Automated facial affect analysis for one-on-one tutoring applications

Nicholas J. Butko, Georgios Theocharous, Matthai Philipose, and Javier R. Movellan

***Abstract*— In this paper, we explore the use of computer vision techniques to analyze students' moods during one-on-one teaching interactions. The eventual goal is to create automated tutoring systems that are sensitive to the student's mood and affective state. We find that the problem of accurately determining a child's mood from a single video frame is surprisingly difficult, even for humans. However when the system is allowed to make decisions based on information from 10 to 30 seconds of video, excellent performance may be obtained.**

I. Introduction

One-on-one tutoring is one of the most effective forms of teaching. Bloom [3] reports that students tutored 1-to-1 outperform their peers taught in classroom settings by as much as two standard deviations (2σ). Computer-based tutoring systems have the potential to radically alter the nature of education by offering a learning experience customized to each individual student [1]. Such systems may reduce the time constraints of human teachers, allowing them to spend quality time with individual students when they need it most.

In the 25 years since Bloom's report, 2σ has become an informal gold-standard against which tutoring technologies are measured. Progress has been made in areas such as middle and high school math and computer programming [10], [5]. However, the literature on automatic tutoring applied to early education is much more sparse [18], [15]. A reason for this is that early education demands forms of physical and social interaction that are beyond the scope of current automatic tutoring technology. For example, the use of physical objects, such as coins, rods, cubes, patterns and other concrete objects, called manipulatives, is critical for teaching abstract and symbolic mathematical concepts in kindergarten and early grades [17], [4], [14], [19]. Just as important is embedding the learning experience in the context of a social interaction [13], [11]. Properly managing the students' moods, including level of attention, interest, frustration, *etc*., is always important but it is particularly critical in early education.

In this paper we focus on the problem of detecting the student's mood along dimensions relevant to teaching. Our goal was not to create new algorithms, but to assess the critical challenges for facial expression recognition in the automated tutoring domain, and to benchmark current approaches to automated facial processing, to see if they provide useful levels of performance. The eventual goal is to create automated tutoring systems that are sensitive to the student's mood and affective state. To this end, we make the following contributions:

1) We describe the collection of a dataset of one-on-one tutoring actions, with rich facial affect information.
2) We determine, with the teacher, which affective states are important for her teaching. These states turn out to be similar to those reported elsewhere [6] but with key differences.
3) We establish ground-truth labels for the dataset based on human coders. These coders had full access to the audio and visual data streams, and used a good deal of context in making their decisions. There was high inter-coder agreement.
4) We assess the difficulty of the perceptual problem of determining affective state in teaching interactions from single frame analysis. This is done by comparing human coder labels of single frames to the labels that the same coders gave to the same frames when they had the full context of the audio and visual data streams.
5) We show that established machine learning and computer vision techniques can achieve at or above human levels in perceiving affective state in learning interactions.

II. The 2σ Dataset

Elusive categories, like mood, are difficult to capture using a set of verbalizable rules, thus the importance of data driven, machine learning approaches. There are already some video databases of teacher-student interaction [16] and of the interaction between students and automated tutoring systems [6]. However, to our knowledge, to date there are no datasets of one-on-one teaching interactions for primary education pupils with expert human teachers that are amenable to computer-vision-based analysis. To this end, we collected such a dataset, and named it the 2σ *Dataset*, to reflect the challenge and promise of personalized tutoring technologies.

We collected a video dataset of one-on-one tutoring sessions between a primary education teacher and students.

The 2σ dataset consists of 10 one-on-one tutoring sessions between a primary education teacher and her students. The 10 different students, one per session, span grades K–2 (about 5–7 years old). In each session, the teacher taught age-appropriate math problems involving manipulating coins. For Kindergartners (K), such problems might be, "Can you show me all the pennies?" while for 2^{nd} graders (2), they might be subtraction problems. Each tutoring session lasted approximately 20 minutes, and was recorded using 4 simultaneous camera views and audio (Figure 1). During the

This work was supported by the National Science Foundation, grants SBE-0542013, IIS INT2-Large 0808767, and CNS-0454233.

