Automatically Recognizing Facial Expression: Predicting Engagement and Frustration

Joseph F. Grafsgaard[1], Joseph B. Wiggins[1], Kristy Elizabeth Boyer[1], Eric N. Wiebe[2], James C. Lester[1]

[1]Department of Computer Science [2]Department of STEM Education
North Carolina State University, Raleigh, NC, USA

{jfgrafsg, jbwiggi3, keboyer, wiebe, lester}@ncsu.edu

ABSTRACT

Learning involves a rich array of cognitive and affective states. Recognizing and understanding these cognitive and affective dimensions of learning is key to designing informed interventions. Prior research has highlighted the importance of facial expressions in learning-centered affective states, but tracking facial expression poses significant challenges. This paper presents an automated analysis of fine-grained facial movements that occur during computer-mediated tutoring. We use the Computer Expression Recognition Toolbox (CERT) to track fine-grained facial movements consisting of eyebrow raising (inner and outer), brow lowering, eyelid tightening, and mouth dimpling within a naturalistic video corpus of tutorial dialogue (N=65). Within the dataset, upper face movements were found to be predictive of engagement, frustration, and learning, while mouth dimpling was a positive predictor of learning and self-reported performance. These results highlight how both intensity and frequency of facial expressions predict tutoring outcomes. Additionally, this paper presents a novel validation of an automated tracking tool on a naturalistic tutoring dataset, comparing CERT results with manual annotations across a prior video corpus. With the advent of readily available fine-grained facial expression recognition, the developments introduced here represent a next step toward automatically understanding moment-by-moment affective states during learning.

Keywords

Facial expression recognition, engagement, frustration, affect, computer-mediated tutoring

1. INTRODUCTION

Over the past decade, research has increasingly highlighted ways in which affective states are central to learning [6, 21]. Learning-centered affective states, such as engagement and frustration, are inextricably linked with the cognitive aspects of learning. Thus, understanding and detecting learner affective states has become a fundamental research problem. In order to identify students' affective states, researchers often investigate nonverbal behavior. A particularly compelling nonverbal channel is facial expression, which has been intensely studied for decades. However, there is still a need to more fully explore facial expression in the context of learning [6].

Recent research has identified facial expressions that are related to self-reported and judged learning-centered affective states [1, 7, 9, 18, 25], which typically include boredom, confusion, engaged concentration, and frustration. However, more research is needed to fully explore the relationships between facial movement and learning-centered affective states. For instance, timing and intensity of facial expressions have only just begun to be explored in the context of learning [18].

The Facial Action Coding System (FACS) [10] has been widely used to study detailed facial movements for decades. FACS enumerates the possible movements of the human face as facial *action units*. Thus, FACS is an objective measure used to identify facial configurations before interpreting displayed affect. Because FACS quantifies facial movements present in displays of emotion, it allows researchers to identify facial components of learning-centered affect, which have been found to be different from those in everyday emotions [4, 6, 7, 9, 18, 27]. Identifying these action units is a time-intensive manual task, but a variety of computer vision tools are in current use, most often focusing on tracking facial feature points [4, 27]. Facial feature tracking tools recognize the presence of a face and then locate facial features such as the corners of the mouth and eyes. Generally, there are two distinct families of tools: low-level tools that track facial features [3] (e.g., which way the head is turned and where points are positioned) and tools that provide affective interpretations [17, 23, 24] (e.g., smiling, emotions). However, the tool used in this study, the Computer Expression Recognition Toolbox (CERT), offers a mid-level alternative. CERT produces intensity values for a wide array of FACS facial action units, thus enabling fine-grained analyses of facial expression [19].

This paper presents an automated facial recognition approach to analyzing student facial movements during tutoring and an examination of the extent to which these facial movements correspond to tutoring outcomes. The novel contributions are two-fold. First, the output of the facial action unit tracking tool, CERT, was validated through comparing CERT output values with manual FACS annotations. The results indicate excellent agreement at the level of presence versus absence of facial movements. Naturalistic video is challenging for computer vision techniques, and this validation is the first of its kind on a naturalistic tutoring video corpus. Second, models were constructed to examine whether the intensity and frequency of facial expressions predict tutoring outcomes. The results show that several specific facial movements predict tutoring outcomes.

For instance, brow lowering intensity (i.e., the magnitude of the CERT output value) was associated with reduced perception of the tutoring session as being worthwhile, and greater self-reported frustration. Additionally, frequency of mouth dimpling predicts increased learning gains and self-reported task success. These results represent a next step toward large-scale analyses and understanding of learning-centered affective states in tutoring.

2. RELATED WORK

D'Mello and colleagues have a longstanding line of research into the mechanisms of facial expression and learning-centered affective states. In recent years, stable correlations between specific facial action units and self-reported or judged affective states have been identified [7, 9]. Brow lowering (AU4) and eyelid tightening (AU7) were correlated with confusion, while inner and outer brow raising (AU1, AU2) were correlated with frustration.

Another prominent line of research is that of Baker and colleagues. After extensively observing student nonverbal behaviors during interactions with tutoring systems, they developed a protocol for judging students' affective states, such as boredom or engagement [1, 22]. This has enabled lightweight annotation of affective states across a wide variety of classrooms. Automated tools extend these approaches to studying student facial expressions, with potential to confirm current hypotheses across large-scale datasets.

