The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions

Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan

Abstract—Student engagement is a key concept in contemporary education, where it is valued as a goal in its own right. In this paper we explore approaches for automatic recognition of engagement from students' facial expressions. We studied whether human observers can reliably judge engagement from the face; analyzed the signals observers use to make these judgments; and automated the process using machine learning. We found that human observers reliably agree when discriminating low versus high degrees of engagement (Cohen's $\kappa = 0.96$). When fine discrimination is required (4 distinct levels) the reliability decreases, but is still quite high ($\kappa = 0.56$). Furthermore, we found that engagement labels of 10-second video clips can be reliably predicted from the average labels of their constituent frames (Pearson $r = 0.85$), suggesting that static expressions contain the bulk of the information used by observers. We used machine learning to develop automatic engagement detectors and found that for binary classification (e.g., high engagement versus low engagement), automated engagement detectors perform with comparable accuracy to humans. Finally, we show that both human and automatic engagement judgments correlate with task performance. In our experiment, student post-test performance was predicted with comparable accuracy from engagement labels ($r = 0.47$) as from pre-test scores ($r = 0.44$).

Index Terms—Student engagement, engagement recognition, facial expression recognition, facial actions, intelligent tutoring systems

1 Introduction

"The test of successful education is not the amount of knowledge that pupils take away from school, but their appetite to know and their capacity to learn." Sir Richard Livingstone, 1941 [36].

Student engagement has been a key topic in the education literature since the 1980s. Early interest in engagement was driven in part by concerns about large drop-out rates and by statistics indicating that many students, estimated between 25% and 60%, reported being chronically bored and disengaged in the classroom [32], [51]. Statistics such as these led educational institutions to treat student engagement not just as a tool for improving grades but as an independent goal unto itself [16]. Nowadays, fostering student engagement is relevant not just in traditional classrooms but also in other learning settings such as educational games, intelligent tutoring systems (ITS) [43], [4], [52], [30], [4], and massively open online courses (MOOCs).

The education research community has developed various taxonomies for describing student engagement. Fredricks, et al. [20] analyzed 44 studies and proposed that there are 3 different forms of engagement: behavioral, emotional, and cognitive. Anderson, et al. [3] organized engagement into behavioral, academic, cognitive, and psychological dimensions. The term *behavioral engagement* is typically used to describe the student's willingness to participate in the learning process, e.g., attend class, stay on task, submit required work, and follow the teacher's direction. *Emotional engagement* describes a student's emotional attitude towards learning – it is possible, for example, for students to perform their assigned work well, but still dislike or be bored by it. Such students would have high behavioral engagement but low emotional engagement. *Cognitive engagement* refers to learning in a way that maximizes a person's cognitive abilities, including focused attention, memory, and creative thinking [3].

The goal of increasing student engagement has motivated the interest in methods to measure it [25]. Currently the more popular tools for measuring engagement include: (1) Self-reports, (2) Observational checklists and ratings scales, and (3) Automated measurements.

Self-reports: Self-reports are questionnaires in which students report their own level of attention, distraction, excitement, or boredom [14], [24], [45]. These surveys need not directly ask the students explicitly how "engaged" they feel but instead can infer engagement as an explanatory latent variable from the survey responses, e.g., using factor analysis [41]. Self-reports are undoubtedly useful. For example, it is of interest to know that between 25% and 60% of middle school students report

- *J. Whitehill is with the Machine Perception Laboratory (MPLab), University of California, San Diego. J.R. Movellan is with the MPLab and also Emotient, Inc. Z. Serpell is with the Department of Psychology at Virginia Commonwealth University. Y-C. Lin, and A. Foster are with the Department of Psychology, Virginia State University.*
E-mail: jake@mplab.ucsd.edu, znserpell@vcu.edu, aysha.foster@gmail.com, linyichen670507@yahoo.com, movellan@emotient.com
- *Support for this work was provided by NSF grants IIS 0968573 SOCS, IIS INT2-Large 0808767, CNS-0454233, and SMA 1041755 to the UCSD Temporal Dynamics of Learning Center, an NSF Science of Learning Center.*

to be bored and disengaged [32], [51]. Yet self-reports also have well-known limitations. For example, some students may think it is "cool" to say they are non-engaged; other students may think it is embarrassing to say so. Self-reports may be biased by primacy and recency memory effects. Students may also differ dramatically in their own sense of what it means to be engaged.

Observational checklists and rating scales: Another popular way to measure engagement relies on questionnaires completed by external observers such as teachers. These questionnaires may ask the teacher's subjective opinion of how engaged their students are. They may also contain checklists for objective measures that are supposed to indicate engagement. For example, do the students sit quietly? Do they do their homework? Are they on time? Do they ask questions? [48]. In some cases, external observers may rate engagement based on live or pre-recorded videos of educational activities [46], [29]. Observers may also consider samples of the student's work such as essays, projects, and class notes [48].

While both self-reports and observational checklists and ratings are useful, they are still very primitive: they lack temporal resolution, they require a great deal of time and effort from students and observers, and they are not always clearly related to engagement. For example, engagement metrics such as "sitting quietly", "good behavior", and "no tardy cards" appear to measure compliance and willingness to adhere to rules and regulations rather than engagement *per se*.

Automated measurements: The intelligent tutoring systems (ITS) community has pioneered the use of automated, real-time measures of engagement. A popular technique for estimating engagement in ITS is based on the timing and accuracy of students' responses to practice problems and test questions. This technique has been dubbed "engagement tracing" [8] in analogy to the standard "knowledge tracing" technique used in many ITS [30]. For instance, chance performance on easy questions or very short response times might be used as an indication that the student is not engaged and is simply giving random answers to questions without any effort. Probabilistic inference can be used to assess whether the observed patterns of time/accuracy are more consistent with an engaged or a disengaged student [8], [26].

Another class of automated engagement measurement is based on physiological and neurological sensor readings. In the neuroscience literature, engagement is typically equated with level of arousal or alertness. Physiological measures such as EEG, blood pressure, heart rate, or galvanic skin response have been used to measure engagement and alertness [23], [18], [39], [49], [9]. However, these measures require specialized sensors and are difficult to use in large-scale studies.

A third kind of automatic engagement recognition – which is the subject of this paper – is based on computer vision. Computer vision offers the prospect of unobtrusively estimating a student's engagement by analyzing cues from the face [42], [29], [10], [11], body posture and hand gestures [24], [29]. While vision-based methods for engagement measurement have been pursued previously by the ITS community, much work remains to be done before automatic systems are practical in a wide variety of settings.

If successful, a real-time student engagement recognition system could have a wide range of applications: (1) Automatic tutoring systems could use real-time engagement signals to adjust their teaching strategy the way good teachers do. So-called *affect-sensitive* ITS are a hot topic in the ITS research community [13], [59], [5], [19], [26], [12], and some of the first fully-automated closed-loop ITS that use affective sensors for feedback are starting to emerge [59], [12]. (2) Human teachers in distance-learning environments could get real-time feedback about the level of engagement of their audience. (3) Audience responses to educational videos could be used automatically to identify the parts of the video when the audience becomes disengaged and to change them appropriately. (4) Educational researchers could acquire large amounts of data to data-mine the causes and variables that affect student engagement. These data would have very high temporal resolution when compared to self-report and questionnaires. (5) Educational institutions could monitor student engagement and intervene before it is too late.

Contributions: In this paper we document one of the most thorough studies to-date of computer vision techniques for automatic student engagement recognition. In particular, we study techniques for data annotation, including the timescale of labeling; we compare state-of-the-art computer vision algorithms for automatic engagement detection; and we investigate correlations of engagement with task performance.

Conceptualization of engagement: Our goal is to estimate perceived engagement, i.e., student engagement as judged by an external observer. The underlying logic is that since teachers rely on perceived engagement to adapt their teaching behavior, then automating perceived engagement is likely to be useful for a wide range of educational applications. We hypothesize that a good deal of the information used by humans to make engagement judgements is based on the student's face.

Our paper is organized as follows: First we study whether human observers reliably agree with each other when estimating student engagement from facial expressions. Next we use machine learning methods to develop automatic engagement detectors. We investigate which signals are used by the automatic detectors and by humans when making engagement judgments. Finally, we investigate whether human and automated engagement judgments correlate with task performance.

2 Dataset collection and annotation for an automatic engagement classifier

The data for this study were collected from 34 undergraduate students who participated in a "Cognitive

Skills Training" experiment that we conducted in 2010-2011 [58]. The purpose of this experiment was to measure the importance to teaching of seeing the student's face. In the experiment, video and synchronized task performance data were collected from subjects interacting with cognitive skills training software. Cognitive skills training has generated substantial interest in recent years; the goal is to boost students' academic performance by first improving basic skills such as memory, processing speed, and logic and reasoning. A few prominent systems include Brainskills (by Learning RX [1]) and FastForWord (by Scientific Learning [2]). The Cognitive Skills Training experiment utilized custom-built cognitive skills training software (reminiscent of BrainSkills) that we developed at our laboratory and installed on an Apple iPad. A webcam was used to videorecord the students; it was placed immediately behind the iPad and aimed directly at the student's face.