N.J. Butko and J.R. Movellan are at the Institute for Neural Computation, 9500 Gilman Dr., #0440, La Jolla, CA 92093-0440 {nick,movellan}@mplab.ucsd.edu

G. Theocharous and M. Philipose are at Intel labs.

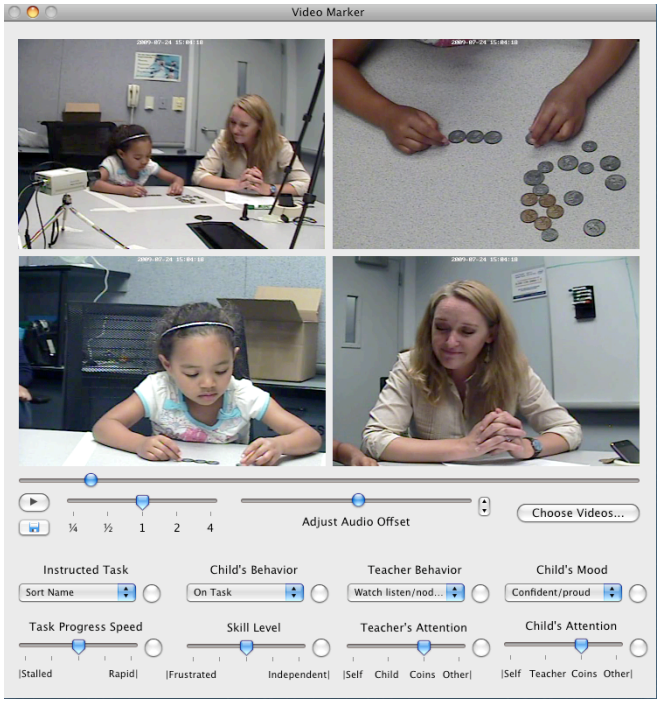

Fig. 1. The 2σ Dataset. Four video cameras simultaneously captured teaching interaction from four angles. A dynamic-context annotation tool, shown in the picture above, allowed frames-by-frame annotation of the student's mood. This tool let the labeler speed up and slow down the video, jump back and forth in the interaction, and hear the voice recordings.

study, we also interviewed the teacher regarding appropriate curricula for the K–2 age groups, about her own teaching habits, about her assessment of each child's learning after a tutoring session, and about students' behavioral indices that were relevant for making moment to moment decisions about how to manage a teaching session.

The research team, which consisted of the teacher, a developmental psychologist, and machine learning researchers, agreed on a common set of labels for the teacher actions and child moods. The teacher actions included setting up coins, reinforcing the child, watching and listening. The child moods, which are the focus of this paper included: interested, thinking, tired, confused, confident, frustrated, distracted.

A. FACS-based computer-vision analysis

The Facial Action Coding System (FACS) is an anatomically inspired, comprehensive, and versatile method to describe human facial expressions [8]. FACS encodes the observed expressions as combinations of Action Unit (AUs). Roughly speaking AUs describe changes in the appearance of the face that are due the effect of individual muscle movements.

In recent years FACS based expression recognition systems have shown dramatic progress; in some cases, these systems are as accurate at predicting mental states as human experts trained in facial analysis and much better than untrained humans [12]. These technologies are now making their way into many applications, such as judging the difference between real and acted pain experiences [12], judging how difficult a student finds a video lecture [21], [22], warning a drowsy driver who is about to fall asleep [20], and improving the veracity of facial expressions made by a robot [23].

In this paper we use the 2σ Dataset to explore whether these methods could be used to analyze the mood of children in tutoring situations.

B. Student Mood States

After the tutoring sessions, we conducted a post-experiment interview with the teacher. In reviewing the videos with her, we asked, "What about the child's mood was important for making decisions about how to manage your interactions?" From her input and careful observation of the video data we extracted a set of key situations that represent the mood state of the child. Below we describe these states and give example scenarios to define their meaning.

