean value by more than three standard deviations. Head proximity typically indicates engagement. Also, his pupils were dilated, which typically indicates interest. AU04 (Brow Lowerer) and AU45 (Blink) were absent, AU25 (Lips Part) was present with low intensity, AU07 (Lid Tightener) and AU26 (Jaw Drop) were present with medium intensity, and AU12 (Lip Corner Puller) was present with high intensity. As this example illustrates, our detector's recognition of pleasant instances is probably influenced by head proximity, pupil size, and smiling. We say "probably" because a random forest's calculations are too complicated to readily analyze the individual influence of all the features. Instead, we described their values relative to their distributions, on the assumption that unusual values are likely to affect the classifier output.

4.1 Quantitative Evaluation

To estimate performance on unseen children, we used leave-1-child-out cross-validation, training the classifier on all but one child and testing it on the held-out child. We performed this process for 5 randomly chosen children and report the median results. **Accuracy** is the percentage of test instances classified correctly. This measure is simplest and has practical significance because it predicts performance on unseen data drawn from the same distribution. However, it is sensitive to that distribution. In contrast, the following weighted measures are weight-averaged across all three classes, and therefore independent of the training set distribution. Each class is assigned a weight equal to the ratio of the number of instances in that class to the total number of instances in the test set. **Weighted precision** is the percentage correct among the instances classified as positive. **Weighted recall** is the percentage correct among the true positive instances. **Weighted F1** is the harmonic mean of weighted recall and weighted precision. As unlabelled data, we used 1007 clips from 20 videos, sampled using the same sampling techniques described in Section 2.

As Table 1 shows, our method beat the supervised learning methods on all four criteria:

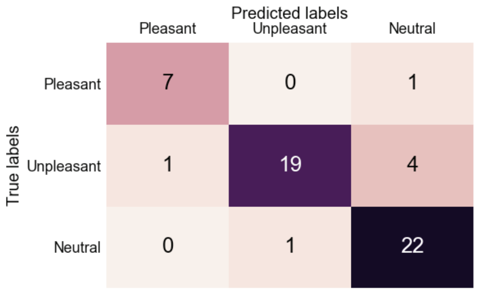
Classifier	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.21	0.52	0.21	0.23
Decision Tree	0.44	0.43	0.44	0.43
SVM	0.51	0.49	0.51	0.46
Adaboost	0.47	0.47	0.47	0.47
Logistic Regression	0.53	0.50	0.53	0.51
KNN	0.53	0.53	0.53	0.53
Random Forest	**0.61**	**0.63**	**0.61**	**0.60**
Our Approach	**0.88**	**0.88**	**0.84**	**0.86**

Table 1: Comparison with supervised learning.

These results were for the consensus data where both judges agreed. We further tested our classifier on all our labeled test data, including 55 consensus clips not used for training and 114 clips on which the judges disagreed. This experiment helped us quantify the degradation in performance due to label noise. We evaluated the accuracy of the prediction compared to both judges' labels and took the average. As expected, average accuracy dropped from 88% to 56% when we included the non-consensus labels, compared to testing on the consensus data alone. The lower accuracy on the unfiltered data reflects the inherent difficulty of replicating subjective judgments on which human experts disagree.

To estimate the effect of cultural differences, we tested our trained classifier on both sets of consensus labels, African and American. Accuracy fell from 88% to 57% when tested on American labels. This difference quantifies the effect of cultural influence on people's facial expressions and other visual cues, and the consequent importance of recruiting judges from the same culture to label their affective states.

4.2 Error Analysis



	Pleasant	Unpleasant	Neutral
Pleasant	7	0	1
Unpleasant	1	19	4
Neutral	0	1	22

Figure 7: Confusion Matrix

Figure 7 shows that our classifier performed well in most cases. However, the classifier incorrectly characterized four unpleasant instances as neutral. Most of these misclassifications involved boredom. Boredom is not easily distinguishable from neutral based on facial features. Indeed, boredom typically lacks facial expression. To detect boredom, we may have to use additional indicators, such as posture and acoustic-prosodic features of speech.

Accuracy is limited by the quality of the visual features input by the classifier from OpenFace, which depend in turn on its face detection. Our data come from authentic settings subject to varying illumination and occlusion. Consequently, OpenFace occasionally (especially in low-light conditions) fails to detect a face when it is present. Inspection of sample videos showed that OpenFace failed to detect a face approximately 2% of the time, typically for a second at a time.

