Automatic Facial Expression Recognition for Intelligent Tutoring Systems

Jacob Whitehill, Marian Bartlett, and Javier Movellan
Machine Perception Laboratory
University of California, San Diego
{jake, movellan}@mplab.ucsd.edu, marni@salk.edu

Abstract

This project explores the idea of facial expression for automated feedback in teaching. We show how automatic real-time facial expression recognition can be effectively used to estimate the difficulty level, as perceived by an individual student, of a delivered lecture. We also show that facial expression is predictive of an individual student's preferred rate of curriculum presentation at each moment in time. On a video lecture viewing task, training on less than two minutes of recorded facial expression data and testing on a separate validation set, our system predicted the subjects' self-reported difficulty scores with mean accuracy of 0.42 *(Pearson* R*) and their preferred viewing speeds with mean accuracy of* 0.29*. Our techniques are fully automatic and have potential applications for both intelligent tutoring systems (ITS) and standard classroom environments.*

1. Introduction

One of the fundamental challenges faced by teachers – whether human or robot – is determining how well his/her students are receiving a lecture at any given moment. Each individual student may be content, confused, bored, or excited by the lesson at a particular point in time, and one student's perception of the lecture may not necessarily be shared by his/her peers. While explicit feedback signals to the teacher such as a question or a request to repeat a sentence are useful, they are limited in their effectiveness for several reasons: If a student is confused, he may feel embarrassment in asking a question. If the student is bored, it may be inappropriate to ask the teacher to speed up the rate of presentation. Some research has also shown that students are not always aware of when they need help [1]. Finally, even when students do ask questions, this feedback may, in a sense, come too late – the student may already have missed an important point, and the teacher must spend lesson time to clear up the misunderstanding.

If, instead, the student could provide feedback at an earlier time, perhaps even subconsciously, then moments of frustration, confusion, and even boredom could potentially be avoided. Such feedback is particularly useful for automated tutoring systems. For example, an interactive tutoring system could dynamically adjust the speed of the instruction to increase when the student's understanding is solid and to slow down during an unfamiliar topic.

In this paper we explore one such kind of feedback signal based on automatic recognition of a student's facial expression. Recent advances in the fields of pattern recognition, computer vision, and machine learning have made automatic facial expression recognition in real-time a viable resource for intelligent tutoring systems (ITS). The field of ITS has already begun to make use of this technology, especially for the task of predicting the student's affective state (e.g., [2, 3, 4, 5]). This paper investigates the potential usefulness of automatic expression recognition for two different tasks: (1) measuring the difficulty as perceived by students of a delivered lecture, and (2) determining the preferred speed at which lesson material should be presented. To this end, we conducted a pilot experiment in which subjects viewed a video lecture at an adjustable speed while their facial expressions were recognized automatically and recorded. Using the "difficulty" scores that the subjects report, the correlations between facial expression and difficulty, and between facial expression and preferred viewing speed, can be assessed.

The rest of this paper is organized as follows: In Section 2, we briefly describe the automatic expression recognition system that we employ in our study. Section 3 describes the experiment we perform on human subjects, and Section 4 presents the results. We end with some concluding remarks about facial expression recognition for ITS.

2. Facial Expression Recognition

Facial expression is one of the most powerful and immediate means for humans to communicate their emotions, cognitive states, intentions, and opinions to each other [6]. In recent years, researchers have made considerable progress in developing automatic expressions classifiers [7, 8, 9]. Some expression recognition systems clas-

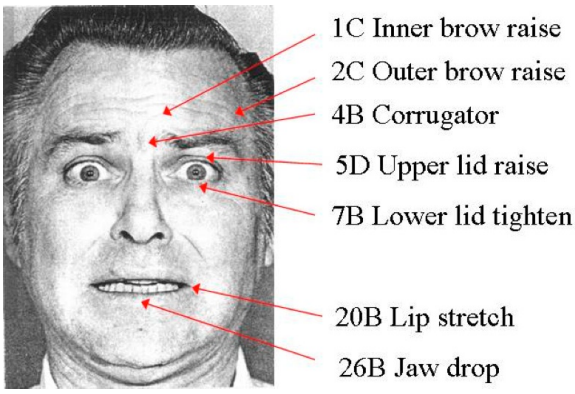

Figure 1. Example of comprehensive FACS coding of a facial expression. The numbers identify the action unit, which approximately corresponds to one facial muscle; the letter (A-E) identifies the level of activation.

