Predicting Learning and Affect from Multimodal Data Streams in Task-Oriented Tutorial Dialogue

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

Learners experience a wide array of cognitive and affective states during tutoring. Detecting and responding to these states is a core problem of adaptive learning environments that aim to foster motivation and increase learning. Recognizing learner affect through nonverbal behavior is particularly challenging, as students display affect across numerous modalities. This study utilizes an automatically extracted set of multimodal nonverbal behaviors and task actions to predict learning and affect in a data set of sixty-three computer-mediated human tutoring sessions. Predictive models of post-session self-reported *engagement*, *frustration*, and *learning* were evaluated with leave-one-out cross-validation. Nonverbal behaviors conditioned on task events and typing were found to be more predictive than incoming student self-efficacy and pretest score. Face and gesture were predictive of *engagement* and *frustration*, while face and posture was predictive of *learning*. The nonverbal model features captured moments when students were most active on the task, such as writing and testing the Java program. These results provide initial evidence linking affect, moment-by-moment multimodal nonverbal behavior, and task performance during tutoring. They improve understanding of learner affect and enable automated tutorial interventions that adapt to student states as a highly effective human tutor would.

Keywords

Affect, engagement, frustration, facial expression recognition, gesture, posture, multimodal nonverbal behavior, computer-mediated tutoring

1. INTRODUCTION

Mastery-oriented one-to-one human tutoring may provide two sigma learning gains [3]. In order to match this high bar of expert human tutoring effectiveness, automated tutorial interventions may need to be designed with both learner knowledge and motivation in mind [5, 9, 12, 31]. Highly effective human tutors simultaneously address cognitive and affective states of learners, adapting to the appropriate level of content difficulty and improving learner motivation through personalized instruction [25]. Just as human tutors consider more than task performance of the student, it may be necessary to bolster automated tutorial interventions with additional information regarding learner affect from nonverbal behavior [32].

Early studies of nonverbal behavior in tutoring relied on manual observations of affect and nonverbal behavior [2, 8, 14, 38]. More recently, automated techniques have been leveraged to track nonverbal behaviors [1, 11, 15, 17, 22]. Most studies have examined individual modalities in detail, such as facial expression [17, 35], posture [10, 18], or gesture [18, 28]. However, a much smaller set of studies has examined multiple modalities of nonverbal behavior [1, 11, 22]. It is likely that a multimodal combination of automatically tracked affective data streams would need to be considered to best adapt to learner affect during tutoring [9].

Facial expression is a particularly informative modality for analysis of affect, as indicated by decades of prior research. Many studies have utilized the Facial Action Coding System (FACS), a coding manual that describes the fine-grained movements of the human face as facial action units (AUs) [13]. Recent automated techniques have enabled FACS-based facial expression recognition. In particular, the Computer Expression Recognition Toolbox (CERT), used in this study, provides a state-of-the-art facial expression recognition tool that identifies facial action units [27]. CERT was trained using databases of spontaneous and posed expressions and has been validated on naturalistic video datasets [16, 26, 27]. Thus, frame-by-frame facial action unit tracking provides detailed affective information that is readily synchronized with additional modalities.

Gestures have been tangentially reported on in the intelligent tutoring systems community, but other phenomena were the primary focus of those studies [14, 38]. Recent work has begun to describe and track cognitive-affective gestures [15, 18, 21, 28]. This study uses an algorithm that processes three-dimensional Kinect™ depth images to identify when one or two hands contact the lower face [15].

Posture has been used as an affective feature in multiple systems, but interpretation of postural movements is very complex [10, 22, 23]. One result replicated across multiple studies is that increases in postural movement are linked with negative affect or disengagement [10, 15, 33, 38]. Early work used expensive pressure-sensitive chairs [22, 38]. Newer techniques rely on computer vision to interpret posture from video [10, 15, 33]. This study uses an algorithm that processes Kinect depth images to identify how far away the student is seated [15].

The analysis reported in this paper combines an automatically extracted set of multimodal nonverbal behaviors and task actions to predict learning and affect in a data set of sixty-three computer-mediated human tutoring sessions. Relative frequencies of nonverbal behaviors contingent on task events and typing statuses were used as predictive features. Model averaging identified the top twenty predictive features per model. Three models were built using stepwise forward linear regression with the Bayesian Information Criterion (BIC) to predict retrospective self-reports of *engagement* and *frustration*, as well as normalized learning gains. The models were evaluated with leave-one-out cross-validation. Nonverbal features were found to be more predictive than incoming student self-efficacy and pretest scores. Face and gesture were predictive of *engagement* and *frustration*, while face and posture were predictive of *learning*. Additionally, the majority of nonverbal predictive features occurred when the student was writing and testing the Java program, which shows that these moments may be most salient to affect. Further studies in this vein can inform the design of automated tutorial interventions in order to adapt to student affect as a highly effective human tutor would.