1) *Interested*. The child is either looking at the coin or teacher and listening attentively.
2) *Thinking*. The child is manipulating coins to solve one of the assigned tasks.
3) *Tired / Bored*. In this state the child has already had enough lessons in the session and is beginning to lose her interest and visibly not paying attention to either the coins or the teacher.
4) *Confused*. The child, seems to be making the wrong moves with the coins on the table. Other signs are hovering over the coins and being undecided as to what to do and also looking up at the teacher trying to get hints as to what is the right thing to do.
5) *Confident / Proud*. The child is smiling, moves coins correctly and fast, and when she finishes a task sits back with a smile and lets the teacher look at her accomplishment. She smiles even more when the teacher gives her reinforcement after a successful task completion.
6) *Frustrated*. The child is unsure how to make progress, and has stopped trying.
7) *Distracted*. The child momentarily switches her attention to different objects and situations other than coins and the teacher, *e.g*. the cameras or the chair.

Examples of the above mood states are shown in Figure 2. They are similar to those previously reported in the literature by D'Mello [7], but there are some differences. As in the current study, previous research also identified boredom, confusion, and frustration as learning-relevant moods. Our "thinking" category is similar to what D'Mello calls "flow," a kind of engaged problem solving. We did not observe anything like "surprise" in our teaching interactions. For our teacher, "confident" and "proud" were similar emotions, and were considered more useful labels than "delight" in D'Mello's system. Our "interested" label probably corresponds most to D'Mello's "neutral," although D'Mello does not give a way to encode the state of "paying attention to the teacher." Finally, D'Mello does not include the "distracted" label.

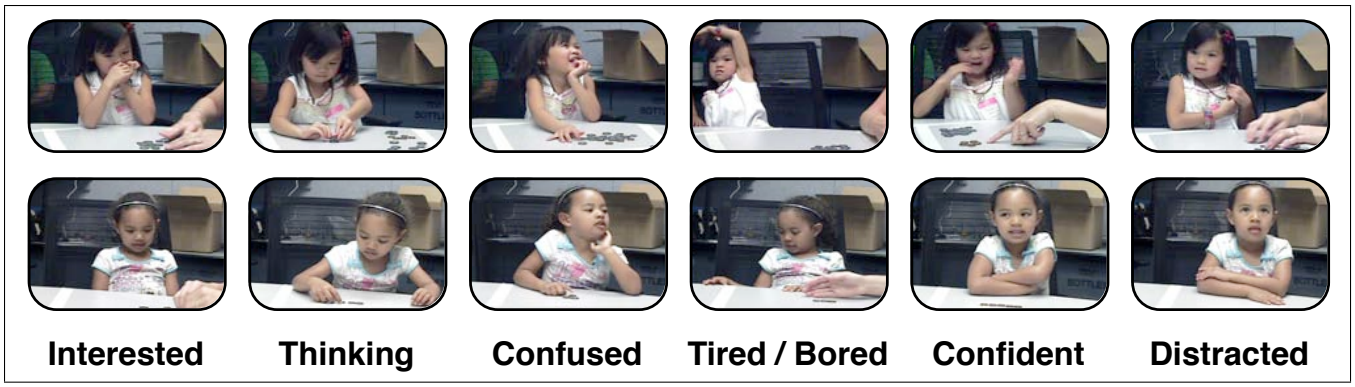

Fig. 2. These images describe visually the different mood states that can be observed in the data for two students. The images of the two children for each mood are similar, and show how information-rich the face is.

In this paper we are concerned with two computer-vision tasks: (A) Correctly predict which of the seven moods the student is showing on each frame: (B) Correctly predict whether a student is showing a positive mood (Interested, Thinking, Confident / Proud) or a negative mood (Tired / Bored, Confused, Frustrated, Distracted).

C. Dynamic-context Labeling

We created a video annotation tool, shown in Figure 1 for labeling many aspects of the student-teacher interaction, including the student's mood. Other labels included the instructed task (macro-teaching behavior), the teacher's momentary actions (micro-teaching behavior), *etc*.

Labeling a single feature of an interaction is a time consuming process, taking much longer than the length of the interaction. Typically the videos are watched at half-speed, with full audio, and even in this rich context, different segments need to be watched multiple times to decide the correct label. Because of this, in this study we focused on the mood labels for two students.

