Automated measurement of children's facial expressions during problem solving tasks

Gwen C. Littlewort, Marian S. Bartlett
Institute for Neural Computation
University of California, San Diego
San Diego, CA, USA
gwen@mplab.ucsd.edu; mbartlett@ucsd.edu

Linda P. Salamanca, Judy Reilly
Department of Psychology
San Diego State University
San Diego, CA, USA

Abstract—**There has been growing recognition of the importance of adaptive tutoring systems that respond to the student's emotional and cognitive state. However little is known about children's facial expressions during a problem solving task. What are the actual signals of boredom, interest, or confusion in real, spontaneous behavior of students? The field also is in need of spontaneous datasets to drive automated recognition of these states. This project aims to collect, measure, and describe spontaneous facial expressions of children during problem solving. We apply the computer expression recognition toolbox (CERT) to videos of behavioral data from children between the ages of 3 and 9, particularly focusing on the spontaneous expressions of uncertainty. From the Facial Action outputs, we analyze changes in facial expression during problem solving, and differences in expression between correct and incorrect trials. Moreover, we demonstrate differences in expression dynamics between older and younger children during problem solving. Future work examines differences between facial action and other modalities such as voice pitch.**

Keywords-facial expression recognition; spontaneous expressions; adaptive tutoring; problem solving; Facial Action Coding System; FACS;

I. INTRODUCTION

There is growing recognition that social interaction between students and teachers plays a crucial role in the effectiveness of learning. One-to-one tutoring has been shown to increase student performance by two standard deviations over conventional group methods of instruction [6]. While major research efforts have focused on cognitive aspects of adaptive one-to-one tutoring such as adjusting to student performance and modeling cognitive processes (e.g.[1] [23] [13]), another major component of one-to-one tutoring is nonverbal behavior. The use of appropriate facial expressions and gestures by teachers has been associated with greater student learning ([28] [12]), student state motivation (Christophel, 1990), and student attendance and participation [29]. These behaviors are negatively correlated with verbal aggression [30] and student resistance [22].

Motivated by this empirical evidence there has been a growing thrust to develop tutoring systems and agents that respond to the students' emotional and cognitive state and interact with them in a social manner. Eg. [24] [13] [31] [25] [11] [27] [34] [10] [8] [9] [2] [21]. For example, AutoTutor is an intelligent tutoring system that interacts with students using natural language to teach physics, computer literacy, and critical thinking skills [20]. The system adapts to the cognitive states of the learner as inferred from dialogue and performance. A new affect sensitive version is presently under development [14]. This system detects four emotions (boredom, flow/engagement, confusion, frustration) by monitoring conversational cues and gross body language.

While the pioneering work has established the potential of socially aware agents, progress has been slowed down due to the lack of datasets that could drive the computational study of nonverbal behavior during learning: What are the actual signals of boredom, interest, or confusion in real, spontaneous behavior of students? This project aims to collect, measure, and describe spontaneous facial expressions of children during problem solving.

II. APPROACH

A. Problem Solving Task

A database of children's facial behavior was collected during a set of problem solving tasks. The tasks were 1. A haptic object recognition task, in which the children were given a 3D object to recognize from touch, with their hands inside a box to prevent viewing. 2. A spatial problem solving task, "rush hour", in which sliding puzzle pieces in a certain sequence will free the desired piece, 3. Arithmetic, and 4. A task designed to elicit frustration, in which there is a toy locked inside a clear plastic box. The children are given a key and told they can play with the toy, and the experimenter leaves the room. However it is the wrong key. The experimenter later produces the correct key.

Video data from 50 children ages 3-9 was collected. The present analysis focuses on task 1, haptic object recognition. Up to 8 different 3D objects plus up to 8 texture samples were presented to each child to recognize by touch. The objects were square block, cylinder, triangle, dog, dinosaur, car, duck, fish. Children in the 3-5 year old range were given just the first five objects, and older children were given 6-8 objects. Thus far, data from 24 children has been analyzed, for a total of 94 trials.

Support for this work was provided by NSF grants SBE-0542013, IIS-0905622, CNS-0454233 and NSF ADVANCE award 0340851. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Figure 1. Haptic object recognition task.

B. *The Computer Expression Recognition Toolbox*

The Computer Expression Recognition Toolbox (CERT) is a fully automated system that analyzes facial expressions from video in real-time ([3] [4] [5] [15] [26]). The system automatically detects frontal faces in the video stream and codes each frame with respect to 20 facial action units (AU's) from the Facial Action Coding System [17].

