The Additive Value of Multimodal Features for Predicting Engagement, Frustration, and Learning during Tutoring

Joseph F. Grafsgaard[1], Joseph B. Wiggins[1], Alexandria K. Vail[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, akvail, keboyer, wiebe, lester}@ncsu.edu

ABSTRACT
Detecting learning-centered affective states is difficult, yet crucial for adapting most effectively to users. Within tutoring in particular, the combined context of student task actions and tutorial dialogue shape the student's affective experience. As we move toward detecting affect, we may also supplement the task and dialogue streams with rich sensor data. In a study of introductory computer programming tutoring, human tutors communicated with students through a text-based interface. Automated approaches were leveraged to annotate dialogue, task actions, facial movements, postural positions, and hand-to-face gestures. These dialogue, nonverbal behavior, and task action input streams were then used to predict retrospective student self-reports of engagement and frustration, as well as pretest/posttest learning gains. The results show that the combined set of multimodal features is most predictive, indicating an additive effect. Additionally, the findings demonstrate that the role of nonverbal behavior may depend on the dialogue and task context in which it occurs. This line of research identifies contextual and behavioral cues that may be leveraged in future adaptive multimodal systems.

Categories and Subject Descriptors
H.1.2 [**User/Machine Systems**]: Human factors; H.5.m [**Information Interfaces and Presentation (e.g., HCI)**]: Miscellaneous

General Terms
Human Factors

Keywords
Affect; multimodal; engagement; frustration; facial expression; gesture; posture; tutorial dialogue

ICMI'14, November 12 - 16, 2014, Istanbul, Turkey

http://dx.doi.org/10.1145/2663204.2663264

1. INTRODUCTION
Combining multiple modalities to predict affect has been found to produce a "consistent, but modest" effect [4]. A majority of such studies aimed to predict basic emotions, which have been the focus of much research over the past several decades. However, recent attempts to apply affective computing techniques to naturalistic domains have considered a broader range of affective states, such as anxiety, boredom, confusion, excitement, and interest [1–3, 10, 12, 13, 18, 24]. With this renewed focus on applying multimodal techniques to affect recognition and understanding, it is important to investigate the relative benefit of multimodal feature sets across non-basic affective states.

Prior studies have examined learning-centered affective states through self-reports and observer judgments [1–3, 13]. Multimodal features such as dialogue, facial expression, posture, and task actions were used to predict momentary affective states, such as boredom, confusion, excitement, and frustration. While these studies have provided insight into moments of judged and self-reported affect, they were not designed to identify how these moments contribute to an overall sense of whether a student is engaged, frustrated, or learning. For instance, if a student is frustrated throughout the majority of a tutoring session, this affective state may be accompanied by cohesive moments of dialogue, nonverbal behavior, and task activity.

A central problem in multimodal interaction for tutoring lies in understanding how observed behavior is associated with persistent affective states. This may be addressed by an approach that analyzes tutorial modalities (e.g., dialogue, nonverbal behavior, task actions) with students' retrospective self-reports of affect. Such an approach is already used in investigations of affect in clinical psychology and inspired the work reported here. In studies of anxiety and depression, nonverbal behavior is examined over a period of time to identify differences in behavior due to psychological conditions [10, 12, 18]. Similarly, learning-centered affective states that occur throughout a tutoring session may coincide with differences in observable behavior.

To address this problem, this paper reports on the first study to investigate how multimodal feature sets can be used to predict whole-session retrospective self-reports of affect and learning gain within human-human tutoring. Multimodal feature sets were constructed from input streams of dialogue, nonverbal behavior, and task actions in computer-mediated one-on-one tutoring. The nonverbal behavior input stream included automatically tracked facial expression, hand-to-face gestures, and posture. Unimodal, bimodal, and trimodal feature sets were

used to predict retrospective self-reports of engagement and frustration during the tutoring session and learning gain. The complete trimodal feature set was most predictive of each of the three tutoring outcomes, and bimodal features with dialogue were most predictive of each tutoring outcome. Importantly, the findings demonstrate that the role of nonverbal behavior may depend on the dialogue and task context in which it occurs. These results provide a promising direction for investigating multimodal feature sets in affective tutorial interaction. Future adaptive multimodal interfaces may leverage such detailed task-contextualized features to disambiguate affective behavior and improve user outcomes.