The videos were collected at 30 frames per second (FPS). In the labeling process, we labeled mood boundaries, in which the student changed from one mood to another. Every frame within that boundary was marked with the mood. In this way, were able to get a dense 30-FPS set of mood labels for the teaching interactions. Each video session consisted of just under 40,000 frames of labeled mood data.

Throughout this paper, we take the labels gathered in this fashion to be "ground truth," *i.e*. they are the standard against which other approaches are compared. We also refer to these labels as the "dynamic-context" case, because each given frame was labeled with knowledge of preceding and subsequent video and audio context surrounding each frame.

In deciding which mood the student showed at each moment, it was helpful to have both the audio signal as well as the preceding and subsequent temporal context. For example, a student manipulating the coins and doing the right thing we might think was "thinking", while a student manipulating the coins and doing the wrong thing we might think was "confused." Without access to the audio, it might be difficult to tell from just the video whether the student was doing the right or wrong thing. Figure 2 shows examples where it might be difficult to tell the difference between "Confident" and "Distracted" based just on single frame images. In light of this difficulty, we were unsure how feasible it would be to use only the facial information available at each moment to tell which of the seven moods the child was in, or even if she was in a "good" learning state, or a "bad" one.

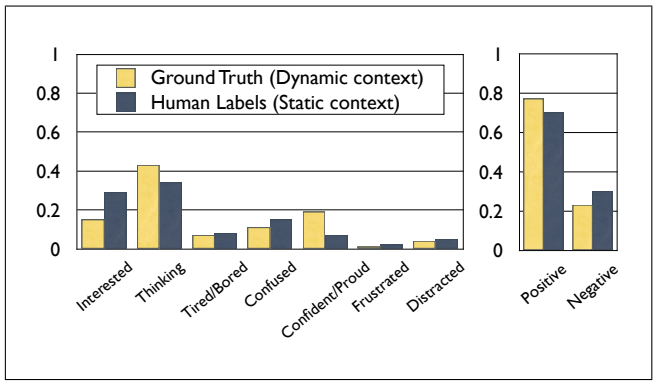

Fig. 3. Mood label histograms. The "Ground Truth" labels were decided in the full context of a dynamic teaching interaction. The "Human Labels" were decided in the context of isolated frames.

III. Baselines

In order to evaluate computer vision approaches to mood detection, we first need to establish performance baselines. The first baseline, the naïve baseline, reflects the performance achievable by simply always guessing the most likely mood. The most common mood, "thinking," occurred during 43% of our teaching interactions. In all, 77% of frames showed positive moods (Thinking, Interested, and Confident / Proud). The full histogram of moods is shown in Figure 3.

In our data, the least common "good" learning state (Interested) was more common than the most common "bad" learning state (Confused). This speaks to the teacher's exceptional ability to manage her students' moods, to keep them in emotional states conducive to learning.

Perceiving affect in single snapshots of a student's face is quite difficult: the frame is stripped of its surrounding

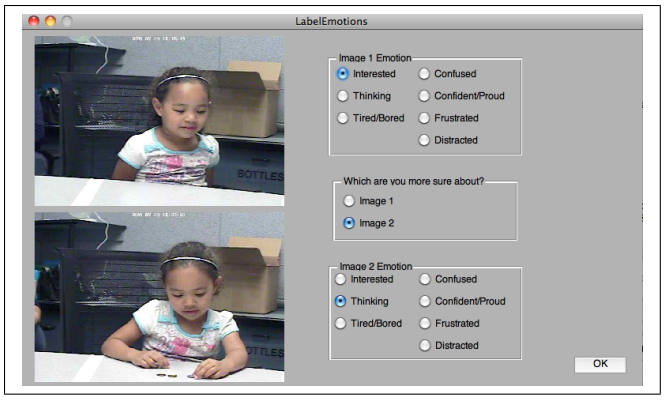

Fig. 4. Our static-context annotation tool for judging the mood of students just from snapshots, in the absence of audio and surrounding video.