CERT employs an appearance-based discriminative approach. Such approaches have proven highly robust and fast for face detection and tracking e.g. [32]. Face detection as well as detection of six internal facial features is first performed on each frame using boosting techniques ([19] [16]). The automatically located faces then undergo a 2D alignment by computing a fast least squares fit between the detected feature positions and a six-feature face model. The unconstrained least squares alignment adjusts for rotation, scale, and shear. The aligned face image is then passed through a bank of Gabor filters 8 orientations and 9 spatial frequencies (2 to 32 pixels per cycle at half octave steps). Output magnitudes were then normalized and passed to facial action classifiers.

Facial action detectors were then developed by training separate linear support vector machines to detect the presence or absence of each facial action. The training set consisted of over 10000 images which were coded for facial actions from the Facial Action Coding System, including over 5000 examples from spontaneous expressions [4]. The output of each facial action detector consists of a real valued number indicating the distance to the hyperplane that separates the two classes (the margin) for each frame of video. System outputs are significantly correlated with the intensity of the facial action, as measured by FACS expert intensity codes [4], and also as measured by naïve observers turning a dial while watching continuous video [33]. Thus the frame-by-frame intensities provide information on the dynamics of facial expression at temporal resolutions that were previously impractical via manual coding.

The present analysis employed CERT version 4.4, which is available for academic use from UCSD. The following set of 19 facial actions was detected for each frame [1 2 4 5 6 7 9 10 12 14 15 17 18 20 23 24 25 26 28]. See Figure 2. Performance on children was tested. See table V.

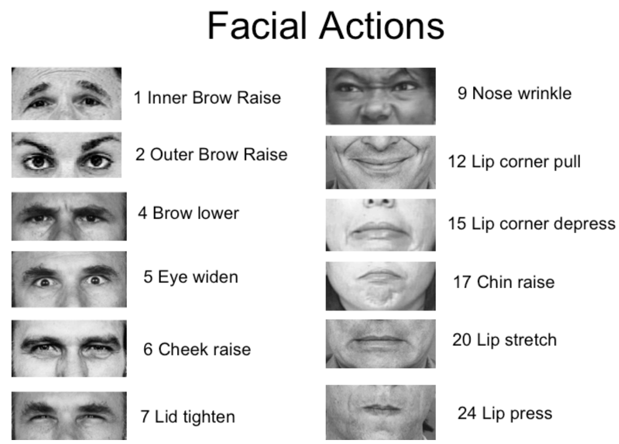

Figure 2. Sample facial actions from the Facial Action Coding System

III. Results

Analysis focused on the latency phase, which is the interval between when the child is handed the object, and when the child begins to respond with an answer. The latency phase ranged from 1-20 seconds per trial. Video data for the 24 subjects were processed by CERT 4.4. This produced a 19-dimensional output stream for each of the 94 trials. The trials grouped by whether the answer provided was correct, with 64 correct trials and 30 error trials. In order to investigate possible developmental differences in facial behavior, the trials were also grouped by age of the subject, 55 in older group ages 6-9, and 39 in the younger group, ages 3-5. Figure 4 shows raw CERT outputs during the latency phase for a 4 year old male subject.

A. *Characterizing Facial Behavior in the Latency Period*

We first characterized the overall latency period behavior by examining which AU's changed significantly from the beginning to the end of the latency, using the average AU intensity in the first and last third of a second of each trial. Changes from the beginning to the end of the latency were then tested using within-subjects paired t-tests. The results are shown in Table 1. Facial actions associated with concentration in the lower face were higher at the start of latency and were reduced at the end of the latency (dimpler,14, chin raise, 17, lip tighten 23). There was also a significant increase in AU25 (mouth open) and AU6 (orbicularis oculi) which has been related to positive emotions, at the end of latency.

Trials with correct and incorrect responses were then compared within subjects. The 63 possible pairs of correct and error trials from the same subject provided differences in the latency period AU change (beginning to end) for the incorrect minus correct. This was tested with paired t-tests (2-tailed). Two actions (9 and 18, nose wrinkle and lip pucker) increased significantly more during error trials than correct trials in the same subject. Moreover, if we allow a 1-tailed test, AU 4 (corrugator), also emerged as increasing significantly more on error trials than correct trials (p=.085). The corrugator was termed the "muscle of concentration" by Darwin and has also been associated with negative emotions [18].