2. RELATED WORK

Several studies have examined multimodal features in learning-centered affect. Two studies have examined multimodal features in order to predict judged affective states. Kapoor & Picard aimed to classify teachers' labels of interest in children's gameplay with Fripple Place, a constraint-satisfaction game [13]. A mixture of Gaussian processes was used to achieve 87% accuracy with combined modalities of facial expression, posture, and task. In a later study, D'Mello & Graesser produced expert judgments of six affective states (boredom, confusion, engagement/flow, frustration, delight, surprise, and neutral) in interactions with the AutoTutor intelligent tutoring system [3]. A multimodal feature set of dialogue, facial expression, and posture performed best at classifying the affective states, with Cohen's κ of 0.33 for fixed emotion judgments and 0.39 for spontaneous ones.

In another line of research with the Wayang Outpost intelligent tutoring system, multimodal features were observed and categorized for their occurrence across learning-centered affective states [24]. This investigation included modalities of head movement, gaze, gesture, facial expression, posture, and vocalizations. This work was followed up with efforts to predict self-reported affective states (confidence, excitement, frustration, and interest) during interactions with Wayang Outpost [1]. Multiple sensor streams were used, including facial expression tracked by the MindReader system, pressure-sensitive mouse, skin conductance bracelet, and pressure-sensitive chair. Stepwise regression models were constructed across feature combinations, with best fit models achieving effect sizes 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 [2]. The validation results showed that the previously used features were only partially generalizable to the new population, with reduced accuracies for most features.

In contrast to these prior studies, this paper presents an analysis of how modalities of dialogue, nonverbal behavior, and task action are predictive of students' whole-session retrospective self-reports of affect. This approach identified observed behaviors that are associated with engagement, frustration, and learning throughout tutoring sessions. Studies in this vein may provide significant insight into persistent tutoring-centric phenomena that analyses of moment-by-moment states would not.

3. MULTIMODAL TUTORING 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.

3.1 JavaTutor Study

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. Recordings of the sessions included database logs, webcam video, skin conductance, and Kinect depth video. For logistical reasons, video and physiology were recorded only at the student workstations. The present analysis examines the database logs, webcam video, and Kinect depth video from the first lesson as a multimodal tutoring corpus. The JAVATUTOR interface is shown in Figure 1.

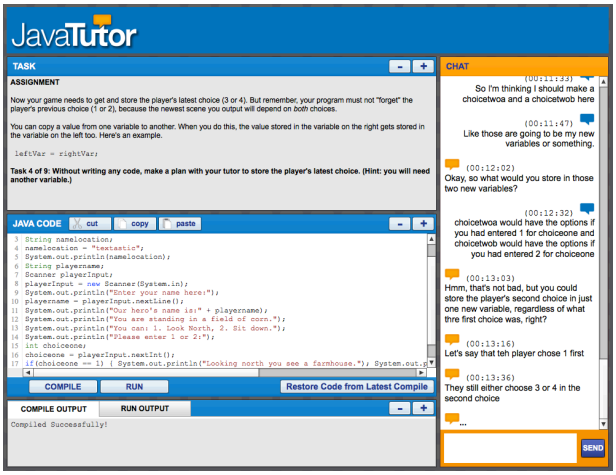


Figure 1. The JAVATUTOR interface

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) [17] and the NASA-TLX workload survey [11], which included an item for Frustration Level. The survey items for retrospective self-report of engagement and frustration are shown in Figure 2. For more information on the UES, see a recent study validating the survey in another task-oriented domain [23].

User Engagement Survey (Likert items):
- I lost myself in this learning experience.
- I was so involved in my learning task that I lost track of time.
- I blocked out things around me when I was working.
- When I was working, I lost track of the world around me.
- The time I spent working just slipped away.
- I was absorbed in my task.
- During this learning experience I let myself go.
- 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.
- I was really drawn into my learning task.
- I felt involved in this task.
- This learning experience was fun.

Frustration Level (0-100 scale):
- How insecure, discouraged, irritated, stressed, and annoyed were you?

Figure 2. Retrospective student affect survey items

3.2 Dialogue Acts

Student-tutor dialogue provides a rich source of data. Analyzing the dialogue transcripts can shed light on the effectiveness of tutoring strategies and participation of the student. Additionally, certain dialogue moves are related to student task progress and learning, as a tutor may need to provide more or less help.