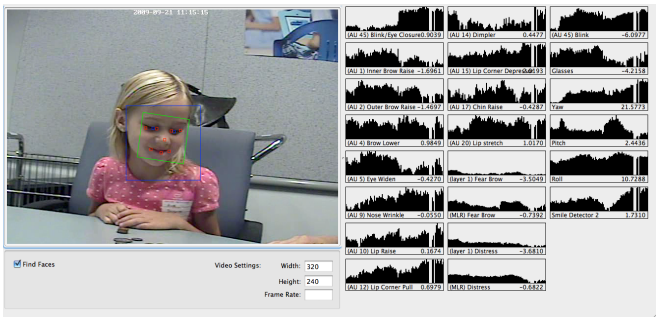

Fig. 5. The Computer Expression Recognition Toolbox automatically extracted 106 facial indices from each frame in the teaching session. These features are used for training the 7 GentleBoost classifiers to perceive the presence/absence of each mood.

video context, and all of its audio context. An important baseline for computer vision is how well humans can do in this situation.

We created a second annotation tool for the static-context, shown in Figure 4. This static annotation tool randomly selected two frames from the video sequences of the same two subjects whose emotions were previously labeled. These were presented in random order to the same human labeler who previously annotated the very same frames in the dynamic-context using the first dynamic annotation tool (Figure 1). The labeler was asked the mood of the student in each frame, and also was asked which label they were more confident in.

The results, shown in Table I, reveal static-context mood judgements to be surprisingly difficult. In the full seven-mood judgement task, labelers were only 36% correct, which was worse than the 43% naïve baseline. Performance goes up slightly to 39% for the more confident frames. Positive / negative affect discrimination was slightly better: 80% of the time, subjects chose correctly a positive or negative mood, slightly better than the 77% naïve baseline. There was no apparent effect of confidence on affect perception.

Table II shows the confusion matrix when the human labeler made mood decisions based on a single video frame. Each row represents the "true" mood of the child, *i.e.*, the judgement made by the observer when he had access to the entire audio-visual context in which a snapshot occurred. Each row is the mood chosen by the observer when he could see a single snapshot. The table reveals that the labeler's errors were not random. For example, if a child was confused, the labeler was 33% likely to say she was thinking, and only 21% likely to correctly say that she was confused.

	Human	Confident	Unconfident	Naïve
7-Mood	0.36	0.39	0.33	0.43
Pos/Neg	0.80	0.80	0.80	0.77

TABLE I

PERFORMANCE BASELINES FOR FRAME-BY-FRAME MOOD PERCEPTION.

When a child was frustrated, most of the time she appeared confused. Distraction was often mistaken for confidence.

IV. STATIC-CONTEXT MOOD DETECTION

The Computer Expression Recognition Toolbox (CERT, shown in Figure 5) [2] is a state of the art system for automatic FACS coding from video. This toolbox was obtained from Machine Perception Techologies,[1] and it is available to academic research groups free of charge upon request. In this experiment, CERT provided 106 indices of facial information for each video frame. These indices included information about facial action units from FACS, head position, and head orientation. We use a version of GentleBoost [9] to train 7 different classifiers. We chose GentleBoost because it has previously used as an effective classifier, but other classifiers have also found to be effective in this domain, and may also perform well. GentleBoost can be viewed as a case of sequential logistic regression with optimally tuned features. Each feature was one of the 106 output dimensions. For each of these features an optimal tuning function is constructed using local kernel regression methods. Features are then sequentially added to a classifier by selecting the one that most reduces a chi-squared error measure. In this paper we stopped the learning process after 4 features were selected. For the remainder of this paper we call this approach "CERT+GentleBoost."

The classifier was trained to detect emotions in the static-context, *i.e.* from each frame independently.

We evaluated the obtained classifiers using a leave-one-out cross-validation method, *i.e.* the 7 classifiers trained on Student A were evaluated on Student B, and the 7 classifiers trained on Student B were evaluated on Student A. This evaluation method is harsh but realistic: when a system is deployed, it is evaluated on students that were not seen in training. Better results can be expected by training and testing on the same student.