2. RELATED WORK

Few studies have examined multimodal nonverbal behavior features in a tutoring context. An initial study by Kapoor and Picard considered prediction of experienced teacher judgments of affect in young student (8-11 years of age) interactions with a game, Fripple Place [22]. Face, posture, and task features were used in a mixture of Gaussian processes. These models performed well at predicting teacher judgments of affect, which was an important initial step toward detecting cognitive-affective states involved in cognitively demanding tasks.

In research on the AutoTutor intelligent tutoring system, multimodal features were used to predict affect labels by expert judges [11]. Emotion labels were manually selected using six affective states (*boredom*, *confusion*, *engagement/flow*, *frustration*, *delight*, *surprise*) and a non-emotional/neutral choice at fixed time intervals and spontaneously across thirty-eight approximately half-hour tutoring sessions. These labels were then predicted using a multimodal feature set including manually annotated Facial Action Coding System facial movements, automatically extracted dialogue features from fifteen seconds prior to an emotion label, and automatically extracted posture features using a pressure-sensitive chair. The fully-featured models of face, dialogue, and posture produced the best levels of agreement, with Cohen's κ of 0.33 for fixed emotion judgments and 0.39 for spontaneous ones.

Another line of research has investigated the use of multiple sensor technologies with the Wayang Outpost intelligent tutoring system [1]. A real-time facial expression analysis tool trained on posed cognitive-affective displays, MindReader, was used to estimate levels of *agreeing*, *concentration*, *interest*, *thinking*, and *unsureness*. Additionally, a pressure-sensitive mouse, skin conductance bracelet, and pressure-sensitive chair were also used. Student cognitive-affective self-reports were given during the tutoring session for states of *confidence*, *excitement*, *frustration*, and *interest*. Stepwise regression models were constructed across combinations of modalities (including tutorial context). The results found that best fit models were achieved through combinations of facial expression and tutoring context (for *confidence*, *excitement*, and *interest*) and posture and tutoring context (for *frustration*). The corresponding model effect sizes for the best fit models ranged from r = 0.54 to 0.83. A follow-up validation study was also conducted with a new set of students from a different school and a lower age group [7]. The results found that the previously used features were only partially generalizable to the validation population, with reduced accuracies for most features. This underscores the necessity of identifying generalizable affective features.

In contrast with prior studies, this paper presents models predicting affective and learning outcomes from moment-to-moment nonverbal behavior and task performance. This line of investigation seeks to identify nonverbal behavioral correlates of both affect and learning. The present results indicate that facial expression, gesture, and posture may have differing affective interpretations based on the tutoring context in which they occur. The nonverbal features were found to be more predictive than incoming student self-efficacy and pretest score. Additionally, the nonverbal features were largely contingent upon student work on the programming task, illustrating that these moments of student task activity may be most salient to affect. Further studies in this vein may produce affect recognition that enables detecting and responding to learner affect as a highly effective human tutor would.

3. TUTORING STUDY

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 database logs, webcam video, and Kinect depth video from the first lesson as a multimodal tutoring corpus, described further in Section 4. The JAVATUTOR interface is shown in Figure 1.

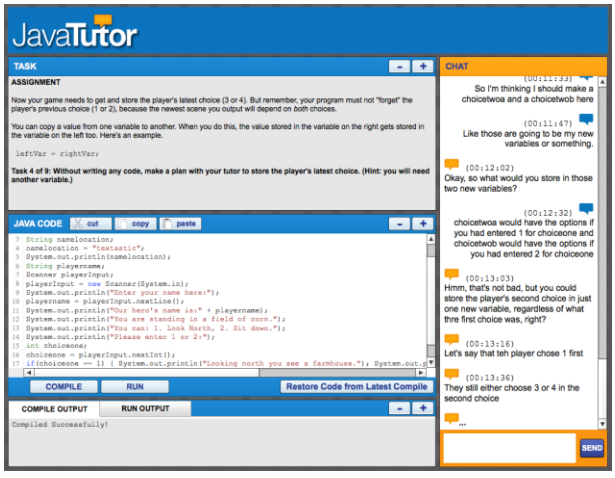

Figure 1. The JAVATUTOR interface

Figure 2. Facial action units recognized by CERT (left to right): AU1 (Inner Brow Raiser) & AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU7 (Lid Tightener), AU14 (Mouth Dimpler)

On a day prior to the first tutoring session, students completed a set of surveys to measure incoming student characteristics. Two of these pre-session survey instruments are analyzed in this paper: computer science domain-specific self-efficacy and general self-efficacy. The computer science self-efficacy measure is comprised of the confidence items from the Computer Science Attitude Survey [37]. General self-efficacy was measured using the New General Self-Efficacy instrument [6]. 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 included the User Engagement Survey (UES) [30] and the NASA-TLX workload survey [20], which included an item for Frustration Level. There is a recent validation of the UES measure with further information [36].