The intelligent tutoring systems community has also begun integrating real-time facial expression tracking into studies of learning-centered affective states [5, 8]. These studies are a parallel line of research to that of understanding student affect. Incorporating nonverbal behavior tracking into intelligent tutoring systems is a necessary step toward meaningful real-time affective interventions. There has also been recent research that may lead to robust sensor-free affect detection [2]. Such an approach identifies patterns of behavior in log data that are associated with observed affective states. Then, models are built from the log data alone to predict affective states.

In prior research toward automated analysis of learning-centered affect, the creators of CERT applied the tool to video corpora taken during demanding tasks [18, 25]. Particularly, the facial expressions of children were investigated in order to compile a set of facial expressions relevant to the younger population [18]. These studies inform the use of automated facial expression recognition. A key difference in the present study is that we are presenting a comparatively much larger scale of analysis (over 80 times the duration of video). Additionally, we conducted a novel validation that compared values of CERT output with manual FACS annotations across a naturalistic tutoring video corpus.

3. TUTORING VIDEO CORPUS

The corpus consists of computer-mediated tutorial dialogue for introductory computer science collected during the 2011-2012 academic year. Students (*N*=67) and tutors interacted through a web-based interface that provided learning tasks, an interface for computer programming, and textual dialogue. The participants were university students in the United States, with average age of 18.5 years (*stdev*=1.5). The students voluntarily participated for course credit in an introductory engineering course, but no prior computer science knowledge was assumed or required. Each student was paired with a tutor for a total of six sessions on different days, limited to forty minutes each session. Recordings of the sessions included database logs, webcam video, skin conductance, and Kinect depth video. This study analyzes the webcam video corpus. The student workstation configuration is shown in Figure 1. The JavaTutor interface is shown on the next page in Figure 3.

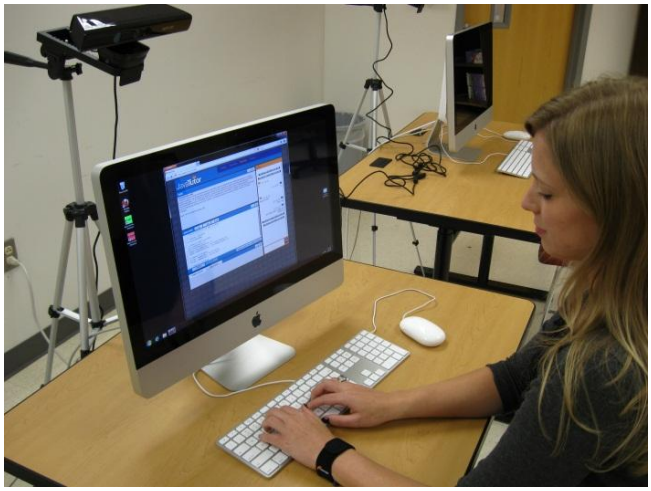


Figure 1. Student workstation with depth camera, skin conductance bracelet, and computer with webcam

Before each session, students completed a content-based pretest. After each session, students answered a post-session survey and posttest (identical to the pretest). The post-session survey items were designed to measure several aspects of engagement and cognitive load. The survey was composed of a modified User Engagement Survey (UES) [20] with Focused Attention, Endurability, and Involvement subscales, and the NASA-TLX workload survey [16], which consisted of response items for Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration Level. Student survey items relevant to the results presented in Section 4 are shown in Figure 2. Students were intentionally not asked about a wider set of emotions in order to avoid biasing their future interactions.

Endurability (UES):
Working on this task was worthwhile.
I consider my learning experience a success.
My learning experience was rewarding.
I would recommend using JavaTutor to my friends and family.

Temporal Demand (NASA-TLX):
How hurried or rushed was the pace of the task?

Performance (NASA-TLX):
How successful were you in accomplishing what you were asked to do?

Frustration Level (NASA-TLX):
How insecure, discouraged, irritated, stressed, and annoyed were you?