We also investigated differences in AU activity at the end of the latency period (average for the last 10 frames) for older versus younger groups and correct versus incorrect. The only significant differences were between age groups, with younger children having more AU10 and 15.

Comparisons of the older versus younger age groups showed the same trends in the increase or decrease from beginning to end of the latency of significant AU's. However developmental differences were discovered in the dynamics as described next.

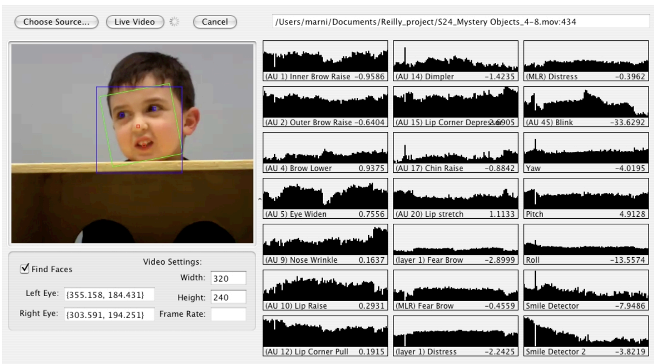

Figure 3. Screen capture of sample child video processed by CERT. Each subplot is a measure of the magnitude of a facial action in each frame of video.

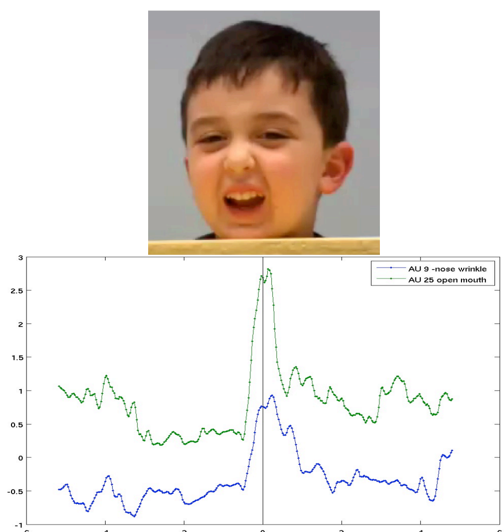

Figure 4. Sample raw CERT output during latency period for Nose wrinkle (top) and mouth open (bottom). The face image is shown at the time point indicated at the vertical line.

TABLE I. INCREASING AND DECREASING AUS AVERAGED OVER 94 TRIALS, COMPARING FIRST 10 FRAMES WITH THE LAST 10 FRAMES OF LATENCY.

AU code	Significant difference	2 tail ttest p-value	Last-first 1/3 sec
4		>0.10	0.07
6	increase	<0.05	0.32
7		>0.10	0.17
9	increase	<0.04	0.33
10	increase	<0.01	0.43
12		>0.10	0.12
14	decrease	<0.05	-0.26
17	decrease	<0.01	-0.37
23	decrease	<0.01	-0.35
24		>0.10	-0.21
25	increase	<0.01	0.53

B. *Expression Dynamics*

The joint dynamics of facial actions during the latency period were first visualized using paired action trajectories, or "phase" plots. This visualization technique facilitates the comparison of the dynamics of two actions, when the changes might be rare events in a large amount of data. The two actions may be synchronous, opposite, change sequentially, be partially correlated or uncorrelated.

For many subjects there appeared to be distinct states with abrupt transitions. The phase plots in Figure 5 show two main states with a rapid shift in between of about 1/10 of a second. The top left plot shows smile and orbicularis oculi. Both are associated with happiness. They both start out low and then there is a rapid shift to higher activation at the end. The top right plot shows brow lower and eye widening. Brow lower starts out high, and eye widening low. As the child is about to report the answer, there is a rapid change, and brow lower releases and the eye widening increases.

The lower two plots show the relation of head motion and face motion. The ordinate is yaw angle, where the child is facing roughly forward. In the bottom left plot, brow lower starts out high. Then both the release of the brow lower and the head motion towards the experimenter occur at the same time. In the bottom right plot, smile starts out low and subject frontal. Here the transition is more stepped, where the smile increases for about 1/3 second and then the head motion towards the experimenter occurs. In contrast, Figure 5 shows phase plots from a younger subject, which appears to be much less organized into states.