sify the face into the set of "prototypical" emotions such as happy, sad, angry, etc. [10]. Others attempt to recognize the individual muscle movements that the face can produce [11] in order to provide an objective description of the face. The best known psychological framework for describing nearly the entirety of facial movements is the Facial Action Coding System (FACS) [12].

2.1. FACS

FACS was developed by Ekman and Friesen as a method to code facial expressions comprehensively and objectively [12]. Trained FACS coders decompose facial expressions in terms of the apparent intensity of 46 component movements, which roughly correspond to individual facial muscles. These elementary movements are called action units (AU) and can be regarded as the "phonemes" of facial expressions. Figure 1 illustrates the FACS coding of a facial expression. The numbers identify the action unit, which approximately corresponds to one facial muscle; the letter (A-E) identifies the level of activation.

2.2. Automatic Facial Expression Recognition

We use the automatic facial expression recognition system presented in [11] for our experiments. This machine learning-based system analyzes each video frame independently. It first finds the face, including the location of the eyes, mouth, and nose for registration, and then employs support vector machines and Gabor energy filters for expression recognition. The version of the system employed here recognizes the following AUs: 1 (inner brow raiser), 2 (outer brow raiser), 4 (brow lowerer), 5 (upper eye lid raiser), 9 (nose wrinkler), 10 (upper lip raiser), 12 (lip corner puller), 14 (dimpler), 15 (lip corner depressor), 17 (chin raiser), 20 (lip stretcher), and 45 (blink), as well as a detector of social Smiles.

3. Experiment

The goal of our experiment was to assess whether there exist significant correlations between certain AUs and the perceived difficulty as well as the preferred viewing speed of a video lecture. To this end, we composed a short composite "lecture" video consisting of seven individual movie clips about a disparate range of topics. The individual clips were excerpts taken from public-domain videos from the Internet. In order, they were:

1. An introductory university physics lecture (46 sec).
2. A university lecture on Sigmund Freud (36 sec).
3. A soundless tutorial on Vedic mathematics (46 sec).
4. A university lecture on philosophy (20 sec).
5. A barely audible sound clip (with a static picture backdrop) of Sigmund Freud (16 sec).
6. A teenage girl speaking quickly while telling a humorous story (21 sec).
7. Another excerpt on physics taken from the same source as the first clip (15 sec).

Representative video frames of all 7 video clips are shown in Figure 2.

3.1. Procedure

Each subject performed the following tasks in order:

1. **Watch the video lecture.** The playback speed could be adjusted continuously by the subject. Facial expression data were recorded.
2. **Take the quiz.** The quiz consisted of 6 questions about specific details of the lecture.
3. **Self-report on the difficulty.** The video lecture was re-played at a fixed speed of 1.0.

For watching the lecture at an adjustable speed we created a special viewing program in which the user can press Up to increase the speed, Down to decrease the speed, and Left to rewind by two seconds. Rewinding the video also set the speed back to the default rate (1.0). The video player was equipped with an automatic pitch equalizer so that, even at high speeds, the lecture audio was reasonably intelligible. Subjects practiced using the speed controls on a separate demo video prior to beginning the actual study. In order to encourage subjects to use their time efficiently and thus to avail themselves of the speed control, we informed them prior to the first viewing that they would take a quiz afterwards, and that their performance on the quiz would be penalized by the amount of time they needed to watch