4. MULTIMODAL TUTORING CORPUS

The tutoring session database logs were combined with automated facial action unit tracking on webcam videos and gesture and posture tracking across Kinect depth image frames. The automated tracking techniques are described in the following subsections. The resulting multimodal features are described in Section 4.3.

4.1 Facial Expression Recognition

A state-of-the-art facial expression recognition tool, the Computer Expression Recognition Toolbox (CERT) [19], was used for 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 [25]. 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).

We previously validated an adjustment to CERT output that produced excellent aggregate agreement with manual FACS annotations across a subset of five action units [16]. The adjustment involves subtraction of the average value for each facial action unit as a baseline in order to reduce systematic tracking error. While any positive output value indicates that CERT recognizes an action unit, we empirically found that a higher threshold may reduce false positives. Thus, we consider an action unit to be present when the baseline-adjusted CERT output is at least 0.25. Examples of the five selected facial action units and their FACS codes (e.g., AU1) are shown in Figure 2.

4.2 Gesture and Posture Detection

Previously developed posture and gesture tracking techniques were applied to the recorded Kinect depth images. The posture tracking algorithm was previously evaluated to be 92.4% accurate, while gesture tracking was found to be 92.6% accurate [15]. The tracking algorithms were run on all sessions, but four sessions had no Kinect recordings due to human error (N=63). Examples of one-hand-to-face and two-hands-to-face gestures are shown in Figure 3.

Figure 3. Examples of hand-to-face gestures

The median head distance of students at each workstation was selected as the "mid" postural position. Distances at one standard deviation (or more) closer or farther than "center" were labeled as "near" or "far," respectively. Additionally, postural movements were identified based on acceleration of the head tracking point. The absolute sum of frame-to-frame acceleration was accumulated in a rolling one-second window at each frame. The average amount of acceleration in a one-second interval was computed across all students. If acceleration in the present interval was above average, it was marked as a postural movement (PosMove). Average frequencies of gesture and posture features are shown in Table 1. Students tended to spend more time in a Mid postural position and most frequently did not display a hand-to-face gesture. Additionally, students moved less than average during each interval, indicating that there were short moments of high movement that raised the average.

Table 1. Average frequency of gesture and posture features

Feature	Avg. Freq.	Feature	Avg. Freq.
NEAR	15%	ONEHAND	16%
MID	68%	TWOHANDS	5%
FAR	17%	NOGESTURE	79%
POSMOVE	29%		
NOMOVE	71%		

4.3 Multimodal Features

The automatically recognized nonverbal behaviors were combined with task-related features in order to form the multimodal tutoring corpus. As students worked on programming tasks, the database logged dialogue messages, typing, and task progress. Tutorial dialogue occurred at any time during the sessions, with student and tutor messages sent asynchronously (STUDENTMSG and TUTORMSG, respectively). As a student completed the programming task, he or she would also press a compile button to convert the Java program code into a format that is ready to run. These compile attempts may be successful (COMPILESUCCESS) or fail due to an error in the program code (COMPILEERROR). The student would also run his or her program (RUNPROGRAM) in order to test the output and interact with it. In parallel with the task events described above, the database logged whether the student was typing at any given moment. The student may not be typing anything (NOTTYPING), working on the program code (CODING), or typing a message to the tutor (TYPINGMSG) at each moment. Additionally, the student was considered to have paused on the task if he or she had made changes to the program and then stopped. This sort of break may be due to the student having resolved the current task, taking a moment to think, or going off-task; therefore, it was introduced as a task event (TASKPAUSE). The average frequency of each task event and typing status is shown in Table 2. The majority of time intervals occurred after tutor messages and when students were not typing. These majority events represent moments when the student may have been reading the task description or reflecting on tutor messages. Tutors were also more active in the dialogue than students, resulting in more time following tutor messages.

Table 2. Average frequency of task events and typing status

Task Event	Avg. Freq.	Typing Status	Avg. Freq.
COMPILEERROR	1.7%	CODING	15%
COMPILESUCCESS	2.1%	TYPINGMSG	12%
RUNPROGRAM	7.9%	NOTTYPING	73%
STUDENTMSG	26.4%		
TUTORMSG	53.1%		
TASKPAUSE	8.8%		

Task events and typing statuses were combined with nonverbal behaviors at one-second intervals across each tutoring session. The most recent event of a given type (nonverbal, task, typing) was counted as the current value at each interval. For instance, if a student had been typing but stopped after half a second into the current interval, the typing status would be assigned to NOTTYPING.