The game software in the experiment consisted of three games – Set, Remember, and Sum – that trained logical, reasoning, perceptual, and memory skills. The games were designed to be mentally taxing. Hard time limits were imposed on each round of the games, and the human trainers who controlled the game software (in either the Wizard-of-Oz or 1-on-1 conditions, as described below) were instructed to "push" students to perform the task more quickly. In this sense, the cognitive skills training domain of our experiment might resemble a setting in which a student is taking a stressful exam. In terms of physical environment, typical ITS and the cognitive skills setting in our study are very similar – a student sits directly in front of a computer or iPad, and a web camera retrieves frontal video of the student. It is possible that the appearance of affective states such as engagement might differ between cognitive skills training and ITS interactions. Nevertheless, it is likely that the methodology of labeling and the computer vision techniques for training automated classifiers could still generalize to more traditional ITS use cases.

The dependent variables during the 2010-2011 experiment were pre- and post-test performance on the Set game. The "Set" game in our study (see Figure 1 **right**) was very similar to the classic card game: the student is shown a board of 9 cards, each of which can vary along three dimensions: size, shape, and color. The objective is to form as many valid sets of 3 cards in the time allotted as possible. A set is valid if and only if the three cards in the set are either all the same or all different for *each* dimension. After forming a valid set, the three cards in that set are removed from the board, and three new cards are dealt. This process then continues until the time elapses.

Experimental data for the engagement study in this paper were taken from 34 subjects from two pools: (a) the 26 subjects who participated in the Spring 2011 version of the Cognitive Skills Training study at a Historically Black College/University (HBCU) in the southern United States. All of these subjects were African-American, and 20 were female. Additional data were collected from (b) the 8 subjects who participated in the Summer 2011 version of the Cognitive Skills Training study at a university in California (UC), all of whom were either Asian-American or Caucasian-American, and 5 of whom were female. The game software used in the UC dataset was identical to the software used in the HBCU except for minor differences in game parameters (e.g., how fast cards are dealt). For the present study, the HBCU data served as the primary data source for training and testing the engagement recognizer. The UC dataset allowed us to assess how well the trained system would generalize to subjects of a different race – a known issue in modern computer vision systems.

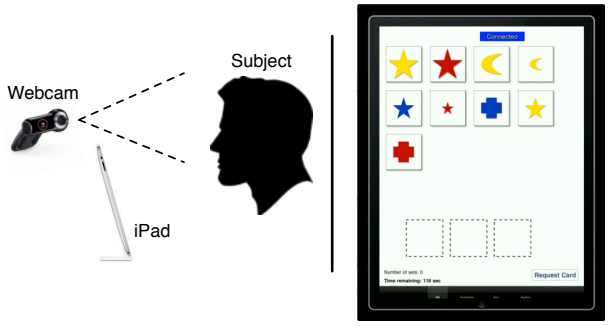


Fig. 1. **Left**: Experimental setup in which the subject plays cognitive skills training software on an iPad. Behind the iPad is a web camera that records the session. **Right**: The "Set" game in the cognitive skills training experiment that elicited various levels of engagement.

In the experimental setup, each subject sat in a private room and played the cognitive skills training software either alone or together with the experimenter. The iPad was placed on a stand and horizontally situated approximately 30 centimeters in front of the subject's face and vertically so that the iPad was slightly below eye level. Behind the iPad pointing towards the subject was a Logitech web camera that recorded the entire session. As described in [58], each subject was assigned either to a Wizard-of-Oz condition or a 1-on-1 condition. In the Wizard-of-Oz condition, the subject sat alone in the room while interacting with the game software, which was controlled remotely by a human wizard who could watch the student in real time. In the 1-on-1 condition, the subject played the games alongside a human trainer who would control the software overtly. In the HBCU dataset, 20 subjects were in the Wizard-of-Oz condition and 6 subjects were in the 1-on-1 condition. In the UC dataset, all subjects were in the Wizard-of-Oz condition.

During each session, the subject gave informed consent and then watched a 3 minute video on the iPad explaining the objectives of the three games and how to play them. The subject then took a 3 minute pre-test on the Set game to measure baseline performance. Test performance was measured as the number of valid

"sets" of 3 cards (according to the game rules) that the student could form within 3 minutes. The particular cards dealt during testing were the same for all subjects. After the pre-test, the subject then underwent 35 minutes of cognitive skills training using the training software. The trainer's goal (in both the Wizard-of-Oz and 1-on-1 conditions) was to help the student maximize his/her test performance on Set. During the training session, the trainer could change the task difficulty, switch tasks, and provide motivational prompts. After the training period, the subject took a post-test on Set and then was done.

2.1 Data annotation

Given the recorded videos of the cognitive training sessions, the next step was to label them for engagement. We organized a team of labelers consisting of undergraduate and graduate students from computer science, cognitive science, and psychology from the two universities where data were collected. These labelers viewed and rated the videos for the appearance of engagement. Note that not all labelers labeled the exact same sets of images/videos. Instead, we chose to balance the goals of obtaining many labels per image/video, and annotating a large amount of data for developing an automated detector. When labeling videos, the audio was turned off, and labelers were instructed to label engagement based only on appearance.

In contrast to the more thoroughly studied domains of automatic basic emotion recognition (happy, sad, angry, disgusted, fearful, surprised, or neutral) [33], [6], [61] or facial action unit classification [37], [27], [7], [47] (from the Facial Action Coding System [17]), affective states that are relevant to learning such as frustration or engagement may be difficult to define clearly [50]. Hence, arriving at a sufficiently clear definition and devising an appropriate labeling procedure, including the timescale at which labeling takes place, is important for ensuring both the reliability and validity of the training labels [50]. In pilot experimentation we tried three different approaches to labeling:

1) Watching video clips (at normal viewing speed) and giving continuous engagement labels by pressing the the Up/Down arrow keys.
2) Watching video clips and giving a single number to rate the entire video.
3) Viewing static images and giving a single number to rate each image.

We found approach (1) very difficult to execute in practice. One problem was the tendency to habituate to each subject's recent level of engagement, and to adjust the current rating relative to that subject's average engagement level of the recent past. This could yield labels that are not directly comparable between subjects or even within subjects. Another problem was how to rate short events, e.g., brief eye closure or looks to the side: should these brief moments be labeled as "non-engagement", or should they be overlooked as normal behavior if the subject otherwise appears highly engaged? Finally, it was difficult to provide continuous labels that were synchronized in time with the video; proper synchronization would require first scanning the video for interesting events, and then re-watching it and carefully adjusting the engagement up or down at each moment in time. We found the labeling task was easier using approaches (2) and (3), provided that clear instructions were given as to what constitutes "engagement".

2.2 Engagement categories and instructions

Given the approach of giving a single engagement number to an entire video clip or image, we decided on the following approximate scale to rate engagement:

1: Not engaged at all – e.g., looking away from computer and obviously not thinking about task, eyes completely closed.
2: Nominally engaged – e.g., eyes barely open, clearly not "into" the task.
3: Engaged in task – student requires no admonition to "stay on task".
4: Very engaged – student could be "commended" for his/her level of engagement in task.
X: The clip/frame was very unclear, or contains no person at all.

Example images for each engagement level are shown in Figure 2. Note that these guidelines pertain certainly to "behavioral engagement" [20] but they also contain elements of cognitive and emotional engagement. For example, whether or not a student is "into" the task is related to her attitude towards the learning task. Also, in our definitions above, the distinction between engagement levels 3 and 4 is related to the student's motivational state.