Figure 2. Subset of student post-session survey items

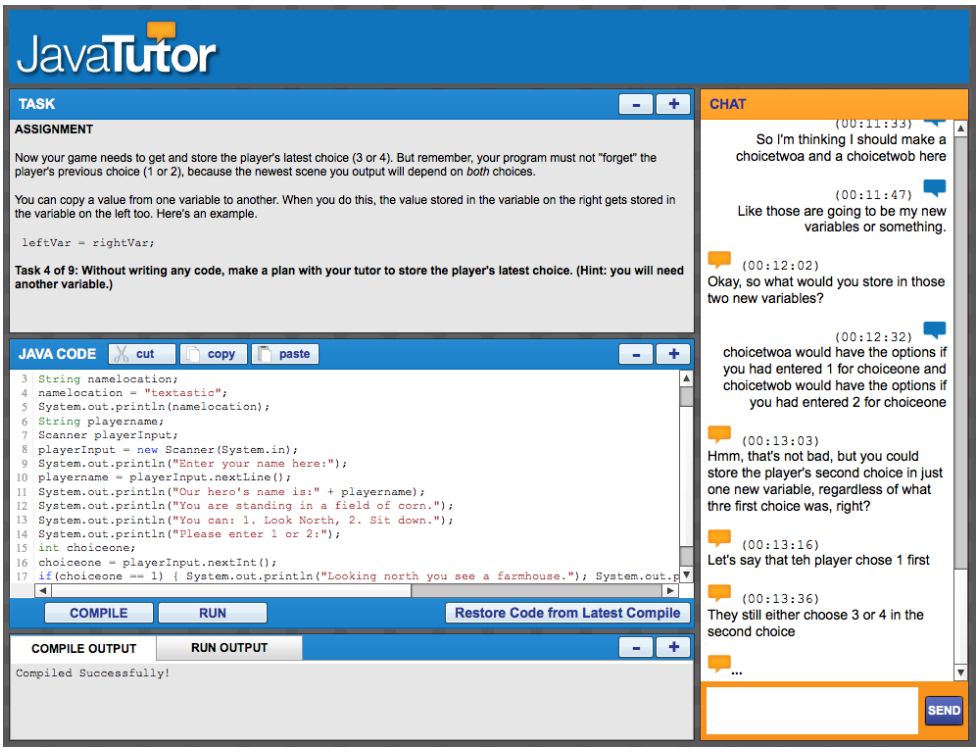


Figure 3. The JavaTutor interface

The tutoring video corpus is comprised of approximately four million video frames totaling thirty-seven hours across the first tutoring session. Two session recordings were missing due to human error (*N*=65). The recordings were taken at 640x480 pixel resolution and thirty frames per second. CERT successfully tracked faces across a great majority of the tutoring video corpus (*mean*=83% of frames tracked, *median*=94%, *stdev*=23%).

3.1 Facial Expression Recognition

The Computer Expression Recognition Toolbox (CERT) [19] was used in this study because it allows frame-by-frame tracking of a wide variety of facial action units. CERT finds faces in a video frame, locates facial features for the nearest face, and outputs weights for each tracked facial action unit using support vector machines. For a detailed description of the technology used in CERT, see [26].

Based on observations from prior studies [12, 13], we selected a subset of the 20 facial action units that CERT detects as the focus of the present analyses. This set of facial action units was informed by a prior naturalistic tutoring video corpus [13], used in this study as a validation set, consisting of approximately 650,000 FACS-annotated video frames and seven tutoring sessions. In this corpus, sixteen facial action units were annotated. The five most frequently occurring action units each occurred in over 10% of the facial expression events. The remaining facial action units occurred substantially less frequently. The five frequently occurring action units were selected for the further analysis presented here on the new corpus. Table 1 shows the relative frequency of each action unit's participation in discrete facial expression events and the number of frames annotated with each action unit from the validation corpus.

A screenshot of CERT processing is shown in Figure 4. In the course of processing videos with CERT, we noted that the range of output values can vary between individuals due to their hair, complexion, or wearing eyeglasses or hats. This has also been noted by the creators of CERT [26]. In order to better capture instances of facial expression displays, we introduce an adjustment procedure for individual tracking differences. First, the average output value for each student was computed for each action unit. These values correspond to individual baselines of facial expression. The average output value per session was subtracted for each action unit, resulting in individually adjusted CERT output. This adjustment was applied to all CERT values presented in this paper. Automatically recognized instances of the selected action units are shown in Figure 5, with corresponding adjusted CERT output. While any positive output value indicates that CERT recognizes an action unit, we used an empirically determined threshold of 0.25 to reduce the potential for false positives. This threshold was based on observations of CERT output in which action unit instances that were more than slightly visible corresponded with output values above 0.25. CERT successfully tracked faces across a large majority of the validation corpus (*mean*=76% of frames tracked, *median*=87%, *stdev*=23%).

Table 1. The five most frequent facial action units in the validation corpus [13]

Facial action unit	Frames	Event Freq.
AU1: Inner Brow Raiser	12,257	15.5%
AU2: Outer Brow Raiser	15,183	21.7%
AU4: Brow Lowerer	127,510	18.6%
AU7: Lid Tightener	9,474	13.2%
AU14: Dimpler	14,462	24.2%

