To appear in Giese,M. Curio, C., Bulthoff, H. (Eds.) Dynamic Faces: Insights from Experiments and Computation. MIT Press. 2009.

Part III

Chapter 14

Insights on spontaneous facial expressions from automatic expression measurement

Marian Bartlett[1], Gwen Littlewort[1], Esra Vural[1,3], Jake Whitehill[1], Tingfan Wu[1], Kang Lee[2], and Javier Movellan[1]

[1]*Institute for Neural Computation, University of California, San Diego*
[2]*Human Development and Applied Psychology, University of Toronto, Canada*
[3]*Engineering and Natural Science, Sabanci University, Istanbul, Turkey*

Correspondence to be addressed to: marni@salk.edu

Introduction

The computer vision field has advanced to the point that we are now able to begin to apply automatic facial expression recognition systems to important research questions in behavioral science. One of the major outstanding challenges has been to achieve robust performance with spontaneous expressions. Systems that performed well on highly controlled datasets often performed poorly when tested on a different dataset with different image conditions, and especially had trouble generalizing to spontaneous expressions, which tend to have much greater noise from numerous causes. A major milestone in recent years is that automatic facial expression recognition systems are now able to measure spontaneous expressions with some success. This chapter describes one such system, the Computer Expression Recognition Toolbox (CERT). CERT will soon be available to the research community. (See http://mplab.ucsd.edu).

This system was employed in some of the earliest experiments in which spontaneous behavior was analyzed with automated expression recognition to extract new information about facial expression that was previously unknown (Bartlett et al., 2008). These experiments are summarized in this chapter. The experiments measured facial behavior associated with faked versus genuine expressions of pain, driver drowsiness, and perceived difficulty of a video lecture. The analysis revealed information about facial behavior during these conditions that were previously unknown, including the coupling of movements such as eye openness with brow raise during driver drowsiness. Automated classifiers were able to differentiate real from fake pain significantly better than naïve human subjects, and to detect driver drowsiness above 98% accuracy. Another experiment showed that facial expression was able to predict perceived difficulty of a video lecture and preferred presentation speed (Whitehill et al., 2008). CERT is also being employed in a project to give feedback on facial expression production to children with autism. A prototype game, called SmileMaze, requires the player to produce a smile to enable a character to pass though doors and obtain rewards (Cockburn et al., 2008).

These are among the first generation of research studies to apply fully automated facial expression measurement to research questions in behavioral science. Tools for automatic expression measurement will bring about paradigmatic shifts in a number of fields by making facial expression more accessible as a behavioral measure. Moreover, automated facial expression analysis will enable investigations into facial expression dynamics that were previously intractable by human coding because of the time required to code intensity changes.

The Facial Action Coding System

The facial action coding system (FACS) (Ekman and Friesen, 1978) is arguably the most widely used method for coding facial expressions in the behavioral sciences. The system describes facial expressions in terms of 46 component movements, which roughly

correspond to the individual facial muscle movements. An example is shown in Figure 1. FACS provides an objective and comprehensive way to analyze expressions into elementary components, analogous to decomposition of speech into phonemes. Because it is comprehensive, FACS has proven useful for discovering facial movements that are indicative of cognitive and affective states. See Ekman and Rosenberg (2005) for a review of facial expression studies using FACS. The primary limitation to the widespread use of FACS is the time required to code. FACS was developed for coding by hand, using human experts. It takes over 100 hours of training to become proficient in FACS, and it takes approximately 2 hours for human experts to code each minute of video. The authors have been developing methods for fully automating the facial action coding system (e.g. Donato et al., 1999; Bartlett et al., 2006). In this chapter we apply a computer vision system trained to automatically detect facial Action Units to data mine facial behavior and to extract facial signals associated with states including: (1) real versus fake pain, (2) driver fatigue (3) easy versus difficult video lectures.

[INSERT FIGURE 1 ABOUT HERE]

Spontaneous Expressions

The machine learning system presented here was trained on spontaneous facial expressions. The importance of using spontaneous behavior for developing and testing computer vision systems becomes apparent when we examine the neurological substrate for facial expression production. There are two distinct neural pathways that mediate facial expressions, each one originating in a different area of the brain. Volitional facial movements originate in the cortical motor strip, whereas spontaneous facial expressions originate in the subcortical areas of the brain (see Rinn, 1984, for a review). These two pathways have different patterns of innervation on the face, with the cortical system tending to give stronger innervation to certain muscles primarily in the lower face, while the subcortical system tends to more strongly innervate certain muscles primarily in the upper face (e.g. Morecraft et al., 2001).