Labelers were instructed to label clips/images for "How engaged does the subject *appear to be*". The key here is the word *appear* – we purposely did not want labelers to try to infer what was "really" going on inside the students' brains because this left the labeling problem too open-ended. This has the consequence that, if a subject blinked, then he/she was labeled as very non-engaged (Engagement $= 1$) because, at that instant, he/she *appeared* to be non-engaged. In practice, we found that this made the labeling task clearer to the labelers and still yielded informative engagement labels. If the engagement scores of multiple frames are averaged over the course of a video clip (see Section 2.4), momentary blinks will not greatly affect the average score anyway. In addition, labelers were told to judge engagement based on the knowledge that subjects were interacting with training software on an iPad directly in front of them. Any gaze around the room or to another person (i.e., the experimenter) should be considered non-engagement (rating of 1) because it implied the subject was not engaging with the iPad. (Such moments occurred at the very beginning or very end of each session when the

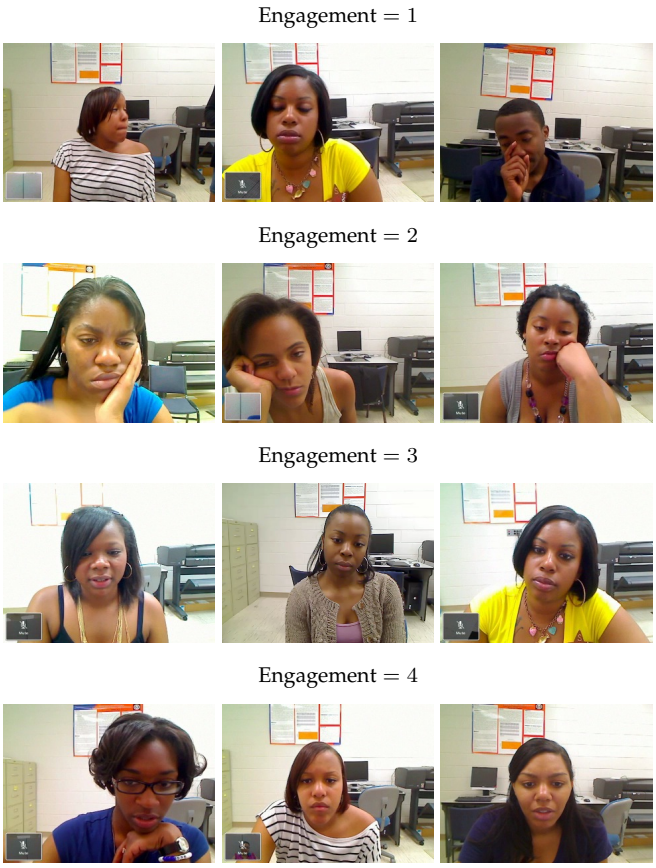


Fig. 2. Sample faces for each engagement level from the HBCU subjects. All subjects gave written consent to publication of their face images.

experimenter was setting up or tearing down the experiment.) The goal here was to help the system generalize to a variety of settings where students should be looking directly in front of them.

2.3 Timescale

An important variable in annotating video is the timescale at which labeling takes place. For approach (2) (described in Section 2.1), we experimented with two different time scales: clips of 60 sec and clips of 10 sec. Approach (3) (single images) can be seen as the lower limit of the length of a video clip. In a pilot experiment we compared these three timescales for inter-coder reliability. As performance metric we used Cohen's κ (see Appendix for more details). Since the engagement labels belong to an ordinal scale ($\{1, 2, 3, 4\}$) and are not simply categories, we used a weighted κ with quadratic weights to penalize label disagreement.

For the 60 sec labeling task, all the video sessions (~ 45 minutes/subject) from the HBCU subjects were watched from start to end in 60 sec clips, and 2 labelers entered a single engagement score after viewing each clip. For the 10 sec labeling task, 505 video clips of 10 sec each were extracted at random timepoints from the session videos and shown to 7 labelers in random order (in terms of both time and subject). Between the 60 sec clips and the 10 sec labeling tasks, we found the 10 sec labeling task more intuitive. When viewing the longer clips, it was difficult to know what label to give if the subject appeared non-engaged early on but appeared highly engaged at the end. The inter-coder reliability of the 60 sec clip labeling task was $\kappa = 0.39$ (across 2 labelers); for the 10 sec clip labeling task $\kappa = 0.68$ (across 7 labelers).

For approach (3), we created custom labeling software in which 7 labelers annotated batches of 100 images each. The images for each batch were video frames extracted at random timepoints from the session videos. Each batch contained a random set of images spanning multiple timepoints from multiple subjects. Labelers rated each image individually but could view many images and their assigned labels simultaneously on the screen. The labeling software also provided a Sort button to sort the images in ascending order by their engagement label. In practice, we found this to be an intuitive and efficient method of labeling images for the appearance of engagement. The inter-coder reliability for image-based labeling was $\kappa = 0.56$. This reliability can also be increased by averaging frame-based labels across multiple frames that are consecutive in time (see Section 2.4).

2.4 Static versus motion information

One interesting question is how much information about students' engagement is captured in the *static pixels* of the individual video frames compared to the *dynamics* of the motion. We conducted a pilot study to examine this question. In particular, we randomly selected 120 video clips (10 sec each) from the set of all HBCU videos. The random sample contained clips from 24 subjects. Each clip was then split into 40 frames spaced 0.25 sec apart. These frames were then shuffled both in time and across subjects. A human labeler labeled these image frames for the appearance of engagement, as described in "approach (3)" of Section 2.1. Finally, the engagement values assigned to all the frames for a particular clip were reassembled and averaged; this average served as an estimate of the "true" engagement score given by that same labeler when viewing that video clip as described in "approach (2)" above. We found that, with respect to the true engagement scores, the estimated scores gave a $\kappa = 0.78$ and a Pearson correlation $r = 0.85$. This accuracy is quite high and suggests that most of the information about the appearance of engagement is contained in the static pixels, not the motion *per se*.

We also examined the video clips in which the reconstructed engagement scores differed the most from the true scores. In particular, we ranked the 120 labeled video clips in decreasing order of absolute deviation of the estimated label (by averaging the frame-based labels) from the "true" label given to the video clip viewed as a whole. We then examined these clips and attempted to explain the discrepancy: In the first clip (greatest absolute deviation), the subject was swaying her head

from side to side as if listening to music (although she was not). It is likely that the coder treated this as non-engaged behavior. This behavior may be difficult to capture from static frame judgments. However, it was also an anomalous case.

In the second clip, the subject turned his head to the side to look at the experimenter, who was talking to him for several seconds. In the frame-level judgments, this was perceived as off-task, and hence non-engaged behavior; this corresponds to the instructions given to the coders that they rate engagement under the assumption that the subject should always be looking towards the iPad. For the video clip label, however, the coder judged the student to be highly engaged because he was intently listening to the experimenter. This is an example of inconsistency on the part of the coder as to what constitutes engagement and does not necessarily indicate a problem with splitting the clips into frames.

Finally, in several clips the subjects sometimes shifted their eye gaze downward to look at the bottom of the iPad screen. At a frame level, it was difficult to distinguish the subject looking at the bottom of the iPad from the subject looking to his/her own lap or even closing his/her eyes, both of which would be considered non-engagement. From video, it was easier to distinguish these behaviors from the context. However, these downward gaze events were rare and can be effectively filtered out by simple averaging.

In spite of these problems, the relatively high accuracy of estimating video-based labels from frame-based labels suggests an approach for how to construct an *automatic* classifier of engagement: Instead of analyzing video clips as video, break them up into their video frames, and then combine engagement estimates for each frame. We used this approach to label both the HBCU and the UC data for engagement. In the next section, we describe our proposed architecture for automatic engagement recognition based on this frame-by-frame design.

3 AUTOMATIC RECOGNITION ARCHITECTURES

Based on the finding from Section 2.4 that video clip-based labels can be estimated with high fidelity simply by averaging frame-based labels, we focus our study on **frame-by-frame** recognition of student engagement. This means that many techniques developed for emotion and facial action unit classification can be applied to the engagement recognition problem. In this paper we proposed a 3-stage pipeline.

1) Face registration: the face and facial landmark (eyes, nose, and mouth) positions are localized automatically in the image; the face box coordinates are computed; and the face patch is cropped from the image [35]. We experimented with 36×36 and 48×48 pixel face resolution.
2) The cropped face patch is classified by four binary classifiers, one for each engagement category $l \in \{1, 2, 3, 4\}$.
3) The outputs of the binary classifiers are fed to a regressor to estimate the image's engagement level.

Stage (1) is standard for automatic face analysis, and our particular approach is described in [35]. Stage (2) is discussed in the next subsection, and stage (3) is discussed in Section 3.11. This architecture is reminiscent of an automated head pose estimation system we developed previously [57], which combines the outputs of multiple binary classifiers to form a real-valued judgment.

3.1 Binary classification

We trained 4 binary classifiers of engagement – one for each of the 4 levels described in Section 2.1. The task of each of these classifiers is to discriminate an image (or video frame) that belongs to engagement level l from an image that belongs to some other engagement level $l' \neq l$. We call these detectors *1-v-other*, *2-v-other*, etc. We compared three commonly used and demonstrably effective feature type + classifier combinations from the automatic facial expression recognition literature:

- GentleBoost with Box Filter features (**Boost(BF)**): this is the approach popularized Viola and Jones in [53] for face detection.
- Support vector machines with Gabor features (**SVM(Gabor)**): this approach has achieved some of the highest accuracies in the literature for facial action and basic emotion classification [35].
- Multinomial logistic regression with expression outputs from the Computer Expression Recognition Toolbox [35] (**MLR(CERT)**): here, we attempt to harness an existing automated system for facial expression analysis to train engagement classifiers.