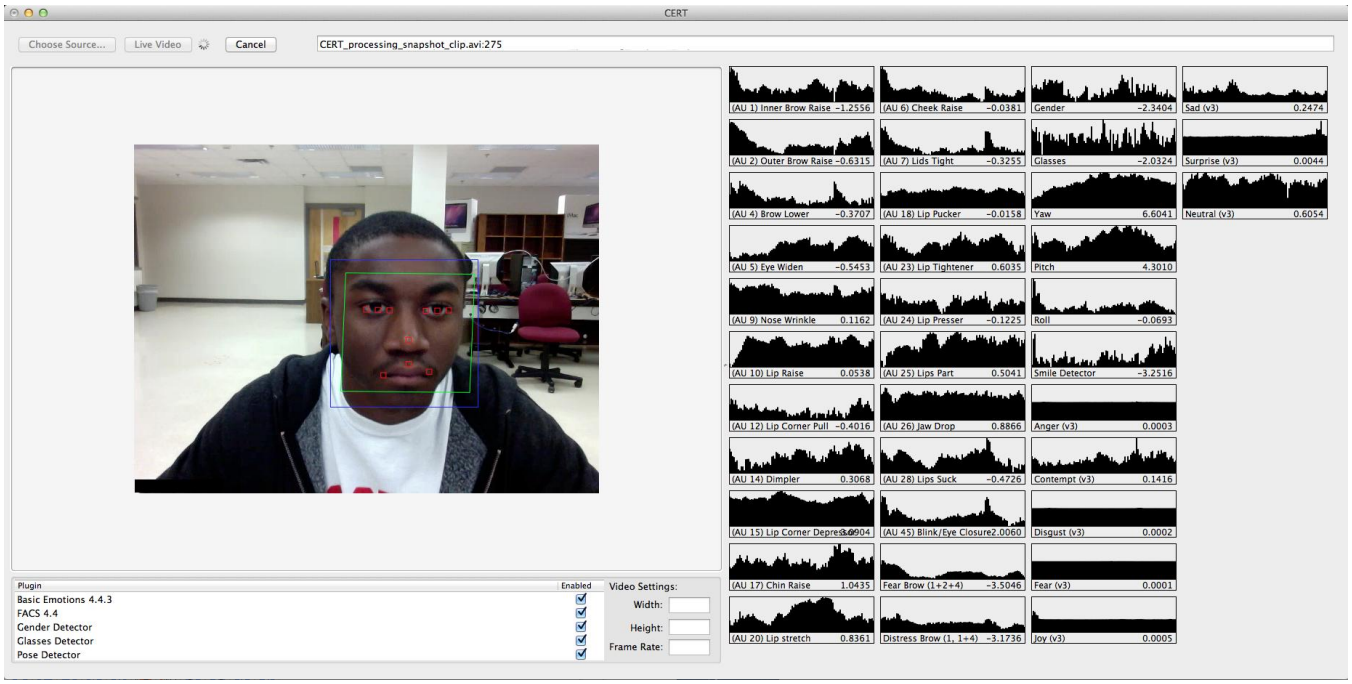
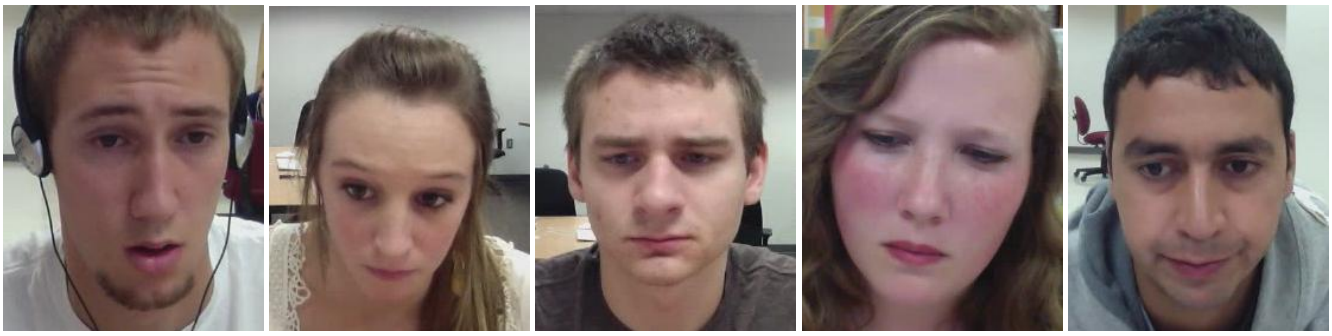


Figure 4. Screenshot of CERT video processing



AU1(0.80) AU2(-0.12) **AU4(0.25)** AU7(-0.23) AU14(-0.06)	AU1(0.17) **AU2(0.27)** AU4(0.08) AU7(-0.09) AU14(-0.53)	AU1(-0.02) AU2(0.07) **AU4(0.47)** AU7(0.08) AU14(-0.85)	AU1(-0.22) AU2(-0.27) AU4(0.11) **AU7(0.26)** AU14(-0.04)	AU1(-0.02) AU2(0.00) AU4(-0.04) AU7(0.03) **AU14(0.46)**
AU1 and AU4: Inner brow raiser and brow lowerer	**AU2: Outer brow raiser**	**AU4: Brow lowerer**	**AU7: Lid tightener**	**AU14: Dimpler**

Figure 5. Automatically recognized facial action units (bold values are above selected threshold of 0.25)

3.2 Validation

CERT was developed using thousands of posed and spontaneous facial expression examples of adults outside of the tutoring domain. However, naturalistic tutoring data often has special considerations, such as a diverse demographic, background noise within a classroom or school setting, no controls for participant clothing or hair, and facial occlusion from a wide array of hand-to-face gesture movements. Therefore, we aim to validate CERT's performance within the naturalistic tutoring domain. CERT's adjusted output was compared to manual annotations from a validation corpus, as described in Section 3.1.