The corpus was previously manually annotated using a fine-grained dialogue act annotation scheme [21]. A J48 decision tree classifier was constructed from these labels, producing an accuracy of 80.11% (Cohen's $K = 0.79$) on a held-out test set. This automatic dialogue act classifier [20] provided annotations of over 8,300 tutor and student messages across the corpus.

Due to space constraints, only the dialogue acts relevant to the results of the analysis are presented (see Table 1 and Table 2). The dialogue acts provide coverage of statements, questions, answers, and feedback of both tutors and students. Within these main categories, there are specific dialogue acts for tutors and students. For instance, tutor dialogue acts include more types of statements and questions, while student dialogue acts include feedback on the student's current understanding of the material.

Table 1. Subset of Student Dialogue Acts (DAs)

Student DA Label	Examples
AEX EXTRA DOMAIN ANSWER	*Pretty good, just a lot of homework.* *I am great.*
E EXPLANATION	*Once you respond with a name…* *I need a write statement.*
FNU NOT UNDERSTANDING FEEDBACK	*I'm not sure if this is right…* *I'm not sure, I guess it would because it is Java?*
FU UNDERSTANDING FEEDBACK	*Oh, alright. Makes sense.* *Ohh, I see!*
GRE GREETING	*Hello!* *See you next time.*
O OBSERVATION	*See, the comment was ignored by Java.* *As you see, we have a bug.*
QI INFORMATION QUESTION	*Why doesn't it stop on the next line in this case?* *How does that work?*

3.3 Nonverbal Behavior

We examined nonverbal behavior modalities as potential indicators of learning-centered cognitive-affective states across tutoring sessions. Recent advances in nonverbal behavior tracking were leveraged to recognize facial expression, hand-to-face gestures, and body posture. Due to human data collection error, the multimodal tutoring corpus includes nonverbal data for sixty-three of the sixty-seven students (N=63).

3.3.1 Facial Expression

The corpus includes facial expression annotations from the Computer Expression Recognition Toolbox (CERT) [15]. This tool recognizes fine-grained facial movements, or facial *action units (AUs)*, detailed in the Facial Action Coding System [6]. 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. A recent article provides a detailed description of the technology used in CERT [25].

Based on the results of a validation study of CERT output, we adopt the method of mean-centering the output of CERT [8]. The validation study showed that this adjustment to CERT output produced excellent aggregate agreement with manual FACS annotations on a subset of five action units, while unadjusted CERT output did not. Also in accordance with that study [8], we use a mean-centered output threshold of 0.25 for presence of a facial action unit. This higher threshold may reduce false positives compared to the default threshold of any positive value. Additionally, we examine the five action units that were previously validated: AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU7 (Lid Tightener), and AU14 (Mouth Dimpler).

Table 2. Subset of Tutor Dialogue Acts (DAs)

Tutor DA Label	Examples
ACK ACKNOWLEDGEMENT	*Okay.* *I see.*
AWH WH-QUESTION ANSWER	*The string PlayerName.* *That depends on the development.*
AYN YES/NO ANSWER	*Yes.* *No, sir.*
E EXPLANATION	*Once you respond with a name…* *I need a write statement.*
FO OTHER FEEDBACK	*That's fine.*
OEX EXTRA DOMAIN OTHER	*Calc is hard.* *I want to thank you for helping us out.*
QEX EXTRA DOMAIN QUESTION	*How are classes going?* *How are you today?*
QO OPEN QUESTION	*How can you fix it?* *How could you solve this problem?*
R REASSURANCE	*We have plenty of time.* *One thing: the more mistakes you make, the more you will learn.*

3.3.2 Posture

Postural features were created using depth image recordings from the Kinect sensor. A posture tracking algorithm identified distances of the head, upper torso, and lower torso. This posture tracking algorithm was previously found to be 92.4% accurate versus manual labels [7].

Head distances were then used to discretize postural distance (POSNEAR, POSMID, and POSFAR), based on workstation-specific median distances and standard deviation. One standard deviation closer or farther than the median (across subjects) was considered "near" or "far," respectively.

Additionally, postural movements were labeled 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; otherwise, it was marked as NOMOVE.