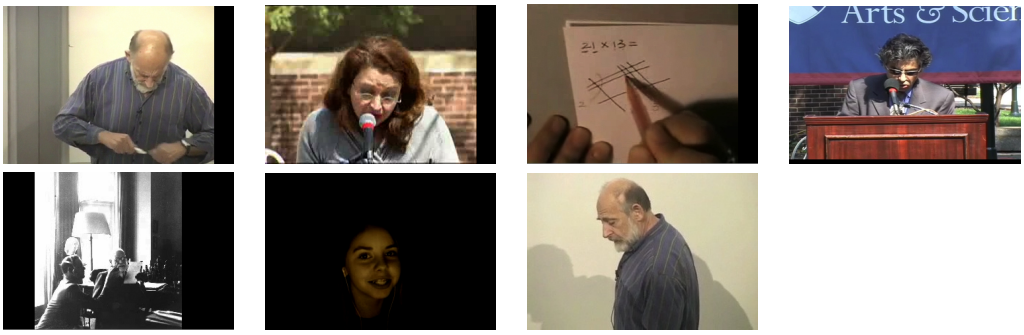

Figure 2. Representative video frames from each of the 7 video clips contained in our "lecture" movie.

the video. We also started a visible, automatic "shut-off" timer when they started watching the lecture to give the impression of additional time pressure. In actuality, the timer provided enough time to watch the whole lecture at normal speed, and the quiz was never graded – these props were meant only to encourage the subjects to modulate the viewing speed efficiently.

While watching the video lecture for the first time, the subject's facial expression data were recorded automatically through a standard Web camera using the automatic face and expression recognition system described in [11]. The experiment was performed in an ordinary office environment inside our laboratory without any special lighting conditions. After watching the video and taking the quiz, subjects were then informed that they would watch the lecture for a second time. During the second viewing, subjects could not change the speed (it was fixed at 1.0), but they instead rated frame-by-frame how difficult they found the movie to be on an integral scale of 0 to 10 using the keyboard (A for "harder", Z for "easier"). This form of continuous audience response labeling was originally developed for consumer research [13]. Subjects were told to consider both acoustic as well as conceptual difficulty when assessing the difficulty of the lecture material. Facial expression information was not collected during the second viewing.

In our experimental design, the fact that subjects adjusted the viewing speed of the lecture video while viewing it may have affected their perception of how difficult the lecture was to understand. Our reason for designing the experiment in this way was to capture both speed control and difficulty information from all subjects. However, we believe that the ability to adjust the speed of the lecture would, if anything, cause the self-reported Difficulty values to be more "flat," thus *increasing* the challenge of the prediction task (predict Difficulty from Expression).

3.2. Human Subjects

Eight subjects (five female, three male) participated in our pilot experiment. Subjects ranged in age from early 20's to mid 30's and were either undergraduate students, graduate students, or administrative or technical staff at our university. Five were native English speakers (American), and three were non-native (one was Northern European, one was Southern European, and one was East Asian). Each subject was paid $15 for his/her participation, which required about 20 minutes in total.

None of the subjects was aware of the purpose of the study or that facial expression data would be captured. Prior to starting the experiment, subjects were informed only that they would be watching a video at a controllable speed and that they would be quizzed afterward. They were not informed of rating the difficulty of the experiment or of watching the video at second time until after the quiz. Subjects were not requested to restrict head movement in any way (though all remained seated throughout the entire video lecture), and the resulting variability in head pose, while presenting no fundamental difficulty for our expression recognition system, may have added some amount of noise. Due to the need to manually adjust the viewing angle of the camera for facial expression recording, it is possible that subjects inferred that their facial behavior would be analyzed.