Our goal is not to judge the effectiveness of each feature type (or each learning method) in isolation, but rather to assess the effectiveness of these state-of-the-art computer vision architectures for a novel vision task. As relatively little research has yet examined how to recognize the emotional states specific to students in real learning environments, it is an open question how well these methods would perform for engagement recognition. We describe each approach in more detail below.

3.1.1 Boost(BF)

Box Filter (BF) features measure differences in average pixel intensity between neighboring rectangular regions of an image. They have been shown to be highly effective for automatic face detection [53] as well as smile detection [56]. For example, for detecting faces, a 2-rectangle Box Filter can capture the fact that the eye region of the face is typically darker than the upper cheeks. At runtime, BF features are fast to extract using the "integral image" technique [53]. At training time, however, the number of BF features relative to the image resolution is very high compared to other image representations (e.g., a Gabor decomposition), which can lead to overfitting. BF features are typically combined with a boosted classifier such as Adaboost [21] or GentleBoost (Boost) [22],

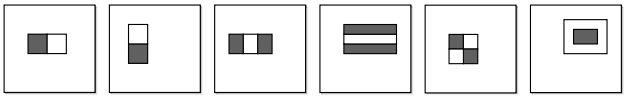


Fig. 3. Box Filter (BF) features, sometimes known as Haar-like wavelet filters, that were used in the study.

which performs both feature selection during training and actual classification at run-time. In our GentleBoost implementation, each weak learner consists of a non-parametric regressor smoothed with a Gaussian kernel of bandwidth σ, to estimate the log-likelihood ratio of the class label given the feature value. Each GentleBoost classifier was trained for 100 boosting rounds. For the features, we included 6 types of Box Filters in total, comprising two-, three-, and four-rectangle features similar to those used in [53], and an additional two-rectangle "center-surround" feature (see Figure 3). At a face image resolution of 48×48 pixels, there were 5397601 BF features; at a face resolution of 36×36 pixels, there were 1683109 features.

3.1.2 SVM(Gabor)

Gabor Energy Filters [44] are bandpass filters with a tunable spatial orientation and frequency. They model the complex cells of the primate's visual cortex. When applied to images, they respond to edges at particular orientations, e.g., horizontal edges due to wrinkling of the forehead, or diagonal edges due to "crow's feet" around the eyes. Gabor Energy Filters have a proven record in a wide variety of face processing applications, including face recognition [31] and facial expression recognition [35]. In machine learning applications Gabor features are often classified by a soft-margin linear support vector machine (SVM) with parameter C specifying how much misclassified training examples should penalize the objective function. In our implementation, we applied a "bank" of 40 Gabor Energy Filters consisting of 8 orientations (spaced at 22.5$\deg$ intervals) and 5 spatial frequencies ranging from 2 to 32 cycles per face. The total number of Gabor features is $N \times N \times 8 \times 5$, where N is the face image width in pixels.

3.1.3 MLR(CERT)

The Facial Action Coding System [17] is a comprehensive framework for objectively describing facial expression in terms of Action Units, which measure the intensity of over 40 distinct facial muscles. Manual FACS coding has previously been used to study student engagement and other emotions relevant to automated teaching [28], [42]. In our study, since we are interested in automatic engagement recognition, we employ the Computer Expression Recognition Toolbox (CERT), which is a software tool developed by our laboratory to estimate facial action intensities automatically [35]. Although the accuracies of the individual facial action classifiers vary, we have found CERT to be useful for a variety of facial analysis tasks, including the discrimination of real from faked pain [34], driver fatigue detection [54], and estimation of students' perception of curriculum difficulty [55]. CERT outputs intensity estimates of 20 facial actions as well as the 3-D pose of the head (yaw, pitch, and roll). For engagement recognition we classify the CERT outputs using multinomial logistic regression (MLR), trained with an L_2 regularizer on the weight vector of strength α. We use the absolute value of the yaw, pitch, and roll to provide invariance to the direction of the pose change. Since we are interested in real-time systems that can operate without baselining the detector to a particular subject, we use the raw CERT outputs (i.e., we do not z-score the outputs) in our experiments.

Internally, CERT uses the SVM(Gabor) approach described above. Since CERT was trained on hundreds to thousands of subjects (depending on the particular output channel), which is substantially higher than the number of subjects collected for this study, it is possible that CERT's outputs will provide an identity-independent representation of the students' faces, which may boost generalization performance.

3.2 Data selection

We started with a pool of 13584 frames from the HBCU dataset. We then applied the following procedure to select training and testing data for each binary classifier to distinguish l-v-other:

1) If the minimum and maximum label given to an image differed by more than 1 (e.g., one labeler assigns a label of 1 and another assigns a label of 3), then the image was discarded. This reduced the pool from 13584 to 9796 images.
2) If the automatic face detector (from CERT [35]) failed to detect a face, or if the largest detected face was less than 36 pixels wide (usually indicative of a erroneous face detection), the image was discarded. This reduced the pool from 9796 to 7785 images.
3) For each of the labeled images, we considered the set of all labels given to that image by all the labelers. If any labeler marked the frame as X (no face, or very unclear), then the image was discarded. This reduced the pool from 7785 to 7574 images.
4) Otherwise, the "ground truth" label for each image was computed by rounding the average label for that image to the nearest integer (e.g., 2.4 rounds to 2; 2.5 rounds to 3). If the rounded label equalled l, then that image was considered a positive example for the l-v-other classifier's training set; otherwise, it was considered a negative example.

In total there were 7574 frames from the HBCU dataset and 16711 from the UC dataset selected using this approach. The distributions of engagement in HBCU were 6.03%, 9.72%, 46.28%, and 37.97% for engagement levels 1 through 4, respectively. For UC, they were 5.37%, 8.46%, 42.31%, and 43.85%, respectively.

3.3 Cross-validation

We used 4-fold subject-independent cross-validation to measure the accuracy of each trained binary classifier. The set of all labeled frames was partitioned into 4 folds such that no subject appeared in more than one fold; hence, the cross-validation estimate of performance gives a sense of how well the classifier would perform on a novel subject on which the classifier was not trained.

3.4 Accuracy metrics

We use the 2AFC [60], [40] metric to measure accuracy, which expresses the probability of correctly discriminating a positive example from a negative example in a 2-alternative forced choice classification task. The 2AFC is an unbiased estimate of the area under the Receiver Operating Characteristics curve, which is commonly used in the facial expression recognition literature (e.g., [37]). A 2AFC value of 1 indicates perfect discrimination, whereas 0.5 indicates that the classifier is "at chance". In addition, we also compute Cohen's κ with quadratic weighting. To compare the machine's accuracy to inter-human accuracy, we computed the 2AFC and κ for human labelers as well, using the same image selection criteria as described in Section 3.2. When computing κ for the automatic binary classifiers, we optimized κ over all possible thresholds of the detector's outputs.

3.5 Hyperparameter selection

Each of the classifiers listed above has a hyperparameter associated with it (either σ, C, or α). The choice of hyperparameter can impact the test accuracy substantially, and it is a common pitfall to give an overly optimistic estimate of a classifier's accuracy by manually tuning the hyperparameter based on the *test* set performance. To avoid this pitfall, we instead optimize the hyperparameters using only the *training* set by further dividing each training set into 4 subject-independent *inner cross-validation folds* in a double cross-validation paradigm. We selected hyperparameters from the following sets of values: $\sigma \in \{10^{-2}, 10^{-1.5}, \ldots, 10^{0}\}$, $C \in \{0.1, 0.5, 2.5, 12.5, 62.5, 312.5\}$, and $\alpha \in \{10^{-5}, 10^{-4}, \ldots, 10^{+5}\}$.

3.6 Results: Binary Classification

Classification results are shown in Table 1 for cropped face resolution of 48×48 pixels. Each cell reports the accuracy (2AFC) averaged over 4 cross-validation folds, along with standard deviation in parentheses. Accuracies at 36×36 pixel resolution were very slightly lower. All results are for subject-independent classification.

From the upper part of Table 1, we see that the binary classification accuracy given by the machine classifiers is very similar to inter-human accuracy. All of the three architectures tested delivered similar performance averaged across the four tasks (1-v-other, 2-v-other, etc.). However, MLR(CERT) performed worse on 1-v-other than the other classifiers. As we discuss in Section 4, many images labeled as Engagement $= 1$ exhibit eye closure. It is possible that CERT's eye closure detector is relatively inaccurate, and in comparison the Boost(BF) and SVM(Gabor) approaches are able to learn an accurate eye closure detector from the training data themselves. On the other hand, CERT performs better than the other approaches for 4-v-other. As described in Section 4, Engagement $= 4$ can be discriminated using pose information. Here, CERT may have an advantage because CERT's pose detector was trained on tens of thousands of subjects.