3.3.3 Gesture

Hand-to-face gestures were also tracked in the Kinect depth sensor recordings. A gesture detection algorithm recognized one or two hands touching the lower face. This algorithm relies on surface propagation from the center of the head, with pixel distances from the center used to identify a round (i.e., normal head shape) or an oblong shape (i.e., surface extends beyond the head) formed by the surface pixels. The hand-to-face gesture detection algorithm was previously found to be 92.6% accurate compared to manual labels [7]. Examples of detected hand-to-face gestures are shown in Figure 3.

Figure 3. Examples of detected hand-to-face gestures

3.4 Multimodal Features

The automatically recognized dialogue acts and nonverbal behaviors were combined with task action 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 (STUDENTMESSAGE and TUTORMESSAGE, respectively). Each of these student and tutor messages has an associated dialogue act label, as described in Section 3.2. These dialogue events comprise the DIALOGUE data stream.

The TASK data stream consists of student task actions. As a student completed the programming task, he or she would 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. The student may also be working on the program code (CODING), or have stopped coding (STOPCODING) at each moment.

The NONVERBAL data stream consists of student facial expression, hand-to-face gestures, and posture. Each of the nonverbal behaviors were tracked at all times, so they were combined in parallel (i.e., each interval records the presence or absence of facial expression, gesture, and posture).

These data streams were discretized into one-second intervals. The most recent event of a given type (DIALOGUE, NONVERBAL, TASK) was used as the current value at each interval. For instance, if a student had been coding but stopped after half a second into the current interval, the task action would be assigned to STOPCODING. These one-second time intervals were used to calculate relative duration following a specific dialogue event or task action, or during a particular nonverbal behavioral display. Thus, each possible feature has a single numerical value (relative duration) for each student session. The simplest feature sets constructed in this way (the *unimodal* sets) consist of a single data stream. The average relative durations of each feature in the unimodal TASK feature set are shown in Table 3.

Table 3. Average relative duration in the TASK feature set

	Avg. Relative Duration (%)
CODING	18
STOPCODING	25
COMPILESUCCESS	9
COMPILEERROR	4
RUNPROGRAM	44

The *bimodal* feature sets each consist of the Cartesian product of two unimodal feature sets. This resulted in three bimodal feature sets: DIALOGUE × NONVERBAL, DIALOGUE × TASK, and NONVERBAL × TASK. Similarly, the complete *trimodal* feature set consists of the Cartesian product of all three unimodal feature sets, DIALOGUE × NONVERBAL × TASK. A final feature set combined all three bimodal feature sets through set union. This BIMODAL UNION feature set allows for comparison of the combined bimodal feature sets versus the complete trimodal feature set.

4. MULTIMODAL FEATURE ANALYSIS

This paper presents two levels of analysis: 1) comparison of feature set performance in predicting tutorial outcomes; and 2) comparison of the most predictive features across feature sets. These were performed for each of the three tutoring outcomes: engagement, frustration, and learning gain.

4.1 Model Construction

Model averaging was used to identify the most generalizable features in models with bimodal or trimodal feature sets and to reduce the feature space [19]. This approach produces average coefficient estimates and standard error across a wide range of models. Ratios of absolute value of the coefficient estimate versus standard error were also computed. These ratios provided a tradeoff between predictive weight and numerical stability, as estimates with lower standard error varied less across models. The features were then sorted using the ratios and the top twenty were selected for use in model building. (In the case of unimodal feature sets, all features were used and model averaging was not performed.) Predictive models were built using forward stepwise linear regression. Features were selected to optimize the leave-one-out cross-validated R^2 value of each model.

4.2 Engagement

The predictive models of engagement across feature sets, shown in Table 4, display a clear advantage of the trimodal feature set combining dialogue, nonverbal behavior, and task actions (D×N×T). The next best model uses the bimodal feature set combining dialogue and task actions (D×T). The union of bimodal feature sets (BIMODAL UNION) performs worse than D×T, but it uses fewer parameters. The trimodal feature set explains 50% more variance in retrospective self-reported engagement compared to the next best model (D×T) and uses fewer parameters.