A tutoring session excerpt is shown in Figure 4. The excerpt shows a rich set of nonverbal behaviors occurring around student work on the programming task. This student produced a variety of facial expressions, particularly when examining and testing the Java program. Additionally, the student performed a one-hand-to-face gesture prior to compiling the program. The corresponding multimodal features for a segment of the excerpt are shown in Figure 5 (top of next page). The multimodal feature vectors cover a twelve-second segment from the excerpt.

26:54	Tutor:	ready?
26:59	Student:	yes! [*Student starts coding*]
28:02	Student:	TASKPAUSE [*Student stops coding*]
28:03	Student:	GESTURE: ONEHANDTOFACE; FACE: AU2 & AU14
28:12	Student:	TASK: COMPILESUCCESS; FACE: AU2
28:14	Student:	FACE: AU14
28:17	Student:	TASK: RUNPROGRAM; FACE: AU1
28:19	Student:	FACE: AU7
28:21	Tutor:	excellent

Figure 4. Tutoring session excerpt

Relative frequencies of nonverbal behavior were calculated separately for task events and typing status. For instance, at each one-second time interval, AU1 was marked as present or absent. Each interval was associated with a task event, with frequency counts tabulated across all task events. The relative frequency of AU1 presence and absence was computed across these task-contingent counts. Thus, the percentages of time intervals occurring with specific task events and particular values of AU1 presence or absence sum to one hundred percent. For instance, one student may have AU1 after RUNPROGRAM 2.12% of the time and NOAU1 after RUNPROGRAM 3.24% of the time. These relative frequencies sum to one hundred percent when combined with the remainder of task-contingent relative frequencies of AU1. Relative frequencies were similarly computed across typing statuses for each nonverbal behavior. Thus, the relative frequencies account for the percent of time in which a student displayed a nonverbal behavior after a specific task event or during a particular typing status (i.e., a student with a 5% relative frequency of ONEHAND after TUTORMSG in a thirty minute session would have displayed a one-hand-to-face gesture for a total of ninety seconds after tutor messages). This resulted in a set of one hundred and sixty-two nonverbal features contingent upon task events and typing statuses. The distribution of these multimodal features across nonverbal behaviors, task events, and typing statuses is shown in Table 3.

Table 3. Counts of multimodal features across nonverbal behaviors, task events, and typing statuses

	Task Event	Typing Status
AU1	12	6
AU2	12	6
AU4	12	6
AU7	12	6
AU14	12	6
GESTURE	18	9
POSTURE	18	9
POSMOVE	12	6

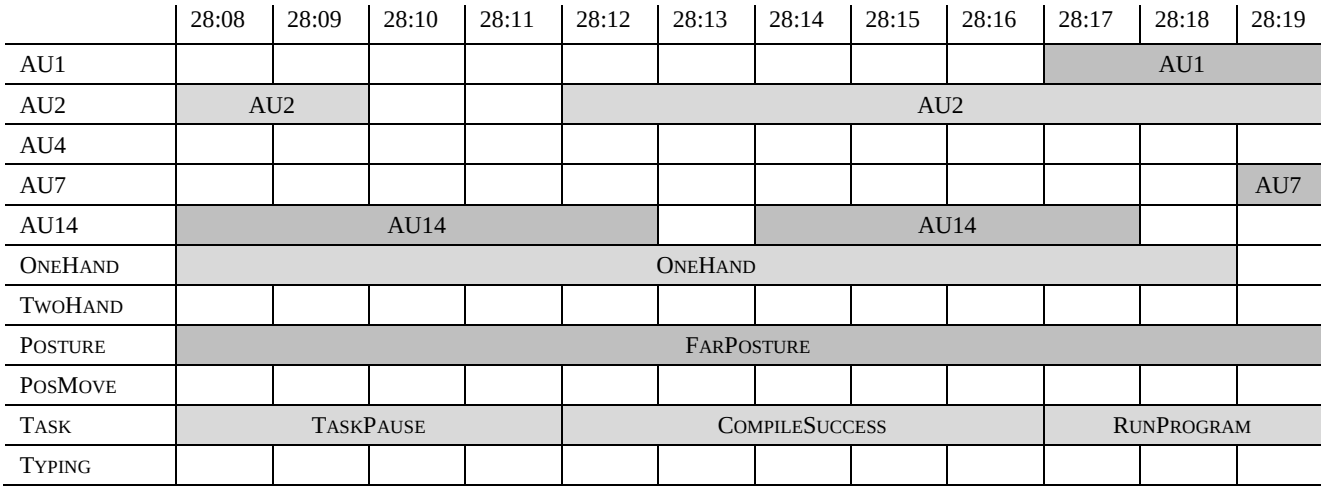

Figure 5. Multimodal feature vectors for a twelve-second segment of tutoring: gray shading indicates presence of a nonverbal behavior, task event, or typing. Time is shown in minutes and seconds from the beginning of the tutoring session.