C. *K-Means Clustering*

In order to investigate the organization of facial behavior into states, K-means clustering was performed. We specifically investigated whether there were differences between older and younger children in the organization of their facial movements into states, as well as whether there were differences in clustering for correct versus incorrect trials. The clustering algorithm was applied to the 19 channel time series for each trial. The number of clusters was chosen as the largest K for which the centers were far enough apart (correlation <c=0.9) and for which the maximum temporal duration of each state was at least T (1/3 second) The results below are insensitive to these criterion parameters c and T. The following variables were tested for significance: t0, the length of the sequence, K, the number of distinct states, N, the number of state changes (provided the duration of each new state >1/3sec), N/K, the mean number of changes per state and N/t0 the mean number of changes per second. The trials from older children were not significantly different from the younger children. See table 2.

The trials where the response was correct differed from the error trials in several ways. See Table 3. Error trials are longer, contain more states, and change more rapidly. Separating the older and younger sets shows that this effect is much stronger for the older children. For example, the older children differ between the trial types on the transitions per state. See table 4. By contrast, the younger children show little difference between correct and error trials. Note that N<K is possible when some states are very short-lived.

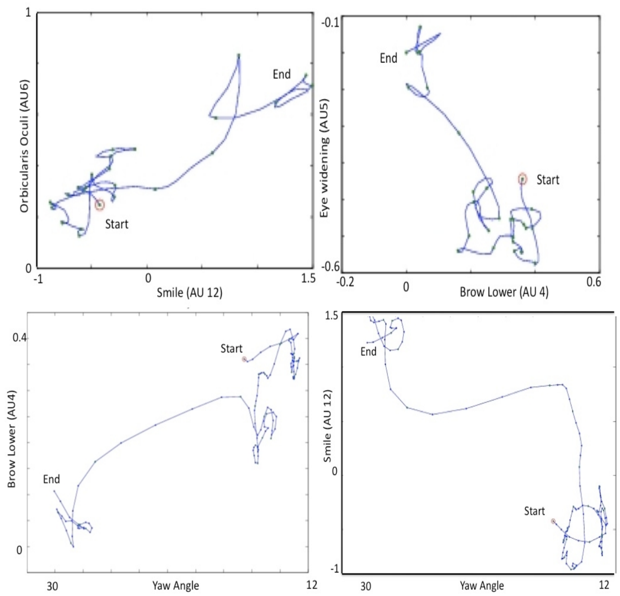

Figure 5. Phase plots showing temporal trajectories of two or more facial actions. ‘Start’ shows where the object is given to the child, and ‘end’ is when he begins to say the answer. Top left: Smile vs orbicularis oculi. Top right: Brow lower (corrugator) vs eye widening. Bottom left: Yaw angle of head against brow lower. Bottom right: Yaw angle against smile.

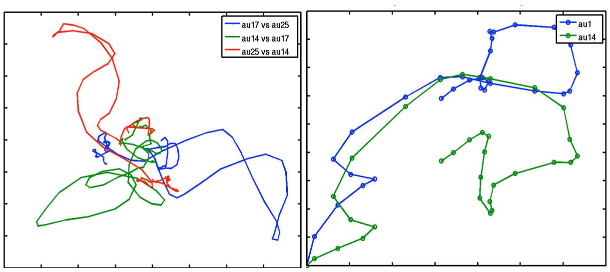

Figure 6. The left plot shows uncoordinated actions in pairs (AU17 vs 25 blue; AU14 vs 17 green; AU25 vs 14 red) and the right plot shows partially coordinated pairs (AU1 vs 12 blue; AU14 vs 12 green).

TABLE II. UNPAIRED 2-TAILED T-TESTS FOR COMPARISONS BETWEEN OLDER AND YOUNGER AGE GROUPS.

measure	p	Age 3-5	Age 6-8
Latency time	ns	2.9 sec	4.0 sec
Number of states	ns	3.5	4.4
Number of changes	ns	4.3	5.8
Changes per state	ns	1.15	1.15
Changes per second	ns	1.39	1.39

TABLE III. UNPAIRED 2-TAILED T-TESTS SHOW THAT THE CORRECT AND ERROR TRIALS DIFFER SIGNIFICANTLY ON ALL THESE MEASURES.

measure	p	Correct trials	Error trials
Latency time	.03	3.1 sec	4.6 sec
Number of states	.05	3.5	5.0
Number of changes	.003	4.1	7.7
Changes per state	.01	1.04	1.38
Changes per second	.02	1.28	1.61

TABLE IV. NUMBER OF CHANGES PER STATE = K/N SHOWING LITTLE DIFFERENCE FOR YOUNGER CHILDREN.