Table 4. Engagement Feature Sets

Feature Set	R^2	# parameters
Unimodal Feature Sets		
DIALOGUE	0.048	3
NONVERBAL	-0.030	2
TASK	0.006	4
Bimodal Feature Sets		
DIALOGUE × NONVERBAL	0.146	3
DIALOGUE × TASK	0.187	9
NONVERBAL × TASK	0.112	5
Trimodal Feature Sets		
BIMODAL UNION	0.157	7
DIALOGUE × NONVERBAL × TASK	0.282	6

The two predictive models of retrospective self-reported engagement incorporating all three modalities of dialogue, nonverbal behavior, and task actions are shown in Table 5. There is slight overlap between the models (features that occur in both models are in bold type). The combination of student observational statements (O-STUDENT) and stopping coding was a negative predictor of engagement in both models. This may indicate moments when the student stopped working on the task in order to make a comment, which is consistent with loss of focus on the task. Mouth dimpling (AU14) was also involved in negatively predictive features in both models. This facial action unit has been associated with frustration and mental effort in prior analyses [8, 14]. Open questions from the tutor (QO-TUTOR) with no student brow lowering (NoAU4) were positively predictive of engagement in the model with the combined bimodal feature set (Bimodal Union). Brow lowering has been previously associated with confusion, frustration, and mental effort [5, 8, 9, 14]. Thus, this may highlight moments when the student was not perplexed by a tutor question. In the model with the trimodal feature set, student informational questions (QI-STUDENT) with outer brow raising (AU2) and running the program were positively predictive of engagement. This may indicate some interest in how the program operates, with the student actively engaging in discussion with the tutor.

A sequence of tutoring events related to engagement is shown in Figure 4. The left image shows the student coding the program. The middle shows when the student had stopped coding and read a yes/no answer from the tutor. In the right image, the student has returned to coding the program.

CODING | AYN-TUTOR, POSNEAR, STOPCODING | CODING

Figure 4. Sequence related to engagement

Table 5. Engagement Feature Comparison

Feature	β	p
BIMODAL UNION, $R^2 = 0.157$		
O-STUDENT, **STOPCODING**	-0.310	0.002
QEX-TUTOR, **AU14**	-0.283	0.005
GRE-STUDENT, COMPILESUCCESS	-0.275	0.007
FO-TUTOR, **AU2**	-0.247	0.061
FO-TUTOR, AU7	-0.195	0.137
QO-TUTOR, NOAU4	0.269	0.008
INTERCEPT	0.003	1
DIALOGUE × NONVERBAL × TASK, $R^2 = 0.282$		
AEX-STUDENT, **AU14**, RUNPROGRAM	-0.333	0.003
O-STUDENT, AU1, **STOPCODING**	-0.201	0.082
TYPINGMESSAGE, AU4, CODING	-0.197	0.110
AYN-TUTOR, POSNEAR, **STOPCODING**	0.256	0.014
QI-STUDENT, **AU2**, RUNPROGRAM	0.707	<0.001
INTERCEPT	0.076	1

4.3 Frustration

A comparison of predictive models of retrospective self-reported frustration across unimodal, bimodal, and trimodal feature sets is shown in Table 6. As with the models of engagement, there is a clear advantage of the trimodal feature sets. The trimodal feature sets greatly improve upon the bimodal feature sets, with the complete trimodal feature set (DIALOGUE × NONVERBAL × TASK) explaining more than three times the variance of the best bimodal feature set (DIALOGUE × TASK). The union of bimodal features (BIMODAL UNION) also explains over two times the variance of the individual bimodal feature sets, demonstrating an additive effect across the combined feature sets.

Table 6. Frustration Feature Sets

Feature Set	R^2	# parameters
Unimodal Feature Sets		
DIALOGUE	-0.033	1
NONVERBAL	-0.010	2
TASK	-0.033	1
Bimodal Feature Sets		
DIALOGUE × NONVERBAL	0.019	2
DIALOGUE × TASK	0.137	2
NONVERBAL × TASK	0.134	5
Trimodal Feature Sets		
BIMODAL UNION	0.347	7
DIALOGUE × NONVERBAL × TASK	0.520	5

Both predictive models of retrospective self-reported frustration that include all three modalities of dialogue, nonverbal behavior, and task actions are shown in Table 7. Interestingly, both models provide only positive predictors of frustration. There is also a fair degree of overlap between the models (shared features are indicated in bold). In both models, higher frustration is predicted by students providing feedback on their current understanding when working on the program. Students often did this after receiving tutor input on how the program works. So, this type of dialogue act may confirm that they understood the new information given by the tutor. However, the underlying trend may be that students had just received help from the tutor

in order to remove a misconception. Thus, the students may feel frustration partially due to the lingering misconception.