5. PREDICTIVE MODELS

Fine-grained analyses of multimodal affective expressions are enabled by automated tracking of nonverbal behavior. Such analyses have the potential to reveal previously undiscovered ways in which affective displays relate to task performance, learning, and affective outcomes within a tutoring context. For instance, the same affective expression may have different causes depending on the tutoring context. As a first step toward examining the fine-grained tutoring context of learner affective displays, predictive models of affective and learning outcomes were constructed using the multimodal tutoring corpus, in which facial expression, gesture, and posture are combined with task actions.

Initial feature selection was performed using model averaging in JMP statistical software, which created regression models for all possible combinations of predictive variables [34]. Model averaging is used to identify and remove weakly predictive variables across all models. Specifically, the twenty most predictive variables were selected using the average coefficient estimate from models with one, two, or three predictive variables. The predictive models were then constructed using minimum Bayesian Information Criterion (BIC) in forward stepwise linear regression. 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. Thus, model averaging was used to identify the most generally predictive variables, while minimum BIC was used to constrain model complexity. Tutoring outcomes (*engagement*, *frustration* and *learning*) were the dependent variables. All variables were standardized (i.e., centered on the mean and scaled to unit standard deviation) to enable comparison. The predictive models shown in the following sub-sections have been constructed using the entire corpus, with associated regression coefficients and R^2 values. Additionally, leave-one-out cross-validated R^2 values were computed using the same predictive variables (but different coefficients in each fold) to examine generalizability of the predictive models.

5.1 Predicting Engagement

Each student's Engagement score was the sum of the Focused Attention, Felt Involvement, and Endurability sub-scales in the User Engagement Survey [30] administered following the tutoring session. This model only uses self-reports of engagement from students who fully completed the User Engagement Survey (N=61). The predictive model of *engagement* was composed of three features, including students' incoming computer science self-efficacy, one-hand-to-face gestures after successful compile, and brow lowering (AU4) after sending a student dialogue message. Each of the nonverbal features explains more variance than the trait-based feature of computer science self-efficacy. This seems to indicate that state-based nonverbal features are more indicative of *engagement*. The cross-validated model effect size was $r = 0.39$. The model is shown in Table 4.

Table 4. Stepwise linear regression model for Engagement

Engagement =	**Partial R^2**	**Model R^2**	***p***
0.31 * OneHand after CompileSuccess	0.10	0.10	0.009
-0.31 * AU4 after StudentMsg	0.09	0.19	0.008
0.27 * Computer Science Self-Efficacy	0.07	0.26	0.020
~0 (intercept)			0.959
RMSE = 0.88 standard deviations in Engagement			
Leave-One-Out Cross-Validated R^2 = 0.15			

5.2 Predicting Frustration

The Frustration Level scale from NASA-TLX [20] was the student's retrospective self-report of how insecure, agitated or upset he or she was during the tutoring session. The predictive model of *frustration* included students' incoming general self-efficacy and two features that accounted for the absence of nonverbal behavior. The sole feature predictive of higher *frustration* corresponded with compile errors, which intuitively may be frustrating. The absence of brow lowering (AU4) after running the Java program reinforces a prior result that indicated AU4 as a marker of *frustration* [17]. Also, students with higher general self-efficacy tended to have less *frustration*, as represented in the model. The cross-validated model effect size was $r = 0.41$. The model is shown in Table 5.

Table 5. Stepwise linear regression model for Frustration

Frustration =	Partial R^2	Model R^2	*p*
-0.42 * General Self-Efficacy	0.14	0.14	0.004
-0.56 * NoAU4 after RunProgram	0.08	0.22	0.004
0.42 * NoGesture after CompileError	0.08	0.30	0.011
~0 (intercept)			1.000
RMSE = 0.85 standard deviation in Frustration Level			
Leave-One-Out Cross-Validated R^2 = 0.17			

5.3 Predicting Learning Gain

Normalized learning gain measures how much a student learned relative to what he or she could have learned [29]. This accounts for relative differences in learning between students who scored high or low on the pretest. 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:

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

The predictive model of normalized learning gain is the only one of the three to include postural features. These features indicate that MID and FAR postural positions are predictive of learning, though whether they are positive or negative predictors is dependent upon the tutoring context. Mouth dimpling (AU14) after running the Java program was predictive of learning. This supports a prior result that AU14 is positively associated with learning [17]. Finally, general self-efficacy predicted higher learning gains. The cross-validated model effect size was r = 0.62. The model is shown in Table 6.