Since inter-coder reliability is commonly reported using Cohen's κ, we report those values as well, both for HBCU and UC datasets, in Table 2. The "Avg" κ is the mean of the κ values for the 4 binary classifiers. As in Table 1, both the SVM(Gabor) and BF(Boost) classifiers

Accuracy 2AFC (and std. dev.) – train on HBCU, test on HBCU

	Classifier			
Task	**MLR(CERT)**	**Boost(BF)**	**SVM(Gabor)**	**Human**
1-v-other	0.862 (0.061)	0.965 (0.012)	0.914 (0.031)	0.909 (0.021)
2-v-other	0.721 (0.130)	0.709 (0.130)	0.711 (0.038)	0.620 (0.143)
3-v-other	0.574 (0.045)	0.607 (0.065)	0.630 (0.075)	0.606 (0.070)
4-v-other	0.697 (0.076)	0.632 (0.111)	0.660 (0.127)	0.650 (0.068)
Avg	0.714	0.728	0.729	0.696

Accuracy 2AFC (and std. dev.) – train on HBCU, test on UC

	Classifier		
Task	**MLR(CERT)**	**Boost(BF)**	**SVM(Gabor)**
1-v-other	0.782 (0.120)	0.845 (0.111)	0.831 (0.091)
2-v-other	0.682 (0.095)	0.597 (0.102)	0.668 (0.067)
3-v-other	0.507 (0.058)	0.464 (0.063)	0.570 (0.055)
4-v-other	0.613 (0.108)	0.469 (0.153)	0.697 (0.041)
Avg	0.646	0.594	0.691

TABLE 1
Top: Subject-independent, within-dataset (HBCU) engagement recognition accuracy (2AFC metric) for each engagement level $l \in \{1, 2, 3, 4\}$ using each of the three classification architectures, along with inter-human classification accuracy. **Bottom**: Engagement recognition accuracy on a different dataset (UC) not used for training.

Cohen's κ (and std. dev.) – train on HBCU, test on HBCU

	Classifier			
Task	**MLR(CERT)**	**Boost(BF)**	**SVM(Gabor)**	**Human**
1-v-other	0.393 (0.094)	0.662 (0.060)	0.528 (0.167)	0.629 (0.245)
2-v-other	0.254 (0.209)	0.246 (0.164)	0.222 (0.132)	0.272 (0.260)
3-v-other	0.154 (0.063)	0.193 (0.090)	0.213 (0.109)	0.209 (0.154)
4-v-other	0.301 (0.098)	0.214 (0.119)	0.261 (0.135)	0.256 (0.109)
Avg	0.275	0.329	0.306	0.341

Cohen's κ (and std. dev.) – train on HBCU, test on UC

	Classifier		
Task	**MLR(CERT)**	**Boost(BF)**	**SVM(Gabor)**
1-v-other	0.329 (0.236)	0.400 (0.258)	0.414 (0.261)
2-v-other	0.123 (0.100)	0.068 (0.040)	0.154 (0.134)
3-v-other	0.078 (0.049)	0.027 (0.022)	0.096 (0.054)
4-v-other	0.137 (0.123)	0.063 (0.058)	0.260 (0.125)
Avg	0.167	0.140	0.231

TABLE 2
Similar to Table 1, but shows Cohen's κ instead of 2AFC. **Top**: HBCU test set. **Bottom**: UC generalization set.

demonstrate performance that is close to inter-human accuracy.

Overall we find the results encouraging that machine classification of engagement can reach inter-human levels of accuracy.

3.7 Generalization to a different dataset

A well-known issue for contemporary face classifiers is to generalize to people of a different race from the people in the training set; in particular, modern face detectors often have difficulty detecting people with dark skin [56]. For our study, we collected data both at HBCU, where all the subjects were African-American, as well as UC, where all the subjects were either Asian-American or Caucasian-American. This gives us the opportunity to assess how well a classifier trained on one dataset generalizes to the other. Here, we measure performance of the binary classifiers described above that were trained on HBCU when classifying subjects from UC.

Results are shown in Tables 1 and 2 (**bottom**) for each classification method. The most robust classifier was SVM(Gabor): average 2AFC fell only slightly from 0.729 to 0.691, and average κ fell from 0.306 to 0.231. Interestingly, the MLR(CERT) architecture was not particularly robust to the change in population, despite being trained on a much larger number of subjects. It is possible that the head pose features that are measured by CERT and are useful for the HBCU dataset do not generalize to the UC dataset. Between the Boost(BF) and SVM(Gabor) approaches, it is possible that the larger number of BF features compared to Gabor features led to overfitting – the Boost(BF) classifiers generalized well to subjects within the same population, but not to subjects of a different population.

3.8 Discrimination of extreme emotion states

In addition to the l-v-rest classification results described above, we also assess the accuracy of the SVM(Gabor) classifier on the task of discriminating between a student who is very engaged (i.e., Engagement $= 4$) from a student who is very non-engaged (i.e., Engagement $= 1$). On this binary task, the accuracy (2AFC) on the HBCU test set was 0.9280. On the UC generalization set, it was 0.7979.

3.9 Effect of data selection procedures

As described in Section 3.2, we excluded images on which there is large label disagreement (step 1). It is conceivable that this could bias the results to be too optimistic because the "harder" images might be ones on which labelers tend to disagree. In a supplementary analysis using just the first training/testing fold for evaluation, we compared SVM(Gabor) classifiers on image sets created *without* excluding images with high disagreement (which resulted in 10409 images in which the face was detected, instead of 7574), to SVM(Gabor) classifiers trained *with* excluding those images. Results were very similar: the average accuracy (2AFC) over all 4 binary engagement classifiers was 0.7632 after excluding the images with high disagreement, and just slightly lower at 0.7570 without those images. This suggests that the larger number of images available for training can compensate for the noisier labels.

Confusion matrix for binary classifiers (HBCU)

	$\mathbf{E=1}$	$\mathbf{E=2}$	$\mathbf{E=3}$	$\mathbf{E=4}$
1-v-other	0.8434	0.3780	0.1014	0.1281
2-v-other	0.4096	0.6525	0.2852	0.2277
3-v-other	0.3849	0.4802	0.7038	0.5028
4-v-other	0.2857	0.3658	0.4961	0.7365

Confusion matrix for binary classifiers (UC)

	$\mathbf{E=1}$	$\mathbf{E=2}$	$\mathbf{E=3}$	$\mathbf{E=4}$
1-v-other	0.7780	0.4729	0.2386	0.2643
2-v-other	0.3603	0.6054	0.3650	0.2646
3-v-other	0.3878	0.4303	0.5482	0.3829
4-v-other	0.2786	0.3674	0.4448	0.7323

TABLE 3
Confusion matrices for the binary SVM(Gabor) classifiers. Each cell is the probability that the classifier l-v-other will classify an image, whose "true" engagement is given by $E = l'$, as engagement l. **Top**: results for HBCU test set. **Bottom**: results for UC generalization set.

3.10 Confusion matrices

An important question when developing automated classifiers is what kinds of mistakes the system makes. For example, does the 1-v-rest classifier ever believe that an image, whose true engagement label is 4, is actually a 1? To answer this question, we must first select a threshold on the real-valued output of each classifier so that it can make a binary decision. Here, we choose the threshold τ to maximize the *balanced error rate* on each test fold, which we define as the average of the false positive rate and the false negative rate. If the l-v-rest classifier's real-valued output on an image is greater than τ, then the classifier decides the image has engagement level l; otherwise, it decides the image has some engagement level *not* l. Using this threshold selection procedure, and averaging results across folds, we computed the confusion matrices on both the HBCU and UC datasets of the SVM(Gabor) engagement classifiers; the matrices are shown in Table 3. Each cell gives the probability that the binary classifier l-v-other will classify an image, whose "true" engagement (the rounded average label over all labels given to a particular image) is given by $E = l'$, as engagement l. Note that neither the rows nor the columns sum to 1 – this is natural because the classifiers are binary, not 4-way.