Table 7. Frustration Feature Comparison

Feature	β	p
BIMODAL UNION, $R^2 = 0.347$		
FNU-STUDENT, **POSNEAR**	0.141	0.210
FNU-STUDENT, AU14	0.202	0.042
QEX-TUTOR, **STOPCODING**	0.260	0.013
STUDENTMESSAGE, RUNPROGRAM	0.260	0.008
OEX-TUTOR, POSMOVE	0.323	0.003
FU-STUDENT, **CODING**	0.456	<0.001
INTERCEPT	0	1
DIALOGUE × NONVERBAL × TASK, $R^2 = 0.520$		
R-TUTOR, NOAU1, **STOPCODING**	0.091	0.549
FU-STUDENT, NOAU2, **CODING**	0.336	<0.001
OEX-TUTOR, AU7, **STOPCODING**	0.429	0.006
FU-STUDENT, **POSNEAR**, **CODING**	0.451	<0.001
INTERCEPT	0	1

Another prominent pattern in both models is the emphasis on moments when the student was coding the program or had stopped coding. The moments of coding aligned with student feedback on his or her current level of understanding (FU-STUDENT). On the other hand, moments when the student stopped coding were associated with tutor messages. These tutor messages were related to off-topic statements (OEX) or questions (QEX), and reassurance (R). In such instances, the tutor is focusing on off-task discussion. Therefore, these features may be related to moments when the student stopped coding and received off-task tutor messages (with causality in either direction).

A sequence of tutoring events related to frustration is shown in Figure 5. First, the student encounters a compile error. The tutor then directs the student on how to fix the problem. Second, the student reports her understanding of the directive and implements the solution. Third, the tutor tells the student to compile and she does so successfully.

COMPILEERROR | FU-STUDENT, POSNEAR, CODING | COMPILESUCCESS

Figure 5. Sequence related to frustration

4.4 Normalized Learning Gain

Normalized learning gain (or percent learning gain) measures how much a student learned relative to what he or she could have learned [16]. This accounts for relative differences in learning between students who scored high or low on the pretest. Normalized learning gain was computed with 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 models of normalized learning gain are shown in Table 8, containing all three modalities of dialogue, nonverbal behavior, and task actions. In contrast with the affective tutoring outcomes, a unimodal feature set has significant predictive power, as the DIALOGUE unimodal feature set is predictive of learning. This underscores the important role of tutorial dialogue in the process of learning. Despite the significant predictive power of the best bimodal feature set (DIALOGUE × NONVERBAL), the complete trimodal feature set (D×N×T) still explains around sixteen percent more variance in learning gains.

The trimodal feature comparison models of normalized learning gain are shown in Table 9. These models overlap on the mouth dimpling facial action unit (AU14), which is a negative predictor of learning in both models. In viewing the tutoring session videos, AU14 appeared to correspond with moments when students were expending mental effort or thinking about the task. In the BIMODAL UNION model, AU14 coincides with tutor extra-domain questions. In the complete trimodal model (D×N×T), AU14 co-occurs with students running the program and remarking on lack of understanding in one trimodal feature, and with student stopping coding and tutor answers to complicated questions (not a "yes" or "no" answer) in another trimodal feature. In each case, the presence of AU14 may involve a reaction to a recent event in the tutoring session, whether it is the behavior of the program or messages from the tutor.

Table 8. Normalized Learning Gain Feature Sets

Feature Set	R^2	# parameters
Unimodal Feature Sets		
DIALOGUE	0.370	10
NONVERBAL	0.037	3
TASK	-0.034	1
Bimodal Feature Sets		
DIALOGUE × NONVERBAL	0.465	6
DIALOGUE × TASK	0.407	13
NONVERBAL × TASK	0.243	10
Trimodal Feature Sets		
BIMODAL UNION	0.460	5
DIALOGUE × NONVERBAL × TASK	0.544	8