The creators of CERT have applied the tool to the problem of understanding children's facial expressions during learning. To validate CERT's output, they compared it with manual FACS annotations across 200 video frames [18]. However, the goal in this analysis is to validate CERT's performance across a validation corpus of approximately 650,000 video frames. It is important to know whether average CERT output values for video frames with a specific facial movement are different from those without that facial movement. If the values are differentiable, then CERT may be an appropriate tool for general use at a large scale. If the values cannot be distinguished, then CERT is likely to provide many false positives and false negatives. Thus, this novel validation analysis provides needed

insight into how well CERT performs across an entire corpus. The design of the validation analysis is shown in Figure 6.

Adjust CERT Output: Adjusting CERT output values with a baseline for each individual allows for comparison across students

↓

Binary Split on AU: Output values for each student are divided between AU-present and AU-not-present and then averaged; these values serve as the input variable for logistic regression

↓

Build Predictive Model: A logistic regression model is built from all students' average values for AU-present and AU-not-present, producing predicted categories of AU/Not-AU

↓

Compare Predictions: Compute Cohen's Kappa and accuracy from the logistic regression model predictions and manual tags to evaluate how well AU/Not-AU was predicted

Figure 6. Design of the validation analysis

Adjusted CERT output was computed for each video frame as described in Section 3.1. The CERT output values were then averaged within five binary splits, one for each facial action unit under consideration. Each binary split was comprised of frames with a specific facial action present and frames without that particular action unit, as labeled in the validation corpus. For example, to evaluate performance on brow lowering (AU4), video frames were divided between presence or absence of AU4 via the manual annotations. Once the binary split was performed, the frames were further subdivided by student. Thus, each student has an average value for frames with a specific action unit present and an average value for frames without that action unit. Logistic regression models were constructed using the average value as the sole parameter. One logistic regression model was built per action unit, for a total of five. The binary response variable categories (action unit present/action unit absent) were produced from each regression model. The predicted categories were compared to the categories from manual annotation, yielding Cohen's κ and percent accuracy.

The validation results show that CERT output has an excellent capability to distinguish facial expression events from baseline across the validation corpus, yielding an average κ across the five action units of 0.82. Naturalistic data is challenging for computer vision techniques, so the validation analysis confirms the accuracy of CERT facial expression recognition. Table 2 displays the validation results.

Table 2. Comparison of agreement on validation corpus [13] Manual FACS vs. logistic regression of CERT output

FACS Coder	AU1	AU2	AU4	AU7	AU14
Manual κ^*	0.88	0.82	0.79	0.78	0.73
CERT κ^*	0.86	0.86	0.68	1	0.71
CERT Accuracy*	93%	93%	85%	100%	86%

*Manual κ on face events; CERT evaluated on avg. output

In order to explore the effectiveness of the correction for individual differences described in Section 3.1, the validation analysis was performed again, this time without corrected output values. With raw CERT output, the logistic regression models could not distinguish between the average values for AU-present versus AU-not-present (Table 3). Thus, agreement with the manual annotations was poor. The validation analyses illustrate that CERT output should be corrected with average values if a comparison across individuals is desired. This correction is straightforward to apply in post-processing. In a real-time application of such a tool, a running average could be computed at each video frame.

Table 3. Secondary validation analysis on raw CERT output

	AU1	AU2	AU4	AU7	AU14
CERT κ	0.14	0.29	0.05	0.29	0.29
CERT Accuracy	57%	64%	54%	64%	64%

A difficulty that remains for facial expression recognition is face occlusion, where the face is covered by an object, hand, etc. One source of face occlusions is hand-to-face gestures [14], where one or two hands touch the lower face. These gestures are particularly prominent in our tutoring video corpus, as students often place a hand to their face while thinking or cradle their head in both hands while apparently tired or bored. These gestures can result in loss of face tracking or incorrect output. Accordingly, our analyses considered only video frames where face tracking and registration were successful (i.e., where CERT produced facial action unit output). Examples of both types of occlusion errors are shown in Figure 7. The CERT adjusted output values for the mostly occluded face frame (in the left image) are [AU1 = 1.34, AU2 = 0.65, AU4 = 0.62, AU7 = 0.30, AU14 = -1.17]. If these values are interpreted with the 0.25 threshold, then they represent presence of multiple action units, but that is clearly not the case when viewing the video. CERT was unable to find the student's face in the partially occluded frame (in the right image), though the presence of brow lowering is apparent. While hand-to-face gestures present a significant complication in naturalistic tutoring data, there has been preliminary progress toward automatically detecting these gestures [14], so their effect may be mitigated in future facial expression tracking research.

Figure 7. Facial recognition errors due to gestures: mostly occluded (left) and partially occluded (right)

4. PREDICTIVE MODELS

Automated facial expression recognition enables fine-grained analyses of facial movements across an entire video corpus.

With such tracking, there is potential to discover previously unidentified ways in which both frequency [7] and intensity [18] of facial expressions inform diagnosis of student affective states. A first step toward this possibility is to quantify facial expressions as they occurred throughout tutoring and compare these with tutorial outcomes. Therefore, predictive models of both affective and learning outcomes were built leveraging both the average intensity and frequency of facial movements. Refer to Figure 5 for example images of the facial action units.