Table 6. Stepwise linear regression model for Normalized Learning Gain

Norm. Learning Gain =	Partial R^2	Model R^2	*p*
0.10 * AU14 after RunProgram	0.11	0.11	0.004
0.10 * General Self-Efficacy	0.08	0.19	0.002
-0.12 * MidPosture after CompileError	0.08	0.27	<0.001
-0.21 * FarPosture during Coding	0.04	0.31	<0.001
0.20 * FarPosture after CompileSuccess	0.18	0.49	<0.001
0.43 (intercept)			<0.001
RMSE = 0.24 std. dev. in Normalized Learning Gain			
Leave-One-Out Cross-Validated R^2 = 0.38			

6. DISCUSSION

The results demonstrate that nonverbal behaviors at specific moments in the tutoring session are predictive of *engagement*, *frustration*, and *learning*. The combination of task events, typing, and nonverbal behaviors in multimodal features is predictive beyond incoming student characteristics, such as pretest score and self-efficacy. Additionally, the affective valence (positive or negative) of the nonverbal behaviors depended upon the tutoring context in which they occurred.

The predictive model of *engagement* was composed of three features, including students' incoming computer science self-efficacy, one-hand-to-face gestures after successful compile, and brow lowering (AU4) after sending a student dialogue message. One-hand-to-face gestures may have different affective valence depending on the physical position of the hand. A student may rest his/her head on the hand as a sign of boredom [2], or touch his/her chin in a moment of contemplation [28]. Here, one-hand-to-face gestures after compile success were predictive of higher post-session self-report of *engagement*. This may coincide with student focus on the programming task. In the moments after updating the program code and compiling it, the student is no longer typing and may then reflect on current progress. Brow lowering (AU4) after the student sends a dialogue message, on the other hand, was a predictor of lower *engagement*. This may indicate that a student is having difficulty with the subject matter, most likely responding to a tutor message (in this corpus, tutor messages were predominant and students rarely took initiative in the dialogue). Both of the nonverbal features were more predictive than computer science domain-specific self-efficacy, which was associated with greater *engagement*.

Frustration was significantly predicted by general self-efficacy. Higher levels of general self-efficacy coincided with lower post-session reports of *frustration*. Students with higher general self-efficacy are more confident in their ability to complete difficult tasks and therefore may be less intimidated by a novel learning task. However, inclusion of two nonverbal features doubled the explanatory power of the model. Each of the nonverbal features captured absence of nonverbal behaviors after specific task events. Absence of brow lowering (AU4) after running the Java program was predictive of lower *frustration*. At this point, the student is testing the program to see whether it matches his/her expectation. A prior result on this tutoring corpus found that AU4 was an indicator of *frustration*. Therefore, the present result supports that finding, but also provides a specific tutoring context (running the program) that is particularly meaningful for *frustration*. The sole feature predictive of higher *frustration* corresponded with compile errors (which occur when the program is incorrect). This correspondence between compiling the program and frustration is similar to results of prior analyses of student emotions during computer programming [4, 24]. Not all students had compile errors, so this feature represents those students who may have found the task to be more difficult. The absence of gestures after compile errors may be due to swift tutor interventions to remediate problems with the program. In this case, a student may feel frustrated due to overly active tutoring strategies.

Normalized learning gain was predicted by a combination of students' incoming general self-efficacy, mouth dimpling (AU14) after running the program, and three posture-related features. The model shows that students with more general confidence in their ability to complete novel and difficult tasks

tended to learn more than their peers. Displays of AU14 after running the program also were predictive of higher learning gain. Two aspects of AU14 discovered in prior results may shed light on this. First, occurrence of AU14 in general was associated with greater learning gain [16]. Second, AU14 in the first five minutes of tutoring was correlated with higher *frustration*, while AU14 in the last five minutes of tutoring was correlated with greater learning gain [17]. Running the program occurs most frequently during the later portion of the session. So, AU14 displays after running the program may also occur toward the end. With this timing-related interpretation, it may be that continued mental effort throughout the tutoring session is reflected in displays of AU14. Further study of AU14 may confirm whether it is generally an indicator of mental effort.