As expected, the matrix diagonals dominate over all other values in the rows, which means that each l-v-rest classifier is most likely to respond to an image whose true engagement level is l. However, we also observe that the binary classifiers sometimes make "egregious mistakes". For example, the 3-v-other classifier on the

Confusion matrix of MLR regressor (HBCU)

	$\mathbf{D=1}$	$\mathbf{D=2}$	$\mathbf{D=3}$	$\mathbf{D=4}$
$\mathbf{E=1}$	0.5961	0.1735	0.1258	0.1047
$\mathbf{E=2}$	0.2039	0.3634	0.3521	0.0807
$\mathbf{E=3}$	0.0313	0.1511	0.4669	0.3507
$\mathbf{E=4}$	0.0379	0.0621	0.3971	0.5029

Confusion matrix of MLR regressor (UC)

	$\mathbf{D=1}$	$\mathbf{D=2}$	$\mathbf{D=3}$	$\mathbf{D=4}$
$\mathbf{E=1}$	0.5351	0.1225	0.1963	0.1461
$\mathbf{E=2}$	0.2588	0.2218	0.2552	0.2642
$\mathbf{E=3}$	0.1525	0.1382	0.3950	0.3143
$\mathbf{E=4}$	0.1254	0.0603	0.3898	0.4244

Confusion matrix of human labelers (HBCU)

	$\mathbf{D=1}$	$\mathbf{D=2}$	$\mathbf{D=3}$	$\mathbf{D=4}$
$\mathbf{E=1}$	0.5943	0.2509	0.1445	0.0103
$\mathbf{E=2}$	0.0274	0.4230	0.4910	0.0586
$\mathbf{E=3}$	0.0029	0.1108	0.6924	0.1939
$\mathbf{E=4}$	0.0011	0.0299	0.5456	0.4234

TABLE 4
Confusion matrices specifying the conditional probability $P(D = l \mid E = l')$ of the automatic MLR-based engagement regressor's output D given the "true" engagement label E. **Top**: results on HBCU test set. **Middle**: results on UC generalization set. **Bottom**: results for human labelers on HBCU test set.

HBCU dataset responded positively on 38.49% of images whose true engagement level was 1.

3.11 Regression

After performing binary classification of the input image for each engagement level $l \in \{1, 2, 3, 4\}$, the final stage of the pipeline is to combine the binary classifiers' outputs into a final engagement estimate. For the binary classifiers, we chose the SVM(Gabor) architecture and used two alternative strategies: (1) linear regression for real-valued engagement regression, and (2) multinomial logistic regression (MLR) for 4-way discrete engagement level classification.

3.12 Results

3.12.1 Linear regression

Subject-independent 4-fold cross-validation accuracy, measured using Pearson's correlation r, was 0.50 on the HBCU test set. For comparison, inter-human accuracy on the same task was 0.71. On the UC generalization set, the mean Pearson correlation (over 4 folds) of the regressor was 0.36.

3.12.2 Multinomial logistic regression (MLR)

As an alternative to linear regression, we used multinomial logistic regression (MLR) to obtain discrete-valued engagement outputs in $\{1, 2, 3, 4\}$. The average Cohen's κ (over all 4 folds) of the MLR, when compared with human labels on the HBCU dataset, was 0.42 (std. dev. $= 0.13$); on the UC generalization set, it was 0.23 (std. dev. $= 0.13$).

Table 4 shows confusion matrices both on the HBCU test set and the UC generalization set using MLR as the regressor. The table shows the conditional probability (averaged over 4 folds) that the detector's output D equals l, given that the "true" engagement E (defined as the rounded average engagement label over all human labelers who labeled that image) equals l'. For example, on the HBCU dataset, the probability that the engagement regressor outputs $D = 1$, given that the true engagement was $E = 1$, is 0.5961. The bottom of the table shows the analogous confusion matrix for human labelers. Overall, the confusion matrix of the automated MLR regressor and the matrix for humans are similar. Note, however, that the automated regressor sometimes makes "egregious" mistakes, e.g., mis-classifying images whose true engagement is 1 as belonging to engagement category 4 ($P(D = 4 \mid E = 1) = 0.1047$ for HBCU).

Finally, on the task of discriminating $E = 1$ from $E = 4$ (similar to Section 3.8), the accuracy of the MLR was $\kappa = 0.72$; for human labelers on this task, $\kappa = 0.96$.

4 Reverse-engineering the labelers

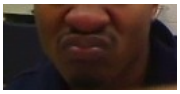
Given that our goal in this project is to recognize student engagement as perceived by an external observer, it is instructive to analyze how the human labelers formed their judgments. We can use the weights assigned to the CERT features that were learned by the MLR(CERT) classifiers to assess quantitatively how the human labelers judged engagement – if the MLR weight assigned to AU 45 (eye closure) had a large magnitude, for example, then that would suggest that eye closure was an important factor in how humans labeled the dataset on which that MLR classifier was trained. In particular, we examined the MLR weights of the 4-v-other MLR(CERT) classifier. Prior to training, the training dataset was first normalized to have unit variance for each feature so that all features had the same scale. After training the MLR, we selected the 5 MLR weights with the highest magnitude; results are shown in Figure 4.

The most discriminating feature was the absolute value of roll (in-plane rotation of the face), with which Engagement $= 4$ was negatively associated (weight of -0.5659). It is possible that the hand-resting-on-hand that is prominent for Engagement $= 2$ also induces roll in the head, and that the MLR(CERT) classifier learned this trend. The second most discriminating facial action was Action Unit 10 (upper lip raiser), which was positively correlated with Engagement $= 4$. However, this correlation could potentially be spurious, as there were many moments when the learners exhibited hand-on-mouth gestures that may have corrupted the AU 10 estimates. Such gestures have been recognized as an important occlusion in automated teaching settings [38].

AU 1 (inner brow raiser), AU 45 (eye closure), and the absolute value of pitch (tilting of the head up and down) were also negatively correlated with Engagement $= 4$. AU 1 has previously been reported to correlate with


AU 1


AU 10


AU 10 (from HBCU)

CERT feature	Weight
abs(Roll)	-0.5659
AU 10	$+0.5089$
AU 1	-0.4430
AU 45 (eye closure)	-0.2851
abs(Pitch)	-0.2644

Fig. 4. Weights associated with different Action Units (AUs) and head pose coordinates to discriminate Engagement $= 4$ from Engagement $\neq 4$, along with examples of AUs 1 and 10. Pictures courtesy of Carnegie Mellon University's Automatic Face Analysis group, http://www.cs.cmu.edu/~face/facs.htm.

Correlations of Engagement with Test Scores		
	Pre-test	**Post-test**
Human labelers		
Mean engagement label	0.52^*	0.37
$P(\text{Engagement} = 1)$	-0.39	-0.22
$P(\text{Engagement} = 2)$	-0.32	-0.18
$P(\text{Engagement} = 3)$	-0.34	-0.40
$P(\text{Engagement} = 4)$	0.57^*	0.47^*
Automatic classifier		
$P(\text{Engagement} = 4)$	0.64^*	0.27

TABLE 5
Engagement statistics that correlated with either pre-test or post-test performance. $P(\text{Engagement} = l)$ denotes the fraction of video frames in which a subject's engagement level was estimated to be l. Correlations with a * are statistically significant ($p < 0.05$, 2-tailed).

students' self-report of frustration [10], but not engagement. The negative correlations with AU 45 and pitch are intuitive – they are suggestive that the student has tuned out (or even fallen asleep), or is looking down away from the screen.

5 Correlation with Test Scores

In this section we investigate the correlation between human and automatic perceptions of engagement with student *test performance* and *learning*. We show results for Pearson correlation. Results for Spearman rank correlation were generally lower and are not reported.

5.1 Test performance

5.1.1 Human labels

We first compared *human* judgments of engagement with test performance by computing the mean engagement label over all labeled frames for each subject in the HBCU dataset, and then correlating these mean engagement labels with pre-test and post-test scores (see Table 5). The Pearson correlation between engagement and pre-test was $r = 0.52$ ($p = 0.0167$, 2-tailed) and between engagement and post-test was $r = 0.37$ ($p = 0.1027$, dof=19, 2-tailed).

We also examined which of the 4 engagement levels was most predictive of task performance by correlating the fraction of frames labeled as Engagement $= 1$, Engagement $= 2$, etc., with student test performance. Only Engagement $= 4$ was positively correlated with pre-test ($r = 0.57, p = 0.0066$, dof=19, 2-tailed) and post-test ($r = 0.47, p = 0.0324$, dof=19, 2-tailed) performance. In fact, the fraction of frames for which a student appeared to be in engagement level 4 (which we denote as $P(\text{Engagement} = 4)$) was a better predictor than the mean engagement predictor described above. All the other engagement levels $l < 4$ were negatively (though non-significantly) correlated with test performance, suggesting that Engagement $= 4$ is the only "positive" engagement state.

For comparison, the correlation between students' pre-test and post-test scores was $r = 0.44$ ($p = 0.0471$, dof=19, 2-tailed), which is slightly (though not statistically significantly) less than the correlation between $P(\text{Engagement} = 4)$ and post-test. In other words, human perceptions of student engagement were just as good of a predictor of post-test performance as the student's pre-test score. A partial correlation between $P(\text{Engagement} = 4)$ and post-test, given pre-test score, gave $r = 0.29$ ($p = 0.2073$, dof=19, 2-tailed).