Age group	Correct	Error
Younger	1.12	1.22
Older	.99	1.48

D. Human FACS code comparison with CERT on children.

This project was the first of several studies of children using CERT. The accuracy question naturally arose, since CERT was trained on adults while children are much more mobile and have different facial morphology. Increased movement results in blurred or empty frames or faces far from frontal. Fifteen percent of the videos were rejected, largely because the subjects faced away from the camera. A sample of about 200 frames were FACS labeled by a human. Faces were found in 97% of these frames by CERT. Up to 3 AUs per frame were coded as present/absent, from a set of 11 actions. The results in table V show that the performance on children is similar to adults. The average hit rate for found frames was 92% while the average ROC was 79% comparing human to CERT FACS on children.

TABLE V. PERFORMANCE OF CERT COMPARED TO HUMAN FACS FOR UPPER FACE, BASED ON 199 FRAMES.

Action Unit	Number of examples	ROC area	Hit rate
AU 1	57	0.78	0.89
AU 4	51	0.87	0.92
AU 5	18	0.92	0.61
AU 6	26	0.89	1.00
AU 9	36	0.87	1.00

LOWER FACE ACTIONS

AU 12	36	0.90	0.94
AU 14	9	0.59	1.00
AU 17	20	0.83	1.00
AU 18	13	0.70	0.69
AU 20	12	0.61	1.00
AU 23	19	0.77	1.00

I. DISCUSSION

There has been growing recognition of the importance of adaptive tutoring systems that respond to the student’s emotional and cognitive state. However little is known about children's spontaneous facial expressions during problem solving. The long-term goal of this project is to provide information about children's behavior during problem solving, and to contribute a dataset of spontaneous expressions to drive automated recognition of states related to learning and problem solving. This research will contribute quantitative descriptions of non-verbal behavior of typically developing children, and will also provide a basis for comparison of clinical populations.

With this work, we have demonstrated that automated FACS coding can be applied to educationally relevant behavioral experiments which involve large amounts of video of spontaneous actions. The CERT system had not been applied to young children before this study. Despite morphological differences and the high mobility of children, it was still possible to uncover trends in behavior and characterize some group differences. We presented preliminary results showing that indicators of uncertainty or concentration can be automatically measured during this problem-solving task. In order to explore possible developmental differences in facial behavior, this study compared two age groups. This is an important consideration for the development of adaptive computer vision systems, as children's nonverbal behavior is often assumed to be the same across age groups.

The automated facial expression measurement system enabled novel investigations of expression dynamics which were previously infeasible due to the time required to manually code expression dynamics. This study is the first to demonstrate differences in expression dynamics of older versus younger children.

A. Ongoing Research

Another 26 subjects are being added to the analysis, including longitudinal data. Similar tasks, such as lock box, are being labeled for comparison. Data from other modalities, including voice pitch, will also be included in the analysis of joint dynamics. Voice pitch was coded using the PRAAT system [7]. Figure 10 shows a clip of video coded for different modalities. The lower graph shows rising voice pitch as the subject says "its like a triangle or something?" The upper graph shows CERT output for two indicator actions (nose and brow wrinkle) that children may use to convey uncertainty. Figure 1 shows a facial expression from the same clip. In this example, an increase in these two facial actions is associated with an increase in pitch. Ongoing research explores how children express uncertainty across modalities, and how these modalities are coordinated in time. The 30 Hz measurements provided by automated facial expression systems such as CERT facilitate such investigations of cross-modal dynamics.

Training classifiers to predict labels is a useful approach because the weights reveal which actions are important for a particular decision or, for example, whether it matters which time window is used for analysis. Classifiers trained to predict the age of a subject based on their expressions early during mystery box latency perform at 0.89 ROC, using Multinomial Logistic Regression. The representation was based on a the intensity, slope and acceleration of AUs in the first half of latency. AU 5, 9, 10, 15 and 25 had positive weight for 8 year olds, where as negative weights were associated with AU 1, 2, 6, 14 and 20.