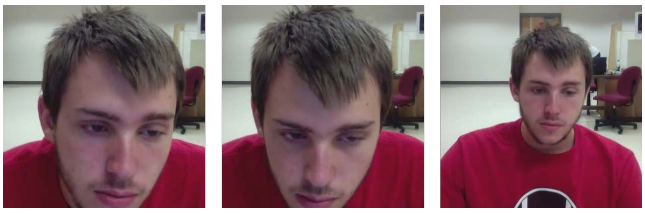
On the side of positive predictors of learning, facial action unit features were involved in both models. In the BIMODAL UNION model, learning is positively predicted by student outer brow raising (AU2) with tutor feedback that is not distinctly positive or negative (FO-TUTOR). In the complete trimodal model (D×N×T), the positive predictors that included facial action units were tutor open questions (QO-TUTOR) after stopping coding with brow lowering (AU4), and tutor acknowledgements with no brow lowering (NOAU4) during coding. In the first case, the student may be effortfully considering the tutor's question. In the second case, the student may be continuing work on the task with no hesitation due to the tutor

acknowledgement message. Both the positive and negative predictors of normalized learning gain demonstrate the importance of upper and lower face expression recognition in task-oriented multimodal systems.

Table 9. Normalized Learning Gain Feature Comparison

Feature	β	p
BIMODAL UNION, $R^2 = 0.460$		
O-STUDENT, **STOPCODING**	-0.478	<0.001
QEX-TUTOR, **AU14**	-0.341	0.001
GRE-STUDENT, **COMPILESUCCESS**	-0.314	0.002
FO-TUTOR, AU2	0.295	0.002
INTERCEPT	0.002	1
DIALOGUE × NONVERBAL × TASK, $R^2 = 0.544$		
FNU-STUDENT, **AU14**, RUNPROGRAM	-0.415	<0.001
AWH-TUTOR, **AU14**, **STOPCODING**	-0.311	0.001
E-TUTOR, POSMOVE, CODING	-0.213	0.019
FU-STUDENT, AU1, **COMPILESUCCESS**	-0.171	0.053
E-STUDENT, NOMOVE, RUNPROGRAM	0.132	0.160
ACK-TUTOR, NOAU4, CODING	0.205	0.028
QO-TUTOR, AU4, **STOPCODING**	0.231	0.011
INTERCEPT	0.002	1

A sequence of events related to learning is shown in Figure 6. In the first image, the student is testing his program. Further in the session, the student has begun coding more of the program. In the second image, the student has stopped coding and displayed AU4 as the tutor asked an open-ended question. The third image shows the student at a farther postural distance after successfully compiling the program.



RUNPROGRAM | QO-TUTOR, AU4, STOPCODING | COMPILESUCCESS

Figure 6. Sequence related to normalized learning gain

5. DISCUSSION

This paper has presented an in-depth comparison of multimodal feature sets related to engagement, frustration, and learning in computer-mediated human tutoring. The results show a distinct additive effect of features across modalities of dialogue, nonverbal behavior, and task actions. Each set of models found an improvement from unimodal to bimodal features and from bimodal to trimodal features.

Prior work has demonstrated mixed results in applying multimodal feature sets to prediction of affect. Often, a particular feature set is useful for one affective state, but not another [4]. However, most prior multimodal studies of tutoring did not involve a strong dialogue component [1, 2, 13]. In the present multimodal tutoring corpus, student-tutor dialogue plays a very significant role in predicting tutoring outcomes. A majority of features involve dialogue in the models built on a union of bimodal features, which had the potential to select nonverbal behavior and task features instead. Additionally, dialogue was the only unimodal feature set that was strongly predictive of a tutoring outcome—in this case, normalized learning gain. The importance of adaptive dialogue has also been shown in studies that examined the advantages of one-to-one tutoring [22].

While dialogue was of primary importance in these findings, the nonverbal behavior and task modalities also provided additional explanatory power. Task actions were fairly straightforward (e.g., the student was working on the task or not), but nonverbal behaviors co-occurred with specific task contexts. For instance, both presence of brow lowering (AU4) and its absence (NOAU4) appeared as positive predictors of normalized learning gain. The task contexts associated with these predictions were tutor utterance (QO vs. ACK) and student task actions (STOPCODING vs. CODING). The first feature (QO-TUTOR, AU4, STOPCODING) describes a moment when a student has stopped coding, has been posed an open-ended question by the tutor, and is thoughtfully reflecting on the question (as evidenced by brow lowering). The second feature (ACK-TUTOR, NOAU4, CODING), in contrast, may highlight a moment when the student is focused on implementing the program after receiving an acknowledgment from the tutor. AU4 may be absent in this context because the student knows how to modify the program and is making the changes with certainty. Similarly, a near postural position (POSNEAR) was predictive of both higher engagement and higher frustration. The divergent contexts of POSNEAR were student/tutor utterance (AYN-TUTOR vs. FU-STUDENT) and student task actions (STOPCODING vs. CODING). In the first case (AYN-TUTOR, STOPCODING), it seems that the student may sit near and stop coding while reading the tutor answer, which may reflect focused concentration. In the second case (FU-STUDENT, CODING), the student may be responding to tutor help and continuing work on the task, in which case the student may be having difficulty with the task, in turn associated with frustration.