4.3 Sensitivity analysis

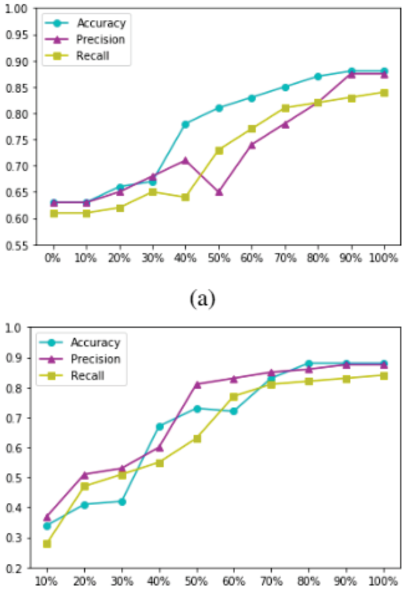
To explore the sensitivity of the results to different factors, we varied the amount of unlabeled data, the amount of labeled data, and the method for selecting unlabeled data.

Effect of amount of unlabeled data: How did test accuracy vary with the amount of unlabeled data? We started with no unlabeled data, i.e., supervised learning, and added 10% of the unlabeled data at each iteration until all of the unlabeled data was utilized. Figure 8(a) shows that as the number of unlabeled instances increased from zero to 1007, accuracy rose asymptotically from 61% to 88%.

Effect of amount of labeled data: To analyze the effect of the amount of labeled data on classifier performance, we varied the percentage of labelled instances used from 10% to 100%, keeping the unlabeled training set constant, i.e. 1007 instances. Figure 8(b) shows that:

1. Accuracy, precision, and recall increased with the amount of labeled data, as expected.
2. Too little labeled data produced poor results, even with all the unlabeled data.

Effect of choice of data to pseudo-label: We conducted an experiment to understand why unlabeled data helped, and where the action was. We hypothesized that the order in which semi-supervised learning chose unlabeled instances to pseudo-label had a substantial effect on the performance



(a)

(b)

Figure 8: Classifier Performance vs. (a) Number of Unlabeled Training Instances (b) Number of Labeled Training Instances

of the resulting classifier. To test this hypothesis, we compared choosing 10 random instances at each iteration versus choosing the 10 instances closest to the instances labeled (or pseudo-labeled) so far.

Data pseudo-labeled	Accuracy	Precision	Recall	F1-Score
10 Random instances	0.78	0.78	0.72	0.74
10 closest instances	**0.88**	**0.88**	**0.84**	**0.86**

Table 2: Effect of order of pseudo-labeling

Table 2 shows that choosing the 10 nearest instances at each iteration performed better than choosing 10 random instances. Why? Semi-supervised learning exploits the continuity assumption that instances near each other are likelier to belong to the same class than other instances assigned to that class based solely on generalization by a classifier trained on incomplete training data.

5 Conclusion

Contributions: We presented an innovative, multi-channel method for automating affect detection in a tablet app solely by integrating visual cues extracted from its front-facing camera input. We used semi-supervised learning to leverage our sparsely labeled training data. We evaluated it against human judges on authentic data from a novel population of children using RoboTutor in natural settings, and analyzed its performance both quantitatively and qualitatively. This work constitutes significant progress in automated affect detection, whether to improve tutor design off-line or to respond to student affect at runtime.

Limitations and future work: To increase inter-rater reliability, we combined confusable affective states into the same class. Future work to distinguish them could enhance RoboTutor's emotional intelligence.

The evaluated method is based solely on input from the tablet's front-facing camera, consistent with our focus on identifying what information about affect we can derive from visual cues. This type of input is more practical in realistic settings than inputs currently available in lab settings, such as EEG, pressure sensors, or even video from external cameras. However, some other types of input are readily available to a tablet tutor.

In particular, tablets input audio. Speech input is pedagogically informative when it can accurately be recognized or analyzed for other properties, such as prosody. Both these uses of audio input are feasible in quiet lab settings. However, in natural settings where multiple children use tablets in close proximity and noise-canceling headset microphones are too fragile or expensive, audio input is liberally contaminated with background speech from other children and their tablets.

The tutor itself could be a fruitful source of information, including its internal states, decisions, and actions, and student input such as screen taps and other gestures. Such data is tutor-specific but informative, and we plan to exploit it in the future, especially to recognize the contextual clues that our judges used to distinguish among boredom, confusion, and frustration.

We implemented our detector on a Windows PC. We can use it off-line to analyze screen-recorded sessions for guidance in redesigning RoboTutor. In principle, it could be applied to any screen capture video that includes camera input of the user, whether from the front-facing camera of a tablet, or the webcam atop a computer monitor.

However, RoboTutor itself runs on Android tablets. Incorporating the detector into RoboTutor will require porting it to an Android tablet to detect facial expressions and other visual cues in real time. We will also need to redesign RoboTutor to respond to detected affective states, evaluate the effects of such responses, and refine them accordingly. These responses should improve RoboTutor's ability to engage children and help them learn, and may generalize usefully to other tutors as well.