Predicting the correctness of a response is more difficult (0.65) than predicting age, but once it is conditioned on the age of the subject the ROC is 0.73 for 4 year olds and 0.81 for 8 year olds for predicting whether the answer will be correct or incorrect. In this case the weights for incorrect were higher for AU2, 6, 17 and 25 in the second half of latency, with negative weights for AU 23, 15, 5 and 4.

In summary, training classifiers indicates that age is predictable based on early latency, whereas correctness is predictable based on later latency particularly in older children.

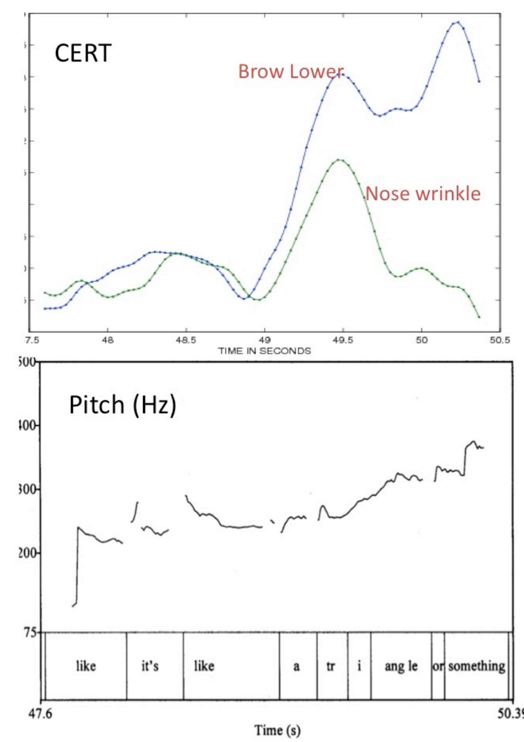

Figure 7. CERT output for two facial actions (top), time locked with pitch measures from Praat, and speech transcription (bottom). Figure 1 shows a facial expression from the same clip.

References

[1] Anderson, J.R., C.F. Boyle, and B.J. Reiser, (1985). Intelligent Tutoring Systems. Science vol. 228, pp. 456-462.

[2] Bailenson, J.N. and N. Yee, Digital Chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. Psychological Science, 2005. 16: p. 814-819.

[3] Bartlett, M. S., G. Littlewort, et al. (2004). Machine Learning Methods for Fully Automatic Recognition of Facial Expressions and Facial Actions. IEEE International Conference on Systems, Man & Cybernetics. The Hague, Netherlands: 592- 597.

[4] Bartlett M.S., Littlewort G.C., Frank M.G., Lainscsek C., Fasel I., and Movellan J.R., (2006). Automatic recognition of facial actions in spontaneous expressions. Journal of Multimedia., 1(6) p. 22-35.

[5] Bartlett, M. S., Littlewort-Ford, G.C., and Movellan, J.R. (2008). “Computer Expression Recognition Toolbox. .” Proc: 8th Int’l IEEE Conference on Automatic Face and Gesture Recognition, Amsterdam.

[6] Bloom, B.S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational Researcher, Vol. 13, No. 6, pp. 4-16.

[7] Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International 5(9/10): 341-345.

[8] Brand, M. Voice puppetry. in ACM SIGGRAPH. 1999: ACM Press/Addison-Wesley Publishing Co.

[9] Burleson, W. and R.W. Picard, Evidence for Gender SpecificApproaches to the Development of Emotionally Intelligent Learning Companions. IEEE Intelligent Systems, Special issue on Intelligent Educational Systems, 2007.

[10] Cassell, J. and K.R. Thórisson, The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. International Journal of Applied Artificial Intelligence, 1999. 13(4-5): p. 519-538.

[11] Chaffar, S. & Frasson, C. (2004). Inducing optimal emotional state for learning in intelligent tutoring systems. In Lester, Vicari, & Paraguacu (Eds.), Proceedings, Intelligent tutoring systems. Springer.

[12] Chesebro, J.L. and McCroskey, J.C. (2001). The relationship of teacher clarity and immediacy with student state receiver apprehension, affect, and cognitive learning. Communication Education, 50, 59-68.

[13] Conati, C., A. Gertner, and K. van Lehn, (2002) “Using Bayesian Networks to Manage Uncertainty in Student Modeling”, User Modeling and User-Adapted Interaction 12: 371-417.