On each frame, the classifiers output the probability for each of the seven moods separately. Normalizing across moods gives a probability distribution of the classifiers'

[1] http://mpt4u.com/

Static \ Dynamic	Interested	Thinking	Tired / Bored	Confused	Confident / Proud	Frustrated	Distracted
Interested	0.18	0.08	0.20	0.18	0.25	0.20	0.08
Thinking	0.11	**0.57**	0.15	**0.33**	0	0.20	0
Tired / Bored	**0.28**	0.05	**0.40**	0.05	0	0	0.17
Confused	0.11	0.15	0.05	0.21	0	**0.60**	0
Confident / Proud	0.18	0.13	0.05	0.15	**0.75**	0	0.33
Frustrated	0	0	0.10	0	0	0	0
Distracted	0.14	0.02	0.05	0.08	0	0	**0.42**

TABLE II

HUMAN LABELS CONFUSION MATRIX. EACH COLUMN, j IS THE TRUE MOOD OF THE CHILD, DETERMINED BY A HUMAN FROM A DYNAMIC-CONTEXT. EACH ROW, i, IS THE MOOD JUDGMENT MADE BY A HUMAN. THE MEANING OF ENTRY (i, j) IS "WHEN THE CHILD WAS IN MOOD j, THE HUMAN VIEWING A SNAPSHOT THOUGHT SHE WAS IN MOOD i." UNDERLINES SHOW THE PROBABILITY OF A CORRECT MOOD JUDGEMENT. BOLDFACE SHOWS THE MOST LIKELY JUDGEMENT. NOTE THAT IN THE STATIC, SINGLE FRAME CONTEXT, HUMANS NEVER CORRECTLY LABELED CONFUSION.

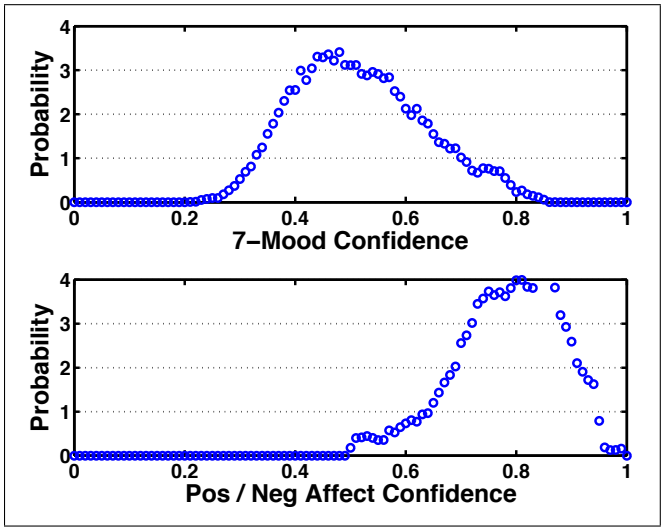

Fig. 6. CERT+GentleBoost confidence histograms for the highest confidence label.

collective beliefs that each mood is present in each frame. The mode of this distribution (the maximum probability mood) was chosen to be the classifier label for that frame.

We defined a confidence measure as the probability of the maximum probability mood. This confidence could, in theory, be as low as .14 (1/7), and as high as 1. In practice, it was between .27 and .87, with a frequency histogram shown in Figure 6. Varying a confidence threshold to either accept only labels with high confidence (low recall) or all confidence levels (high recall) gives a precision-recall curve (PRC, shown in Figure 7).

We compare this PRC to human baselines using the labeler's confidence: The overall labeler performance is at 100% recall, while the performance on the one-half of frames that were marked as "more confident" corresponds to 50% recall. When CERT+GentleBoost is forced to label a mood for every frame, even when it is unconfident, the performance is slightly better than humans, but below the Naïve baseline. However, if CERT+GentleBoost can choose to only label high-confidence frames, its 7-way performance improves dramatically.