3.3. Data Collection and Processing

While the subjects watched the video, their faces were analyzed in real-time using the expression recognition system presented in [11]. The output of 12 action unit detectors (AUs 1, 2, 4, 5, 9, 10, 12, 14, 15, 17, 20, 45) as well as the smile detector were time-stamped and saved to disk. The muscle movements to which the above-listed AUs correspond are shown in Table 1. Speed adjustment events (Up, Down, and Rewind) were used to compute an overall Speed data series. A Difficulty data series was likewise computed using the difficulty adjustment keyboard events (A and Z). Since all Expression, Difficulty, and Speed events were timestamped, and since the video player itself timestamped the display time of each video frame, we were able to time-align pairwise the Expression and Difficulty, and

Description of Facial Action Units

AU #	Description
1	Inner brow raiser
2	Outer brow raiser
4	Brow lowerer
5	Upper eye-lid raiser
9	Nose wrinkler
10	Upper lip raiser
12	Lip corner puller
14	Dimpler
15	Lip corner depressor
17	Chin raiser
20	Lip stretcher
45	Blink
Smile	"Social" smile (not part of FACS)

Table 1. List of FACS Action Units (AUs) employed in this study.

Expression and Speed time series, and then analyze them for correlations.

4. Results

We performed correlation analyses between individual AUs and both the Difficulty and Speed time series. We also performed multiple regression over *all* AUs to predict both the Difficulty and Speed time series. Local quadratic regression was employed to smooth the AU values. The smoothing width for each subject was taken as the average length of time for which the user left the Difficulty value unchanged during the second viewing of the video. The exact number of data points in the Expression data series varied between subjects since they required different amounts of time to watch the video, but for all subjects at least 790 data points (approximately 4 per second) were available for calculating correlations.

For each subject there were a number of AUs that were significantly correlated (we required $p < 0.05$) with perceived difficulty, and also a number of AUs correlated with viewing speed. We report the 3 AUs with the highest correlation magnitude for each prediction task (Difficulty, Viewing Speed). Results are shown in Tables 2 and 3.

These results indicate substantial inter-subject variability on which AUs correlated with perceived difficulty, and on which AUs correlated with viewing speed. The only AU which showed both a significant and consistent correlation (though not necessarily in the top 3) with difficulty was AU 45 (blink) – for 6 out of 8 subjects their difficulty labels were negatively correlated with blink, meaning these subjects blinked less during the more difficult sections of video. This finding is consistent with evidence from experimental psychology that blink rate decreases when interest or mental load is high [14, 15]. To our surprise, AU 4 (brow lowerer), which is associated with concentration and consternation, was not consistently positively correlated with difficulty.

Correlations between AUs and Self-reported Difficulty

Subj.	3 AUs Most Correlated with Self-Reported Difficulty	Overall Corr. (R)
1	**4** (+.42), **9** (-.40), **2** (-.35)	0.84
2	**5** (-.34), **15** (-.30), **17** (-.25)	0.73
3	**20** (+.66), **5** (+.45), **45** (-.42)	0.76
4	**20** (-.51), **5** (-.47), **9** (-.47)	0.85
5	**10** (-.31), **12** (-.28), **2** (-.25)	0.60
6	**5** (-.65), **4** (-.55), **15** (-.49)	0.88
7	**17** (-.53), **1** (-.47), **14** (-.43)	0.74
8	**17** (-.22), **5** (+.19), **45** (+.18)	0.56
Avg		**0.75**

Table 2. *Middle column*: The three significant correlations with the highest magnitude between difficulty and AU value for each subject. *Right column*: the overall correlation between predicted and self-reported Difficulty value, when using linear regression over the whole set of AUs for prediction.

Correlations between AUs and Viewing Speed

Subj.	3 AUs Most Correlated with Viewing Speed
1	**9** (+.29), **45** (+.26), **4** (-.24)
2	**17** (+.21), **2** (-.16), **Smile** (+.16)
3	**14** (-.46), **2** (-.44), **1** (-.42)
4	**20** (+.42), **2** (-.37), **17** (-.36)
5	**1** (-.21), **20** (-.20), **15** (-.19)
6	**9** (-.48), **4** (+.40), **15** (+.39)
7	**17** (+.35), **14** (+.34), **Smile** (+.32)
8	**15** (-.53), **17** (-.47), **12** (-.46)

Table 3. The three significant correlations with highest magnitude between preferred viewing speed and AU value for each subject.