Predictive models were constructed using minimum Bayesian Information Criterion (BIC) in forward stepwise linear regression, using JMP statistical software. These models are conservative in how they select predictive features because the explanatory value of added parameters must offset the BIC penalty for model complexity. Tutoring outcomes (affective and learning) were the dependent variables. Therefore, a model was constructed to predict each of the post-session survey scales and normalized learning gain (ten in total). The models for which facial action unit features were significantly explanatory are described below.

4.1 Facial Action Units and Affective Outcomes

Endurability was the student's self-report of whether he or she found the tutoring session to be worthwhile and whether he or she would recommend JavaTutor tutoring to others. Endurability was predicted by inner brow raising (AU1) intensity and brow lowering (AU4) intensity. AU1 was a positive predictor, while AU4 was negative. After adjusting for degrees of freedom (i.e., the number of model parameters), the model effect size was r = 0.37. The model is shown in Table 4.

Table 4. Stepwise linear regression model for Endurability

Endurability =	**Partial R^2**	**Model R^2**	***p***
-10.58 * *AU4_Intensity*	0.088	0.088	0.004
6.60 * *AU1_Intensity*	0.075	0.162	0.023
16.61 (intercept)			<0.001
RMSE = 10.01% of range in Endurability scale			

Temporal demand captures the student's self-report of whether he or she felt rushed or hurried during the session. Temporal demand was negatively predicted by outer brow raising (AU2) frequency; that is, students with higher frequency of this action unit reported feeling more rushed during the session. The adjusted model effect size was r = 0.23. The model is shown in Table 5.

Table 5. Stepwise linear regression model for Temporal Demand

Temporal Demand =	**Partial R^2**	**Model R^2**	***p***
-103.15 * *AU2_Freq*	0.068	0.068	0.037
34.90 (intercept)			<0.001
RMSE = 19.69% of range in Temporal Demand scale			

Performance was the student's self-report of how successful he or she felt in accomplishing the task. Performance was positively predicted by frequency of mouth dimpling (AU14), so students who displayed AU14 more frequently reported a higher sense of performance. The adjusted model effect size was r = 0.26. The model is shown in Table 6.

Table 6. Stepwise linear regression model for Performance

Performance =	**Partial R^2**	**Model R^2**	***p***
64.65 * *AU14_Freq*	0.081	0.081	0.022
72.74 (intercept)			<0.001
RMSE = 8.50% of range in Performance scale			

Frustration was the student's self-report of how insecure, agitated or upset he or she was during the tutoring session. Frustration was positively predicted by intensity of brow lowering (AU4); that is, students who displayed more intense AU4 reported feeling more insecure, agitated, or upset. The adjusted model effect size was r = 0.29. The model is shown in Table 7.

Table 7. Stepwise linear regression model for Frustration

Frustration =	**Partial R^2**	**Model R^2**	***p***
77.27 * *AU4_Intensity*	0.098	0.098	0.011
-15.34 (intercept)			0.165
RMSE = 17.05% of range in Frustration scale			

4.2 Facial Action Units and Learning Gain

We considered whether facial movements predicted learning gains. Normalized learning gain was computed using the following formula if posttest score was greater than pretest score:

$$NLG = \frac{Posttest - Pretest}{1 - Pretest}$$

Otherwise, normalized learning gain was computed as follows:

$$NLG = \frac{Posttest - Pretest}{Pretest}$$

Normalized learning gain was predicted by outer brow raising (AU2) intensity and mouth dimpling (AU14) frequency. AU2 was a negative predictor and AU14 was a positive predictor; that is, lower AU2 intensity corresponded to lower learning gain, while greater AU14 frequency corresponded to higher learning gain. The adjusted model effect size was r = 0.43. The model is shown in Table 8.

Table 8. Stepwise linear regression model for Normalized Learning Gain

Norm. Learn Gain =	**Partial R^2**	**Model R^2**	***p***
-2.29 * *AU2_Intensity*	0.145	0.145	<0.001
2.13 * *AU14_Freq*	0.064	0.208	0.031
0.73 (intercept)			0.053
RMSE = 29.49% of range in Normalized Learning Gain			

5. DISCUSSION

The results highlight that specific facial movements predict tutoring outcomes of engagement, frustration, and learning. Particular patterns emerged for almost all of the facial action units analyzed. We discuss each of the results in turn along with the insight they provide into mechanisms of engagement, frustration, and learning as predicted by facial expression.

Average intensity of brow lowering (AU4) was associated with negative outcomes, such as increased frustration and reduced desire to attend future tutoring sessions. Brow lowering (AU4) has been correlated with confusion in prior research [7, 9] and

interpreted as a thoughtful state in other research [12, 18]. Here, the average intensity of brow lowering is found to be a positive predictor of student frustration and a negative predictor of students finding the tutoring session worthwhile. It may be that the tutor and student were unable to overcome student confusion, resulting in frustration instead of deep learning [9]. This interpretation is compatible with the theory of cognitive disequilibrium, which maps possible transitions from confusion to deep learning when a new concept is successfully acquired or to frustration when the concept cannot be reconciled with the student's present understanding. It is also possible that in some cases, AU4 displays represent an angry or agitated affective state. AU4 is a key component of the prototypical display of anger [11]. Further study that accounts for student progress through the programming task may reveal whether there is a significant cognitive aspect to this result.