By summing the probability CERT+GentleBoost assigns to all positive classes, we can get a confidence that the student is showing a positive mood. In theory, the confidence of the most likely positive / negative affect label can range from 0.5 to 1, and in practice it ranged from 0.5001 to 0.9881, with a frequency histogram shown in Figure 6. Varying a confidence threshold gives a precision-recall curve (Figure 7). In positive / negative affect perception, CERT+Gentleboost was slightly better than humans for moderate confidence labels, slightly worse than humans for low confidence labels, and always better than baseline. At very high confidences, performance jumps to nearly 100% correct.

Table III shows the CERT+GentleBoost performance at different average query intervals, in seconds. The videos were collected at 30 Frames-Per-Second (FPS). A recall of 1 corresponds to CERT+GentleBoost giving outputs of the child's mood on average every 1/30th second, while a recall of 3% corresponds to outputting the child's mood 1 time per second on average. Depending on how often the teacher needs to query the mood of the student, an appropriate point on the precision-recall curve can be chosen. When given an average 10 second window of video, the system could perfectly classify the mood as being positive or negative. When given 30 seconds it achieved 83% accuracy on the 7-mood categorization task.

	1/30 s	1/3 s	1 s	3 s	10 s	30 s
7-Mood	0.36	0.50	0.63	0.72	0.75	0.83
Pos/Neg	0.79	0.81	0.81	0.81	1.00	1.00

TABLE III

PERFORMANCE OF CERT+GENTLEBOOST AT DIFFERENT AVERAGE QUERY TIMES, IN SECONDS.

V. CONCLUSION

Automatic detection of the student moods during learning may be critical for automatic tutoring systems applied to children. We identified seven key moods that are relevant to expert teachers when they make moment to moment decisions about how to proceed in their lessons. We began to explore the use of frame-by-frame facial analysis for determining a student's mood during one-on-one teaching. We showed that frame by frame mood detection is extremely

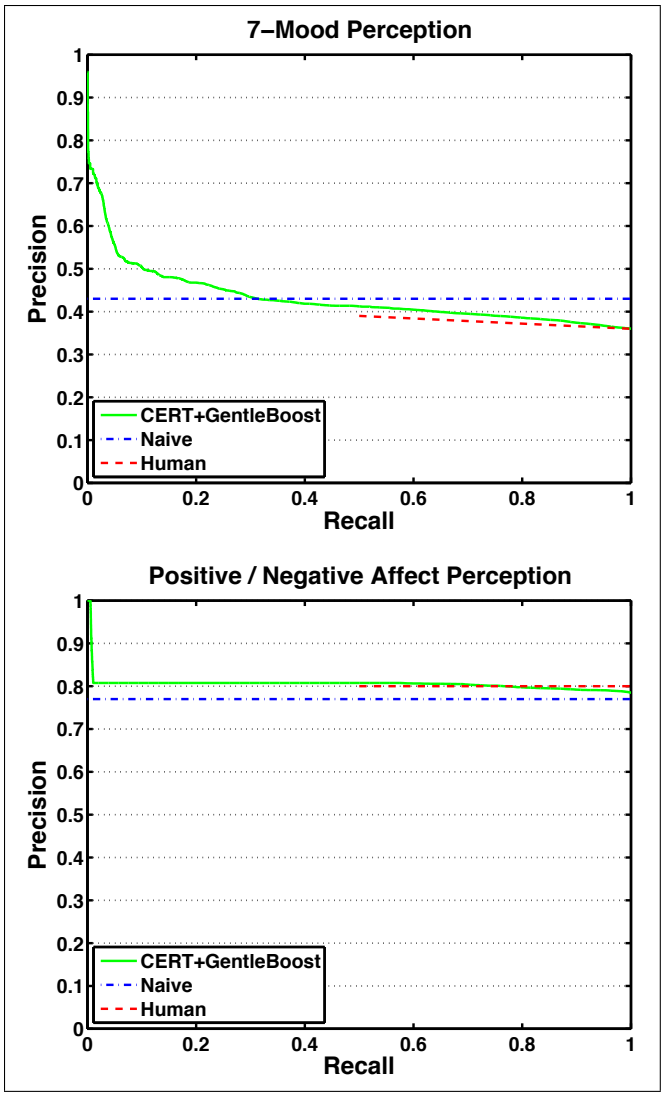

Fig. 7. CERT+GentleBoost Precision-Recall curves for 7-Mood and Positive/Negative Affect perception, compared to Naïve and Human baselines.

difficult for both humans and machines. However when given 10 to 30 seconds of temporal context current computer vision technology can provide excellent performance levels.