The posture-related features included both MID and FAR distances. These postural positions may encode information beyond whether a student is sitting at a certain distance from the computer. For instance, when a student is sitting at MID distance, the shoulders may be hunched or the student may have a straight back. FAR postural position was both predictive of higher learning gain (when occurring after compile success) and lower learning gain (when present during coding). It may be that bored students slouched in a FAR position during coding, while relaxed (but active) students were similarly farther back. New tracking methods may be developed to disambiguate these subtleties of posture. Interestingly, postural position was predictive of learning, but moment-to-moment postural movement was not. Discretization across one-second intervals may not have adequately captured brief postural movements.

The predictive models largely include nonverbal features that occur around moments of student work on the programming task. These may be pivotal moments on a student's path to learning, as students are actively working on the task and confirming whether the program works as intended. Prior results in analysis of skin conductance on this tutoring corpus showed that students' physiological responses to compile attempts and failures were associated with *learning* and *frustration* [19]. The predictive models presented in this paper further underscore the importance of tutoring context in interpretation of nonverbal behavior.

7. CONCLUSION

This paper presented a multimodal analysis of automatically recognized nonverbal behaviors and task events. State-of-the-art facial expression recognition was leveraged, along with depth video-based gesture detection and posture tracking algorithms, in order to automatically annotate nonverbal behaviors across a corpus of sixty-three tutoring sessions. Multimodal feature vectors were constructed at one-second intervals, including facial expression, gesture, posture, the most recent task event, and whether the student was typing. These features were then used to predict post-session *engagement*, *frustration*, and *learning* outcomes. The results show that multimodal nonverbal behavior features are predictive of affect and learning beyond student incoming characteristics, such as self-efficacy and pretest scores.

These results are a first step toward understanding the relationship between affect, moment-by-moment nonverbal behavior, and task performance during tutoring. The multimodal data streams included nonverbal behavior (facial expression, gesture, posture) and task logs (discrete task events, typing status) across time intervals. This approach provides a basis for triangulating learner affect from multimodal time sequence data. The fine-grained data collected on task performance and nonverbal behavior provides an estimation of learners' underlying real-time cognitive and affective processes.

Further research may identify how facial expressions co-occur and provide further validation of fine-grained tracking of facial movements. Additionally, spatiotemporal features of gesture and posture have only just begun to be explored. Future work may disambiguate between different types of one-hand-to-face and two-hands-to-face gesture, as well as tracking more detailed postural information, such as slouching and leaning. Human tutors innately employ knowledge of nonverbal behavior, thus research in this vein brings the capabilities of automated tutorial intervention closer to those of human tutors. This line of investigation informs our understanding of learner affect and enables affective interventions that intelligently model nonverbal behavior and task actions, as a highly effective human tutor would.

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 Grant CNS-1042468 (STARS Alliance). 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] Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K. and Christopherson, R.M. 2009. Emotion Sensors Go To School. *14th International Conference on Artificial Intelligence in Education* (2009), 17–24.

[2] 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 (Apr. 2010), 223–241.

[3] Bloom, B.S. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*. 13, 6 (1984), pp. 4–16.

[4] Bosch, N., D'Mello, S.K. and Mills, C. 2013. What Emotions Do Novices Experience during Their First Computer Programming Learning Session? *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (2013), 11–20.

[5] du Boulay, B., Avramides, K., Luckin, R., Martinez-Miron, E., Mendez, G.R. and Carr, A. 2010. Towards Systems That Care: A Conceptual Framework based on Motivation, Metacognition and Affect. *International Journal of Artificial Intelligence in Education*. 20, 3 (2010).

[6] Chen, G., Gully, S.M. and Eden, D. 2001. Validation of a New General Self-Efficacy Scale. *Organizational Research Methods*. 4, 1 (2001), 62–83.

[7] 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* (2010), 135–146.

[8] Craig, S.D., D'Mello, S.K., Witherspoon, A. and Graesser, A.C. 2008. Emote Aloud during Learning with AutoTutor: Applying

the Facial Action Coding System to Cognitive-Affective States during Learning. *Cognition & Emotion*. 22, 5 (2008), 777–788.

[9] 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.

[10] D'Mello, S.K., Dale, R. and Graesser, A.C. 2012. Disequilibrium in the Mind, Disharmony in the Body. *Cognition & Emotion*. 26, 2 (2012), 362–374.

[11] D'Mello, S.K. and Graesser, A.C. 2010. Multimodal Semi-automated Affect Detection From Conversational Cues, Gross Body Language, and Facial Features. *User Modeling and User-Adapted Interaction*. 20, 2 (May 2010), 147–187.