[14] D’Mello, Graesser, & Picard, (2007). Towards an Affect sensitive Auto tutor. IEEE Intelligent Systems. July/AUG p. 53- 61.

[15] Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P. & Sejnowski, T.J. (1999). “Classifying facial actions.” IEEE Trans. Pattern Analysis and Machine Intelligence 21(10): 974-989.

[16] Eckhardt, M., Fasel, I.R. and Movellan, J.R. (2009). Towards Practical Facial Feature Detection. International Journal of Pattern Recognition and Artificial Intelligence 23(3) pp. 379- 400.

[17] Ekman, P. and W. Friesen (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto, CA, Consulting Psychologists Press.

[18] Ekman, P. (2003). Darwin, deception, and facial expression. Annals New York Academy of sciences, Vol. 1000, pp. 205-221.

[19] Fasel, I., Fortenberry B., Movellan J.R. “A generative framework for real-time object detection and classification.,” Computer Vision and Image Understanding 98, 2005.

[20] Graesser, A.C., Jackson, G.T., & McDaniel, B. (in press). AutoTutor holds conversations with learners that are responsive to their cognitive and emotional states. Educational Technology.

[21] Gratch, Wang, Gerten, Fast, & Duffy (2007). Creating rapport with virtual agents. Int. Conf. on Intelligent Virtual Agents, Paris, France.

[22] Kearney, P., Plax, T.G., Smith, V.R., and Sorensen, G. (1988). Effects of teacher immediacy and strategy type on college student resistance to on-task demands. Communication Education, 37, 54-67.

[23] Koedinger K.R and John R. Anderson J.R., "Intelligent Tutoring Goes to School in the Big City", Int'l Journal of Artificial Intelligence in Education (1997), 8, 30-43.

[24] Kort, B., R. Reilly, and R.W. Picard, “An Affective Model of Interplay between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion,” Proc. 2nd IEEE Int’l Conf. Advanced Learning Technologies (ICALT 2001), IEEE CS Press, 2001, p. 43.

[25] Litman, D.J. and K. Forbes-Riley, “Predicting Student Emotions in Computer-Human Tutoring Dialogues,” Proc. 42nd Ann. Meeting Assoc. for Computational Linguistics, Assoc. for Computational Linguistics, 2004, pp. 352_359.

[26] Littlewort G.C., Bartlett M.S., Fasel I, Susskind J, Movellan “An automatic system for measuring facial expression in video.” Computer Vision and Image Understanding: Special Issue on Face Processing in Video 24(6) p615-625,2006

[27] McQuiggan, S. W., and Lester, J. C. (2007). Leveraging Affect for Narrative-Centered Guided Discovery Learning Environments. Supplementary Proceedings of the 13th Inter- national Conference on Artificial Intelligence in Education (AIED 2007), (pp 67-74).

[28] Mottet, T.P., Parker-Raley, J., Cunningham, C., Beebe, S.A., and Raffeld, P. C. (2006). Testing the neutralizing effect of instructor immediacy on student course workload expectancy violation and tolerance for instructor unavailability. Communication Education, 55, 147-166.

[29] Rocca, K.A. (2004). College student attendance: Impact of instructor immediacy and verbal aggression. Communication Education, 53, 185-195.

[30] Rocca, K.A. and McCroskey, J. C. (1999). The interrelationship of student ratings of instructors’ immediacy, verbal aggressiveness, homophily, and interpersonal attraction. Communication Education, 48, 308-316.

[31] Sheldon-Biddle, E., Malone, L., McBride, D.: Objective measurement of student affect tooptimize automated instruction. In: Proceedings of Workshop on Modelling User Attitudes and Affect, User Modeling (2003).

[32] Viola, P. & Jones, M. (2004). Robust real-time face detection. J. Computer Vision, Vol. 57, No. 2, pp. 137-154.

[33] Whitehill, J., Bartlett, M.S., and Movellan, J.R. Measuring the Difficulty of a Lecture-Using Automatic Facial Expression Recognition. Proceedings of the International Conference on Intelligent Tutoring Systems 2008 (short paper).

[34] Woolf, B., Burelson, W., and Arroyo, I. (2007). Emotional Intelligence for Computer Tutors. Supplementary Proceedings of the 13th International Conference on Artificial In-telligence in Education (AIED 2007), (pp 6-15).