4.1. Predicting Difficulty from Expression Data

To assess how much overall signal is available in the AU outputs for predicting self-reported difficulty values, we performed linear regression over all AUs and targeted Difficulty labels as the dependent variable. The correlations between the predicted difficulty values and the self-reported values are shown in right column of Table 2. A graphical representation of the predicted difficulty for Subject 6 is shown in Figure 3. The average correlation between predicted difficulty values and self-reported values of 0.75 suggests that AU outputs are a valuable signal for predicting a student's perception of difficulty. In Section 4.2, we extend this analysis to the case where a Difficulty model is learned from a training set separate from the validation data.

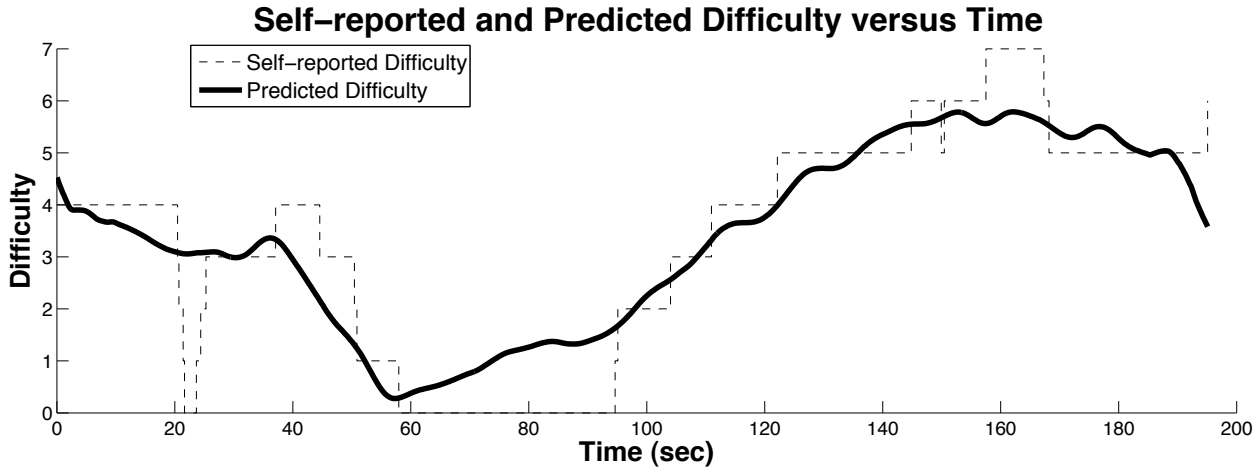

Figure 3. The self-reported difficulty values, and the predicted difficulty values computed using linear regression over all AUs, for Subj. 6.

4.2. Learning to Predict

Given the high inter-subject variability in which AUs correlated with difficulty and with viewing speed, it seems likely that subject-specific models will need to be trained in order for facial expression recognition to be useful for predicting difficulty and viewing speed. We thus trained a linear regression model to predict both Difficulty and Viewing Speed scores for each subject. In our model we regressed over both the AU outputs themselves and their temporal first derivatives. The derivatives might be useful since it is conceivable that sudden changes in expression could be predictive of changes in difficulty and viewing speed. We also performed a variable amount of smoothing, and we introduced a variable amount of time lag into the entire set of captured AU values to account for a possible delay between watching the video and reacting to it with facial expression. The smoothing and lag parameters were optimized using the training data, as explained later in this section.

For assessing the model's ability to learn, we divided the time-aligned AU and Difficulty data into disjoint training and validation sets: Each subject's data were divided into 16 alternating bands of approximately 15 seconds each. The first band was used for training, the second for validation, the third for training, and so on.

Given the set of training data (AUs, their derivatives, and Difficulty values over all training bands), linear regression was performed to predict the Difficulty values in the training set. A grid search over the lag and smoothing parameters was performed to minimize the training error. Given the trained regression model and optimized parameters, the validation performance on the validation bands was then computed. This procedure was conducted separately for each subject.