Finally, it is worth noting that another interpretation of the correlation between engagement and pre-test is that a student's pre-test score is predictive of his/her engagement level during the subsequent learning session.

5.1.2 Automatic estimates

We also computed the correlation between *automatic* judgments of engagement and student pre- and post-test performance. Since the best predictor of test performance from human judgments was from the fraction of frames labeled as Engagement $= 4$, we focused on the output of the 4-v-other classifier. In particular, we correlated the fraction of frames over each subject's entire video session that the 4-v-other detector predicted to be a "positive" frame by thresholding with τ, where τ is the median detector output over all subjects' frames. In other words, frames on which the detector's output exceeded τ was considered to be a "positive" frame for engagement level 4. The correlation with this automatic $P(\text{Engagement} = 4)$ predictor and pre-test performance was 0.64 ($p = 0.0023$, dof=19, 2-tailed); for post-test performance, it was $r = 0.27$ ($p = 0.2436$, dof=19, 2-tailed). This is the same pattern of correlations as in Section 5.1.1 – engagement was more predictive of pre-test than of post-test.

5.2 Learning

In addition to raw test performance, we also examined correlations between engagement and learning. The average difference between the post-test and pre-test scores

(across 21 subjects) was 2.81 sets, which was statistically significant ($t(20) = 4.3746$, $p = 0.0002$, 2-tailed), and which suggests that students were learning. However, we did not find significant correlations between engagement and learning, either using human labels or automatically estimated engagement labels.

5.3 Discussion

The correlation between engagement and pre-and post-test scores is of interest. Particularly telling is that post-test performance can be predicted just as accurately by looking at students' faces during learning ($r = 0.47$) as by looking at their pre-test scores (0.44). These results are consistent with [15], who found a positive correlation between "student energy" (valence) and math pre-test score as well as a positive correlation between a student being "on-task" and math post-test scores,

The lack of correlation between engagement and learning was somewhat disappointing but we believe it is an important clue for planning future research. The more engaged students have higher pre-test scores, which suggests there may be ceiling effects. It is possible, for example, that improving a test score from 10 to 11 is more difficult than improving from 1 to 2. We explored this hypothesis by optimizing the correlation between engagement and learning gains over different monotonic transformations both of engagement and of test scores. In particular, by searching over all monotonic mappings from $\{1, \ldots, 4\}$ into $\{0, \ldots, 4\}$ for engagement, and from $\{0, \ldots, 12\}$ (the range of test scores observed in our experiment) into $\{0, \ldots, 20\}$ for test scores, we identified a transformation that gave moderate ($r = 0.44$, $p = 0.0458$, dof=19, 2-tailed) but statistically significant correlations between learning and engagement. This non-linear monotonic transformation was effectively "undoing" the ceiling effect, weighting learning gains more heavily that started from larger pre-test baselines. However, we tried a large number of monotonic transformations, and thus the statistical significance of this analysis should be taken with a grain of salt. We also note that the correlation between engagement and learning might become significant if the number of subjects were increased.

Finally, and most importantly, in short term laboratory studies such as ours, most students are quite motivated and engaged. Indeed, while examining the videos, we rarely found periods of prolonged non-engagement. This is obviously different from classroom situations in which some students are consistently engaged and some students consistently disengaged across days, months, and years. Future work would benefit from focusing on long-term learning situations where variance in engagement is more likely to be observed and the effect of engagement on learning is more likely to become apparent.

6 Conclusions

Increasing student engagement has emerged as a key challenge for teachers, researchers, and educational institutions. Many of the current tools used to measure engagement – such as self-reports, teacher introspective evaluations, and checklists – are cumbersome, lack the temporal resolution needed to understand the interplay between engagement and learning, and in some cases capture student compliance rather than engagement.

In this paper we explored the development of real-time automated recognition of engagement from students' facial expressions. The motivating intuition was that teachers constantly evaluate the level of their students' engagement, and facial expressions play a key role in such evaluations. Thus, understanding and automating the process of how people judge student engagement from the face could have important applications.

Our work extends prior research on engagement recognition using computer vision [42], [29], [10], [11] and is arguably the most thorough study on this topic to date: We collected a dataset of student facial expressions while performing a cognitive training task. We experimented with multiple approaches for human observers to assess student engagement. We found that interobserver reliability is maximized when the length of the observed clips is approximately 10 seconds. Shorter clips do not provide enough context and reliability suffers. Longer clips tend to be harder to evaluate because they often mix different levels of engagement. When discriminating low v. high levels of engagement, interobserver reliability was high (Cohen's $\kappa = 0.96$). We also found that the engagement judgments of 10-second clips could be reliably approximated (Pearson $r = 0.85$) by averaging single frame judgments over the 10 seconds. This indicates that static expressions contain the bulk of the information observers use to assess student engagement. We found that observers rely on head pose, and elementary facial actions like brow raise, eye closure, and upper lip raise to make their judgments.

Our results suggest that machine learning methods could be used to develop a real-time automatic engagement detector with comparable accuracy to that of human observers. We showed that both human and automatic engagement judgments correlate with task performance. In particular, student post-test performance was predicted just as accurately (and statistically significantly) by observing the face of the student during learning ($r = 0.47$) as from the pre-test scores ($r = 0.44$). We failed to find significant correlations between perceived engagement and learning. However, a-posteriori statistical analysis suggests this may be due to ceiling effects and a fundamental limitation of short-term laboratory studies such as ours. In such studies, most students tend usually to be quite engaged, which is quite different from the long-term engagement or disengagement found in classrooms. This points to the importance of long-term studies that approximate the classroom ecology in which some students are engaged and others are chronically disengaged for days, months, and years.

While the progress made here is modest, it reinforces the idea that automatic recognition of student engage-

ment is possible and could potentially revolutionize education as we know it. For example, using computer vision systems, a set of low-cost, high-resolution cameras could monitor engagement levels of entire classrooms, without the need for self-report or questionnaires. The temporal resolution of the technology could help understand when and why students get disengaged, and perhaps to take action before it is too late. Web-based teachers could obtain real-time statistics of the level of engagement of their students across the globe. Educational videos could be improved based on the aggregate engagement signals provided by the viewers. Such signals would indicate not only whether a video induces high or low engagement, but most importantly, which parts of the videos do so. Our work underlines the importance of focusing on long-term field studies in real-life classroom environments. Collecting data in such environments is critical to train more reliable and ecologically valid engagement recognition systems. More importantly, sustained, long-term studies in actual classrooms are needed to gain a better understanding of the interplay between engagement and learning in real life.

APPENDIX: INTER-HUMAN ACCURACY

The classifiers in this paper were trained and evaluated on the average label, across all human labelers, given to each image. To enable a fair comparison between inter-human accuracy and machine-human accuracy, we assess accuracy (using Cohen's κ, Pearson's r, or 2AFC) of each human labeler by comparing his/her labels to the average label, over all other labelers, given to each image. We then average the individual accuracy scores over all labelers and report this as the inter-human reliability. Note that this "leave-one-labeler-out" agreement is typically higher than the average pair-wise agreement.