Acknowledgements

We thank the team that developed RoboTutor and the children who used it. We would like to express our gratitude to Fortunatus Massewe for recording many hours of RoboTutor screen video, and Dr. Leonora Anyango Kivuva, Joash Gambarage, Sheba Naderzad, and Andrew McReynolds for labeling so much data. We would also like to thank Rachel Burcin, Dr. John Dolan and Mikayla Trost for tirelessly supporting us throughout the summer. The first author was a Robotics Institute Summer Scholar supported by an S.N. Bose Scholarship.

References

Baltrušaitis, T.; Robinson, P.; and Morency, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10. IEEE.

Bidwell, J., and Fuchs, H. 2011. Classroom analytics: Measuring student engagement with automated gaze tracking. *Behav Res Methods* 49:113.

Ceballos, D., and Sorrosal, M. 2002. Time aggregation problems in financial time series. In *MS'2002 International Conference on Modelling and Simulation in Technical and Social Sciences*, 25–27.

Chapelle, O.; Scholkopf, B.; and Zien, A. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks* 20(3):542–542.

Craig, S.; Graesser, A.; Sullins, J.; and Gholson, B. 2004. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of educational media* 29(3):241–250.

Craig, S. D.; D'Mello, S.; Witherspoon, A.; and Graesser, A. 2008. Emote aloud during learning with autotutor: Applying the facial action coding system to cognitive–affective states during learning. *Cognition and Emotion* 22(5):777–788.

D'Mello, S., and Graesser, A. 2009. Automatic detection of learner's affect from gross body language. *Applied Artificial Intelligence* 23(2):123–150.

D'Mello, S. K., and Graesser, A. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction* 20(2):147–187.

D'Mello, S., and Graesser, A. 2012. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(4):23.

D'Mello, S.; Picard, R. W.; and Graesser, A. 2007. Toward an affect-sensitive autotutor. *IEEE Intelligent Systems* 22(4):53–61.

Du, S.; Tao, Y.; and Martinez, A. M. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* 111(15):E1454–E1462.

Egger, H. L.; Pine, D. S.; Nelson, E.; Leibenluft, E.; Ernst, M.; Towbin, K. E.; and Angold, A. 2011. The nimh child emotional faces picture set (nimh-chefs): a new set of children's facial emotion stimuli. *International journal of methods in psychiatric research* 20(3):145–156.

Faria, D. R.; Vieira, M.; Faria, F. C.; and Premebida, C. 2017. Affective facial expressions recognition for human-robot interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 805–810. IEEE.

Haq, Z. A., and Hasan, Z. 2016. Eye-blink rate detection for fatigue determination. In *2016 1st India International Conference on Information Processing (IICIP)*, 1–5. IEEE.

Hess, U.; Adams, R. B.; and Kleck, R. E. 2007. Looking at you or looking elsewhere: The influence of head orientation on the signal value of emotional facial expressions. *Motivation and Emotion* 31(2):137–144.

Kaulard, K.; Cunningham, D. W.; Bülthoff, H. H.; and Wallraven, C. 2012. The mpi facial expression database—a validated database of emotional and conversational facial expressions. *PloS one* 7(3):e32321.

Kret, M. E. 2018. The role of pupil size in communication. is there room for learning? *Cognition and Emotion* 32(5):1139–1145.

Matsumoto, D. 1991. Cultural influences on facial expressions of emotion. *Southern Journal of Communication* 56(2):128–137.

McDaniel, B.; D'Mello, S.; King, B.; Chipman, P.; Tapp, K.; and Graesser, A. 2007. Facial features for affective state detection in learning environments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

Michel, P., and El Kaliouby, R. 2003. Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces*, 258–264. ACM.

Mostow, J. 2019. robotutor.org.

Nojavanasghari, B.; Baltrušaitis, T.; Hughes, C. E.; and Morency, L.-P. 2016. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th acm international conference on multimodal interaction*, 137–144. ACM.

Partala, T., and Surakka, V. 2003. Pupil size variation as an indication of affective processing. *International journal of human-computer studies* 59(1-2):185–198.

Petrantonakis, P. C., and Hadjileontiadis, L. J. 2009. Emotion recognition from eeg using higher order crossings. *IEEE Transactions on Information Technology in Biomedicine* 14(2):186–197.

Picard, R. W. 2000. *Affective computing*. MIT press.

PlayStore, G. 2018. Az screen recorder.

Reddy, R. P.; Krishna, P. M.; Narayanan, V.; and Lalitha, S. 2018. Affective state recognition using image cues. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 928–933. IEEE.

Stanley, D. 2013. Measuring attention using microsoft kinect.

Valstar, M., and Pantic, M. 2010. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 65. Paris, France.

Westlund, J. K.; D'Mello, S. K.; and Olney, A. M. 2015. Motion tracker: Camera-based monitoring of bodily movements using motion silhouettes. *PloS one* 10(6):e0130293.

Woolf, B.; Burleson, W.; Arroyo, I.; Dragon, T.; Cooper, D.; and Picard, R. 2009. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology* 4(3-4):129–164.

XPRIZE. 2015. learning.xprize.org.