[12] D'Mello, S.K., Lehman, B., Pekrun, R. and Graesser, A.C. 2012. Confusion Can Be Beneficial for Learning. *Learning & Instruction*. (2012).

[13] Ekman, P., Friesen, W. V. and Hager, J.C. 2002. *Facial Action Coding System*. A Human Face.

[14] Grafsgaard, J.F., Boyer, K.E., Phillips, R. and Lester, J.C. 2011. Modeling Confusion: Facial Expression, Task, and Discourse in Task-Oriented Tutorial Dialogue. *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (2011), 98–105.

[15] 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* (2012), 145–152.

[16] Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N. and Lester, J.C. 2013. Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. *Proceedings of the 6th International Conference on Educational Data Mining* (Memphis, Tennessee, 2013).

[17] Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N. and Lester, J.C. 2013. Automatically Recognizing Facial Indicators of Frustration: A Learning-Centric Analysis. *Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction* (2013), 159–165.

[18] Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N. and Lester, J.C. 2013. Embodied Affect in Tutorial Dialogue: Student Gesture and Posture. *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (Memphis, Tennessee, 2013).

[19] Hardy, M., Wiebe, E.N., Grafsgaard, J.F., Boyer, K.E. and Lester, J.C. 2013. Physiological Responses to Events During Training: Use of Skin Conductance to Inform Future Adaptive Learning Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2013), 2101–2105.

[20] 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.

[21] El Kaliouby, R. and Robinson, P. 2005. The Emotional Hearing Aid: an Assistive Tool for Children with Asperger Syndrome. *Universal Access in the Information Society*. 4, 2 (Aug. 2005), 121–134.

[22] Kapoor, A. and Picard, R.W. 2005. Multimodal Affect Recognition in Learning Environments. *Proceedings of the 13th Annual ACM International Conference on Multimedia* (2005), 677–682.

[23] Kleinsmith, A. and Bianchi-Berthouze, N. 2012. Affective Body Expression Perception and Recognition: A Survey. *IEEE Transactions on Affective Computing*. (2012).

[24] Lee, D.M., Rodrigo, M.M.T., Baker, R.S.J. d., Sugay, J. and Coronel, A. 2011. Exploring the Relationship Between Novice Programmer Confusion and Achievement. *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction* (2011), 175–184.

[25] Lepper, M.R. and Woolverton, M. 2002. The Wisdom of Practice: Lessons Learned from the Study of Highly Effective Tutors. *Improving Academic Achievement*. J. Aronson, ed. Elsevier. 135–158.

[26] 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* (2011), 30–35.

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

[28] Mahmoud, M. and Robinson, P. 2011. Interpreting Hand-Over-Face Gestures. *Proceedings of the International Conference on Affective Computing and Intelligent Interaction* (2011), 248–255.

[29] Marx, J.D. and Cummings, K. 2007. Normalized Change. *American Journal of Physics*. 75, 1 (2007), 87–91.

[30] 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 (2010), 50–69.

[31] Pekrun, R. 2006. The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review*. 18, 4 (2006), 315–341.

[32] 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 (Oct. 2004), 253–269.

[33] Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P.W. and Paiva, A. 2011. Automatic Analysis of Affective Postures and Body Motion to Detect Engagement with a Game Companion. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (2011), 305–311.

[34] Symonds, M.R.E. and Moussalli, A. 2010. A Brief Guide to Model Selection, Multimodel Inference and Model Averaging in Behavioural Ecology using Akaike's Information Criterion. *Behavioral Ecology and Sociobiology*. 65, 1 (Aug. 2010), 13–21.

[35] Whitehill, J., Serpell, Z., Foster, A., Lin, Y.-C., Pearson, B., Bartlett, M.S. and Movellan, J.R. 2011. Towards an Optimal Affect-Sensitive Instructional System of Cognitive Skills. *Proceedings of the Computer Vision and Pattern Recognition Workshop on Human Communicative Behavior* (Jun. 2011), 20–25.

[36] Wiebe, E.N., Lamb, A., Hardy, M. and Sharek, D. 2014. Measuring Engagement in Video Game-based Environments: Investigation of the User Engagement Scale. *Computers in Human Behavior*. 32, (Mar. 2014), 123–132.

[37] Wiebe, E.N., Williams, L., Yang, K. and Miller, C. 2003. Computer Science Attitude Survey. *North Carolina State University Technical Report TR-2003-1*. (2003).

[38] Woolf, B.P., Burleson, W., Arroyo, I., Dragon, T., Cooper, D.G. and Picard, R.W. 2009. Affect-Aware Tutors: Recognising and Responding to Student Affect. *International Journal of Learning Technology*. 4, 3-4 (2009), 129–164.