Results are shown in Table 4. For all subjects except Subject 2, the model was able to predict both the validation Difficulty and Viewing Speed scores with a correlation significantly ($p < 0.05$) above 0. Upon inspecting the AU available for Subject 2, we noticed that the face detection component of the expression recognition system could not find the face for a large stretches of time (the subject may have moved his head slightly out of the camera's view); this effectively decreases the amount of expression data for training and makes the learning task more difficult.

The average validation correlation across all subjects between the model's difficulty output and the self-reported difficulty scores was 0.42. This result is significantly above 0 (Wilcoxon sign rank test, $p < 0.05$), which would be the expected correlation if the expression data contained no useful signal for difficulty prediction. The average validation correlation for predicting preferred viewing speed was 0.29, which was likewise significantly above 0 (Wilcoxon sign rank test, $p < 0.05$), regardless of whether Subject 2 was included or not. While these results show room for improvement, they are nonetheless an encouraging indicator of the utility of facial expression for difficulty prediction, preferred speed estimation, and other important tasks in the ITS domain.

5. Conclusions

Our empirical results indicate that facial expression is a valuable input signal for two concrete tasks important to intelligent tutoring systems: estimating how difficult the student finds a lesson to be, and estimating how fast or slow the student would prefer to watch a lecture. Currently available automatic expression recognition systems can already be used to improve the quality of interactive tutoring programs. As facial expression recognition technology improves in accuracy, the range of its application will grow, both in ITS and beyond. One particular application we are

Facial Expression to Predict Difficulty and Speed (Pearson correlation R):		
Subject	**Difficulty**	**Speed**
1	0.41	0.23
2	0.28	0.04
3	0.44	0.32
4	0.85	0.11
5	0.27	0.44
6	0.56	0.28
7	0.32	0.19
8	0.24	0.68
Avg	**0.42**	**0.29**

Table 4. Accuracy (Pearson R) of predicting the perceived Difficulty, as well as the preferred viewing Speed, of a lecture video from automatic facial expression recognition channels. All results were computed on a validation set not used for training.

currently developing is a "smart video player" which modulates the video speed in real-time based on the user's facial expression so that the rate of lesson presentation is optimal for the current user.

References

[1] V. Aleven and K.R. Koedinger. Limitations of student control: Do students know when they need help? In *Intelligent Tutoring Systems: 5th International Conference*, 2000.

[2] A. Kapoor, W. Burleson, and R. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8), 2007.

[3] H. Rio, A.L. Soli, E. Aguirr, L. Guerrer, and J.P. Alberto Santa. Facial expression recognition and modeling for virtual intelligent tutoring systems. In *Proceedings of the Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence*, 2000.

[4] S.K. D'Mello, R.W. Picard, and A.C. Graesser. Towards an affect-sensitive autotutor. *IEEE Intelligent Systems, Special issue on Intelligent Educational Systems*, 22(4), 2007.

[5] A. Sarrafzadeh, S. Alexander, F. Dadgostar, C. Fan, and A. Bigdeli. See me, teach me: Facial expression and gesture recognition for intelligent tutoring systems. In *Innovations in Information Technology*, 2006.

[6] P. Ekman. *Emotion in the Human Face*. Cambridge University Press, New York, 2 edition, 1982.

[7] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), 2001.

[8] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Proceedings of the IEEE Conference on Automatic Facial and Gesture Recognition*, 2006.

[9] M. Pantic and J.M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man and Cybernetics*, 34(3), 2004.

[10] 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), 2006.

[11] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and J.R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 2006.

[12] 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.

[13] I. Fenwick and M. D. Rice. Reliability of continuous measurement copy-testing methods. *Journal of Advertising Research*, 1991.

[14] M.K. Holland and G. Tarlow. Blinking and mental load. *Psychological Reports*, 31(1), 1972.

[15] H. Tada. Eyeblink rates as a function of the interest valueof video stimuli. *Tohoku Psychologica Folica*, 45, 1986.