Average intensity of inner brow raising (AU1) was positively associated with students finding the tutoring session worthwhile. At first glance, this finding seems to be in marked contrast to prior research that implicated both inner and outer brow raising as components of frustration displays [7]. However, intensity of the facial expressions was not considered in the prior work. AU1 is also a component of prototypical expressions of surprise or sadness [11]. From among these possible affective states—frustration, sadness, and surprise—surprise may be most likely to explain higher ratings of endurability. Students may have found the tutoring session to be surprising because it was a first exposure to computer programming. Surprise displays were observed while processing the videos through CERT and there were numerous such displays in the validation corpus. However, further study is required to disambiguate this result.

Lower frequency of outer brow raising (AU2) predicted a lesser sense of being hurried or rushed; in contrast, greater intensity of displays of AU2 predicted reduced learning gains. Outer brow raising (AU2) has been associated with frustration in prior research [7]. As frustrated students may not achieve high learning gains, the intensity of AU2 may be indicative of frustration. However, AU2 was not predictive of students' self-reported frustration levels, so this may be capturing a subtly different phenomenon. An alternative interpretation comes from research into facial expressions of anxiety, in which "fear brow" facial movements were found to occur more often during anxiety [15]. The prototypical "fear brow" includes AU1, AU2, and AU4 present in combination [11]. An example of this facial expression is shown as AU2 in Figure 5. Greater anxiety during tutoring may result in feeling rushed or hurried and may also negatively impact learning. Thus, anxiety is consistent with the results for AU2. However, the other action units expected in facial expressions of anxiety, AU1 and AU4, did not have the same results. This is likely due to the conflicting nature of brow raising and brow lowering, as the CERT values for AU1 and AU4 may be reduced during their combined movement in the "fear brow" (see Figure 5). Further analyses of combined facial movements would provide insight into this complication of automated facial expression recognition.

Frequency of mouth dimpling (AU14) predicted increased student self-reports of task success, as well as increased learning gains. There have not been conclusive associations of mouth dimpling (AU14) and learning-centered emotions. However, this action unit has been implicated as being involved in expressions of frustration [7] and concentration [18]. In this study, frequency of AU14 was positively predictive of both self-reported performance and normalized learning gains. While the effect appears to be fairly subtle (effect size below 0.3 for both), it appears to be a display of concentration. This leads to the interesting question of whether AU4 or AU14 better represents a thoughtful, contemplative state. Further research in this vein may resolve the question.

While eyelid tightening (AU7) was not added to any of the predictive models, there appear to be reasons for this. Observation of CERT processing and the results of the validation analysis indicate a way to adjust CERT's output of AU7, enabling refined study of the action unit. AU7 is an important facial movement to include, as it has been correlated with confusion [7]. Our proposed method for correcting AU7 output was informed by observing that CERT tends to confuse AU7 with blinking or eyelid closing. In prior manual annotation efforts, we explicitly labeled AU7 only when eyelid movements tightened the orbital region of the eye (as in the FACS manual). Thus, manual annotation seems more effective due to this complication of eye movements. However, note that CERT's AU7 output perfectly agreed with manual annotations in our validation analysis. Thus, CERT clearly tracks eyelid movements well. The problem may be that CERT's AU7 output is overly sensitive to other eyelid movements. One way to mitigate this problem may be to subtract other eye-related movements from instances of AU7. For instance, if AU7 is detected, but CERT also recognizes that the eyelids are closed, the detected AU7 event could be discarded.

The results demonstrated predictive value not only for frequency of facial movements, but also intensity. The relationship between facial expression intensity and learning-centered affect is unknown, but perhaps action unit intensity is indicative of higher-arousal internal affective states. Additionally, it is possible that intensity will inform disambiguation between learning-centered affective states that may involve similar action units (e.g., confusion/frustration and anxiety/frustration). Lastly, intensity of facial movements may be able to aid diagnosis of low arousal affective states. For instance, a model of low intensity facial movements may be predictive of boredom, which current facial expression models have difficulty identifying.

6. CONCLUSION

This paper presented an automated facial recognition approach to analyzing student facial movements during tutoring using the Computer Expression Recognition Toolbox (CERT), which tracks a wide array of well-defined facial movements from the Facial Action Coding System (FACS). CERT output was validated by comparing its output values with manual FACS annotations, achieving excellent agreement despite the challenges imposed by naturalistic tutoring video. Predictive models were then built to examine the relationship between intensity and frequency of facial movements and tutoring session outcomes. The predictive models highlighted relationships between facial expression and aspects of engagement, frustration, and learning.

This novel approach of fine-grained, corpus-wide analysis of facial expressions has great potential for educational data mining. The validation analysis confirmed that CERT excels at tracking specific facial movements throughout tutoring sessions. Future studies should examine the phenomena of facial expression and learning in more detail. Temporal characteristics

of facial expression can also be examined, such as how rapidly an expression appears and how quickly it vanishes. Additionally, with these results in hand, it will be important to conduct an analysis of the broader set of facial action units tracked by CERT to build a comprehensive understanding of the interplay between learning and affect.