Some nonverbal behaviors were more consistently predictive of tutoring outcomes. Mouth dimpling (AU14) appeared in multiple predictive features as an indicator of lower engagement, higher frustration, and reduced learning gain. Additionally, postural movement (POSMOVE) was associated with lower learning gain and increased frustration. Despite the alignment of these nonverbal behaviors toward negative affect, it is important to note that they were conditioned upon specific tutorial contexts. Thus, further analyses are necessary to infer generalizable situations in which these nonverbal behaviors occur.

6. CONCLUSION

It has long been hypothesized that combining multiple modalities improves multimodal interfaces' interpretive capabilities. In the study reported here, human tutors communicated with students through a text-based interface. Automated approaches were leveraged to annotate dialogue, task actions, facial movements, postural positions, and hand-to-face gestures. These dialogue, task progress, and nonverbal behavior input streams were then used to predict retrospective student self-reports of engagement and frustration, as well as pre/post-test learning gains. The overall finding is that the complete trimodal feature set is most predictive of each of the three tutoring outcomes. Additionally, the affective models showed large improvements from unimodal to bimodal and bimodal to

trimodal feature sets. Dialogue played a large role in predicting learning gain, so the magnitude of improvement due to multimodal features was less for this outcome, though the bimodal and trimodal feature sets did outperform unimodal dialogue. Close examination of model features revealed that bimodal features with dialogue were most predictive of each tutoring outcome and that the role of nonverbal behavior may depend on the dialogue and task context in which it occurs.

These results presented here suggest a promising direction for investigating multimodal feature sets in affective tutorial interaction. Detailed task-contextualized multimodal information can provide insight not only into the moment-by-moment affective states experienced by users, but also on the states that may pervade the users' retrospective perception of their interaction. Future adaptive multimodal interfaces may leverage this detailed task-contextualized information to disambiguate affect and intervene effectively.

ACKNOWLEDGEMENTS

This work is supported in part by the North Carolina State University Department of Computer Science along with the National Science Foundation through Grants DRL-1007962 and IIS-1409639. 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*, 17–24.

[2] 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.

[3] 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, 147–187.

[4] D'Mello, S.K. and Kory, J. 2012. Consistent but Modest: A Meta-Analysis on Unimodal and Multimodal Affect Detection Accuracies from 30 Studies. *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, 31–38.

[5] D'Mello, S.K., Lehman, B., Pekrun, R. and Graesser, A.C. 2014. Confusion Can Be Beneficial for Learning. *Learning & Instruction*. 29, 153–170.

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

[7] 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.

[8] 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*, 43–50.

[9] 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*, 159–165.

[10] 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.

[11] 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.

[12] Joshi, J., Goecke, R., Parker, G. and Breakspear, M. 2013. Can Body Expressions Contribute to Automatic Depression Analysis? *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 1–7.

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

[14] 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.

[15] 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*, 298–305.

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

[17] 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.

[18] Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A. and Morency, L.-P. 2013. Automatic Behavior Descriptors for Psychological Disorder Analysis. *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 1–8.

[19] 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, 13–21.

[20] Vail, A.K. and Boyer, K.E. 2014. Adapting to Personality Over Time: Examining the Effectiveness of Dialogue Policy Progressions in Task-Oriented Interaction. *Proceedings of the 15th Annual SIGDIAL Meeting on Discourse and Dialogue*, 41–50.

[21] Vail, A.K. and Boyer, K.E. 2014. Identifying Effective Moves in Tutoring: On the Refinement of Dialogue Act Annotation Schemes. *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, 199–209.

[22] VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A. and Rosé, C.P. 2007. When Are Tutorial Dialogues More Effective Than Reading? *Cognitive Science*. 31, 1, 3–62.

[23] 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, 123–132.

[24] 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, 129–164.

[25] 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.