REFERENCES

[1] LearningRx, 2012. www.learningrx.com.
[2] Scientific Learning, 2012. www.scilearn.com.
[3] A. Anderson, S. Christenson, M. Sinclair, and C. Lehr. Check and connect: The importance of relationships for promoting engagement with school. *Journal of School Psychology*, 42:95–113, 2004.
[4] J. R. Anderson. Acquisition of cognitive skill. *Psychological Review*, 89(4):369–406, 1982.
[5] I. Arroyo, K. Ferguson, J. Johns, T. Dragon, H. Meheranian, D. Fisher, A. Barto, S. Mahadevan, and B. Woolf. Repairing disengagement with non-invasive interventions. *Proceedings of the 2007 Conference on Artificial Intelligence in Education*, pages 195–202, 2007.
[6] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan. Real time face detection and facial expression recognition: development and applications to human computer interaction. In *Proceedings of the CVPR Workshop on Human-Computer Interaction*, 2003.
[7] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 2006.
[8] J. Beck. Engagement tracing: using response times to model student disengagement. In *Proceedings of the 2005 Conference on Artificial Intelligence in Education*, pages 88–95, 2005.
[9] M. Chaouachi, P. Chalfoun, I. Jraidi, and C. Frasson. Affect and mental engagement: towards adaptability for intelligent systems. In *FLAIRS*, 2010.
[10] S. D'Mello, S. Craig, and A. Graesser. Multimethod assessment of affective experience and expression during deep learning. *International Journal of Learning Technology*, 4(3):165–187, 2009.
[11] S. D'Mello and A. Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20(2):147–187, 2010.
[12] S. D'Mello, B. Lehman, J. Sullins, R. Daigle, R. Combs, K. Vogt, L. Perkins, and A. Graesser. A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, pages 245–254, 2010.
[13] S. D'Mello, R. Picard, and A. Graesser. Towards an affect-sensitive AutoTutor. *IEEE Intelligent Systems, Special Issue on Intelligent Educational Systems*, 22(4):53–61, 2007.
[14] S. K. D'Mello, S. Craig, J. Sullins, and A. Graesser. Predicting affective states expressed through an emote-aloud procedure from AutoTutors mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education*, 16(1):3–28, 2006.
[15] T. Dragon, I. Arroyo, B. Woolf, W. Burleson, R. el Kaliouby, and H. Eydgahi. Viewing student affect and learning through classroom observation and physical sensors. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, pages 29–39, 2008.
[16] J. Dunleavy and P. Milton. What did you do in school today? exploring the concept of student engagement and its implications for teaching and learning in canada. *Canadian Education Association (CEA),*, pages 1–22, 2009.
[17] P. Ekman and W. Friesen. *The Facial Action Coding System: A Technique For The Measurement of Facial Movement*. Consulting Psychologists Press, Inc., San Francisco, CA, 1978.
[18] S. Fairclough and L. Venables. Prediction of subjective states from psychophysiology: a multivariate approach. *Biological psychology*, 71:100–110, 2006.
[19] K. Forbes-Riley and D. Litman. Adapting to multiple affective states in spoken dialogue. In *Proceedings of the Special Interest Group on Discourse and Dialogue*, pages 217–226, 2012.
[20] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1):59–109, 2004.
[21] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.
[22] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28(2):337–407, 2000.
[23] B. Goldberg, R. Sottilare, K. Brawner, and H. Holden. Predicting learner engagement during well-defined and ill-defined computer-based intercultural interactions. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, pages 538–547, 2011.
[24] J. Grafsgaard, R. Fulton, K. Boyer, E. Wiebe, and J. Lester. Multimodal analysis of the implicit affective channel in computer-mediated textual communication. In *Proceedings of the International Conference on Multimodal Interaction*, pages 145–152, 2012.
[25] L. Harris. A phenomenographic investigation of teacher conceptions of student engagement in learning. *The Australian Educational Researcher*, 5(1):57–79, 2008.
[26] J. Johns and B. Woolf. A dynamic mixture model to detect student motivation and proficiency. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI06)*, pages 2–8, 2006.
[27] T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pages 46 – 53, March 2000.
[28] A. Kapoor, S. Mota, and R. Picard. Towards a learning companion that recognizes affect. In *AAAI Fall Symposium*, 2001.
[29] A. Kapoor and R. Picard. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 677–682, 2005.
[30] 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.
[31] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:300–311, 1993.
[32] R. Larson and M. Richards. Boredom in the middle school years: Blaming schools versu blaming students. *American journal of education*, 99:418–443, 1991.
[33] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan.

Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.
[34] G. Littlewort, M. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797–1803, 2009.
[35] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. Computer expression recognition toolbox. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG'11)*, pages 298–305, 2011.
[36] R. Livingstone. *The future in education*. Cambridge University Press, 1941.
[37] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended Cohn-Kanade dataset: A complete dataset for action unit and emotion-specied expression. In *CVPR Workshop on Human-Communicative Behavior*, pages 94–101, 2010.
[38] M. Mahmoud and P. Robinson. Interpreting hand-over-face gestures. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 248–255, 2011.
[39] S. Makeig, M. Westerfield, J. Townsend, T.-P. Jung, E. Courchesne, and T. J. Sejnowski. Functionally independent components of early event-related potentials in a visual spatial attention task. *Philosophical Transactions of the Royal Society: Biological Science*, 354:1135–44, 1999.
[40] S. Mason and A. Weigel. A generic forecast verification framework for administrative purposes. *Monthly Weather Review*, 137:331–349, 2009.
[41] G. Matthews, S. Campbell, S. Falconer, L. Joyner, J. Huggins, and K. Gilliland. Fundamental dimensions of subjective state in performance settings: task engagement, distress, and worry. *Emotion*, 2(4):315–340, 2002.
[42] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser. Facial features for affective state detection in learning environments. In *Proceedings of the 29th Annual Cognitive Science Society*, pages 467–472, 2007.
[43] J. Mostow, A. Hauptmann, L. Chase, and S. Roth. Towards a reading coach that listens: Automated detection of oral reading errors. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI93)*, pages 392–397, 1993.
[44] J. R. Movellan. Tutorial on gabor filters. Technical report, MPLab Tutorials, UCSD MPLab, 2005.
[45] H. O'Brien and E. Toms. 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, 2010.
[46] J. Ocumpaugh, R. S. Baker, and M. M. T. Rodrigo. Baker-Rodrigo observation method protocol 1.0 training manual. Technical report, EdLab, Manila, Philippines, 2012.
[47] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *Systems, Man, and Cybernetics – Part B: Cybernetics*, 36(2), 2006.
[48] J. Parsons and L. Taylor. Student engagement: What do we know and what should we do. Technical report, University of Alberta, 2011.
[49] A. Pope, E. Bogart, and D. Bartolome. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, 40:187–195, 1995.
[50] K. Porayska-Pomsta, M. Mavrikis, S. D'Mello, C. Conati, and R. S. Baker. Knowledge elicitation methods for affect modelling in education. *International Journal on Artificial Intelligence in Education*, 2013.
[51] D. Shernof, M. Csikszentmihalyi, B. Schneider, and E. Shernoff. Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*, 18(2):158–176, 2003.
[52] K. VanLehn, C. Lynch, K. Schultz, J. Shapiro, R. Shelby, and L. Taylor. The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, 15(3):147–204, 2005.
[53] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.
[54] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan. Drowsy driver detection through facial movement analysis. In *Proceedings of the IEEE International Conference on Human-Computer Interaction*, pages 6–18, 2007.
[55] J. Whitehill, M. Bartlett, and J. R. Movellan. Automatic facial expression recognition for intelligent tutoring systems. In *Proceedings of the CVPR 2008 Workshop on Human Communicative Behavior Analysis*, pages 1–6, 2008.
[56] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, 2009.
[57] J. Whitehill and J. Movellan. A discriminative approach to frame-by-frame head pose tracking. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.
[58] J. Whitehill, Z. Serpell, A. Foster, Y.-C. Lin, B. Pearson, M. Bartlett, and J. Movellan. Towards an optimal affect-sensitive instructional system of cognitive skills. In *Computer Vision and Pattern Recognition Workshop on Human Communicative Behavior*, pages 20–25, 2011.
[59] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3):129–164, 2009.
[60] T. Wu, N. Butko, P. Ruvolo, J. Whitehill, M. Bartlett, and J. Movellan. Multilayer architectures for facial action unit recognition. *IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics*, 42(4):1027–1038, 2012.
[61] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

Jacob Whitehill is a researcher in machine learning, computer vision, and their applications to education. He holds a PhD from the University of California, San Diego, a MSc from the University of the Western Cape, and a BS from Stanford. Since 2012 he is a cofounder and research scientist at Emotient.

Zewelanji Serpell is an Associate Professor in the Department of Psychology at Virginia Commonwealth University. Her research focuses on developing and evaluating interventions to improve students' executive functioning and optimize learning in school contexts. She has a BA in Psychology from Clark University in Worcester MA, and a Ph.D. in Developmental Psychology from Howard University in Washington, DC.

Yi-Ching Lin is a Modeling and Simulation Fellowship recipient at Old Dominion University where she is pursuing her PhD in Occupational and Technical Studies in the STEM Education and Professional Studies Department. Her current research centers on the use of systems dynamic modeling to assess the influence of various factors on students' choice of STEM majors. She holds a MS in Psychology from Virginia State, a BS in Psychology from University of Missouri Columbia, and a BS in Chemical Engineering from the National Taipei University of Technology.

Aysha Foster is a social science researcher whose interests include effective learning strategies and mental health for minority youth. She received her MS and PhD in psychology from Virginia State University and a BS in Biology from Prairie View A&M University. She is currently a research coordinator at Virginia Commonwealth University for a project examining the malleability of executive control in elementary students.

Javier Movellan founded the Machine Perception Laboratory at UCSD, where he is a research professor. He is also founder and lead researcher at Emotient. Javier's research interests include machine learning, machine perception, automatic analysis of human behavior, and social robots. His team helped develop the first commercial smile detection system, which was embedded in consumer digital cameras. He also pioneered the development of social robots and their use for early childhood education. Prior to his UCSD position he was a Fulbright Scholar at UC Berkeley, where he received his PhD.