ACKNOWLEDGMENTS

This work is supported in part by the North Carolina State University Department of Computer Science and the National Science Foundation through Grant DRL-1007962 and the STARS Alliance Grant CNS-1042468. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

REFERENCES

[1] Baker, R.S.J. d., D'Mello, S.K., Rodrigo, M.M.T. and Graesser, A.C. 2010. Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies*. 68, 4, 223–241.

[2] Baker, R.S.J. d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G.W., Ocumpaugh, J. and Rossi, L. 2012. Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *Proceedings of the 5th International Conference on Educational Data Mining*, 126–133.

[3] Baltrusaitis, T., Robinson, P. and Morency, L.-P. 2012. 3D Constrained Local Model for Rigid and Non-Rigid Facial Tracking. *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2610–2617.

[4] Calvo, R.A. and D'Mello, S.K. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing*. 1, 1, 18–37.

[5] Cooper, D.G., Muldner, K., Arroyo, I., Woolf, B.P. and Burleson, W. 2010. Ranking Feature Sets for Emotion Models used in Classroom Based Intelligent Tutoring Systems. *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization*, 135–146.

[6] D'Mello, S.K. and Calvo, R.A. 2011. Significant Accomplishments, New Challenges, and New Perspectives. *New Perspectives on Affect and Learning Technologies*. R.A. Calvo and S.K. D'Mello, eds. Springer. 255–271.

[7] D'Mello, S.K., Craig, S.D. and Graesser, A.C. 2009. Multi-Method Assessment of Affective Experience and Expression during Deep Learning. *International Journal of Learning Technology*. 4, 3/4, 165–187.

[8] 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.

[9] D'Mello, S.K., Lehman, B., Pekrun, R. and Graesser, A.C. Confusion Can Be Beneficial for Learning. *Learning & Instruction*. (in press)

[10] Ekman, P. and Friesen, W. V. 1978. *Facial Action Coding System*. Consulting Psychologists Press.

[11] Ekman, P., Friesen, W. V. and Hager, J.C. 2002. *Facial Action Coding System: Investigator's Guide*. A Human Face.

[12] Grafsgaard, J.F., Boyer, K.E. and Lester, J.C. 2011. Predicting Facial Indicators of Confusion with Hidden Markov Models. *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, 97–106.

[13] Grafsgaard, J.F., Boyer, K.E. and Lester, J.C. 2012. Toward a Machine Learning Framework for Understanding Affective Tutorial Interaction. *Proceedings of the 11th International Conference on Intelligent Tutoring Systems*, 52–58.

[14] Grafsgaard, J.F., Fulton, R.M., Boyer, K.E., Wiebe, E.N. and Lester, J.C. 2012. Multimodal Analysis of the Implicit Affective Channel in Computer-Mediated Textual Communication. *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, 145–152.

[15] Harrigan, J.A. and O'Connell, D.M. 1996. How Do You Look When Feeling Anxious? Facial Displays of Anxiety. *Personality and Individual Differences*. 21, 2, 205–212.

[16] Hart, S.G. and Staveland, L.E. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload*. P.A. Hancock and N. Meshkati, eds. Elsevier Science. 139–183.

[17] Kaliouby, R. and Robinson, P. 2005. Generalization of a Vision-Based Computational Model of Mind-Reading. *Proceedings of the First International Conference on Affective Computing and Intelligent Interfaces*, 582–589.

[18] Littlewort, G., Bartlett, M.S., Salamanca, L.P. and Reilly, J. 2011. Automated Measurement of Children's Facial Expressions during Problem Solving Tasks. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 30–35.

[19] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J. and Bartlett, M. 2011. The Computer Expression Recognition Toolbox (CERT). *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 298–305.

[20] O'Brien, H.L. and Toms, E.G. 2010. The Development and Evaluation of a Survey to Measure User Engagement. *Journal of the American Society for Information Science and Technology*. 61, 1, 50–69.

[21] Picard, R.W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D. and Strohecker, C. 2004. Affective Learning — A Manifesto. *BT Technology Journal*. 22, 4, 253–269.

[22] Rodrigo, M.M.T. and Baker, R.S.J.d. 2011. Comparing Learners' Affect while using an Intelligent Tutor and an Educational Game. *Research and Practice in Technology Enhanced Learning*. 6, 1, 43–66.

[23] Ruf, T., Ernst, A. and Kublbeck, C. 2011. Face Detection with the Sophisticated High-speed Object Recognition Engine (SHORE). *Microelectronic Systems*. 243–252.

[24] den Uyl, M.J. and van Kuilenburg, H. 2008. The FaceReader: Online Facial Expression Recognition. *Proceedings of Measuring Behavior 2005*, 589–590.

[25] Whitehill, J., Bartlett, M. and Movellan, J. 2008. Automatic Facial Expression Recognition for Intelligent Tutoring Systems. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1–6.

[26] Wu, T., Butko, N.J., Ruvolo, P., Whitehill, J., Bartlett, M.S. and Movellan, J.R. 2012. Multi-Layer Architectures for Facial Action Unit Recognition. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, 42, 4, 1027-1038.

[27] Zeng, Z., Pantic, M., Roisman, G.I. and Huang, T.S. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 31, 1, 39–58.