REFERENCES

[1] J. R. Anderson and B. J. Reiser. The lisp tutor: it approaches the effectiveness of a human tutor. *BYTE*, 10(4):159–175, 1985.

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

[3] B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16, June/July 1984.

[4] J. S. Bruner. *Toward a Theory of Instruction*. Harvard University Press, Cambridge, MA, 1966.

[5] A. T. Corbett. Cognitive computer tutors: Solving the two-sigma problem. In *Proceedings of the 8th International Conference on User Modeling 2001*, volume 2109 of *Lecture Notes In Computer Science*, pages 137–147. Springer-Verlag, 2001.

[6] S. D'Mello, R. W. Picard, and A. Graesser. Towards an affect-sensitive autotutor. *IEEE Intelligent Systems, Special Issue on Intelligent Educational Systems*, 22(4):53–61, July 2007.

[7] S. D'Mello, R. Taylor, K. Davidson, and A. Graesser. Self versus teacher judgments of learner emotions during a tutoring session with AutoTutor. In *Ninth International Conference on Intelligent Tutoring Systems*, pages 9–18, 2008.

[8] P. Ekman, W. Friesen, and J. Hager. *Facial Action Coding System (FACS): Manual and Investigator's Guide*. A Human Face, Salt Lake City, UT, 2002.

[9] I. Fasel, B. Fortenberry, and J. R. Movellan. A generative framework for real-time object detection and classification. *Computer Vision and Image Understanding*, 98:182–210, 2005.

[10] K. R. Koedinger and J. R. Anderson. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8:30–43, 1997.

[11] P. K. Kuhl. Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5:831–843, 2004.

[12] G. C. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797 – 1803, 2009. Visual and multimodal analysis of human spontaneous behaviour:.

[13] A. N. Meltzoff, P. K. Kuhl, T. J. Sejnowski, and J. R. Movellan. Foundations for a new science of learning. *Science*, 325(5938):284–288, 2009.

[14] M. Montessori. *The Montessori Method*. Schoken Books, New York, 1964.

[15] J. Movellan, M. Eckhart, M. Virnes, and A. Rodriguez. Sociable robot improves toddler vocabulary skills. *Proceedings of the 2009 International Conference on Human Robot Interaction*, 2009.

[16] R. Pea, R. Lindgren, and J. Rosen. Computer-supported collaborative video analysis. In *7th International Conference of the Learning Sciences*, 2006.

[17] J. Piaget and A. Szeminska. *The Childs Conception of Number*. W. W. Norton Co., New York, 1941.

[18] P. Ruvolo, J. Whitehill, M. Virnes, and J. R. Movellan. Building a more effective teaching robot using apprenticeship learning. *International Conference on Development and Learning*, 2008.

[19] E. J. Sowell. Effects of manipulative materials in mathematics instruction. *Journal for Research in Mathematics Education*, 20(5):498–505, November 1989.

[20] E. Vural, M. Cetin, A. Ercil, G. C. Littlewort, M. S. Bartlett, and J. R. Movellan. Automated drowsiness detection for improved driving safety. In *Proc. 4th international conference on automotive technologies*, 2008.

[21] J. Whitehill, M. Bartlett, and J. Movellan. Automatic facial expression recognition for intelligent tutoring systems. *Computer Vision and Pattern Recognition*, 2008.

[22] J. Whitehill, M. S. Bartlett, and J. R. Movellan. Measuring the difficulty of a lecture using automatic facial expression recognition. In *Intelligent Tutoring Systems*, 2008.

[23] T. Wu, N. J. Butko, P. Ruvulo, M. S. Bartlett, and J. R. Movellan. Learning to make facial expressions. In *Proceedings of the International Conference on Development and Learning (ICDL)*, Shanghai, China, 2009.