The facial expressions mediated by these two pathways have differences both in which facial muscles are moved and in their dynamics (Ekman, 2001; Ekman & Rosenberg, 2005). Subcortically initiated facial expressions (the spontaneous group) are characterized by synchronized, smooth, symmetrical, consistent, and reflex-like facial muscle movements whereas cortically initiated facial expressions (posed expressions) are subject to volitional real-time control and tend to be less smooth, with more variable dynamics (Rinn, 1984; Frank, Ekman, & Friesen, 1993; Schmidt, Cohn & Tian, 2003; Cohn & Schmidt, 2004). Given the two different neural pathways for facial expression production, it is reasonable to expect to find differences between genuine and posed expressions of states such as pain or drowsiness. Moreover, it is crucial that the computer vision model for detecting states such

as genuine pain or driver drowsiness is based on machine learning of expression samples when the subject is actually experiencing the state in question.

The Computer Expression Recognition Toolbox

The Computer Expression Recognition Toolbox (CERT), developed at University of California, San Diego, is a fully automated system that analyzes facial expressions in real-time. CERT is based on 15 years experience in automated facial expression recognition (e.g. Bartlett et al., 1996, 1999, 2006; Donato et al., 1999; Littlewort et al., 2006). This line of work originated from a collaboration between Sejnowski and Ekman, in response to an NSF planning workshop on automated facial expression understanding (Ekman et al., 1993). The present system automatically detects frontal faces in the video stream and codes each frame with respect to 40 continuous dimensions, including basic expressions of anger, disgust, fear, joy, sadness, surprise, contempt, a continuous measure of head pose (yaw, pitch, and roll), as well as 30 facial action units (AU's) from the Facial Action Coding System (Ekman & Friesen, 1978). See Figure 2.

System overview

The technical approach to CERT is an appearance-based discriminative approach. Such approaches have proven highly robust and fast for face detection and tracking (e.g. Viola & Jones, 2004). Appearance-based discriminative approaches don't suffer from initialization and drift, which presents challenges for state of the art tracking algorithms, and take advantage of the rich appearance-based information in facial expression images. This class of approaches achieves a high level of robustness through the use of very large datasets for machine learning. It is important that the training set is similar to the proposed applications in terms of noise. A detailed analysis of machine learning methods for robust detection of one facial expression, smiles, is provided in Whitehill et al., (in press).

[INSERT FIGURE 2 ABOUT HERE]

The design of CERT is as follows. Face detection and detection of internal facial features is first performed on each frame using boosting techniques in a generative framework (Fasel et al. 2005), extending work by Viola and Jones (2004). The automatically located faces then undergo a 2D alignment by computing a fast least squares fit between the detected feature positions and a six feature face model. The least squares alignment allows rotation, scale, and shear. The aligned face image is then passed through a bank of Gabor filters 8 orientations and 9 spatial frequencies (2 to 32 pixels per cycle at 1/2 octave steps). Output magnitudes were then normalized and passed to facial action classifiers.

Facial action detectors were then developed by training separate support vector machines to detect the presence or absence of each facial action. The training set consisted

of over 10000 images which were coded for facial actions from the Facial Action Coding System, including over 5000 examples from spontaneous expressions.

In previous work, we conducted empirical investigations of machine learning methods applied to the related problem of classifying expressions of basic emotions. We compared image features (e.g. Donato et al., 1999), classifiers such as AdaBoost, support vector machines, and linear discriminant analysis, as well as feature selection techniques (Littlewort et al., 2006). Best results were obtained by selecting a subset of Gabor filters using AdaBoost and then training Support Vector Machines on the outputs of the filters selected by AdaBoost. An overview of the system is shown in Figure 2a.

Benchmark performance

In this chapter, performance for expression detection is assessed using a measure from signal detection theory, area under the ROC curve (A'). The ROC curve is obtained by plotting true positives against false positives as the decision threshold for deciding 'expression present' shifts from high (0 detections and 0 false positives) to low (100% detections and 100% false positives). A' ranges from 0.5 (chance) to 1 (perfect discrimination). We employ A' instead of percent correct since A' can change for the same system depending on the proportion of targets and nontargets in a given test set. A' can be interpreted in terms of percent correct on a 2 alternative forced choice task in which two images are presented on each trial and the system must decide which of the two is the target.

Performances on a benchmark dataset (Cohn-Kanade) shows state of the art performance for both recognition of basic emotions and facial actions. Performance for expressions of basic emotion was .98 area under the ROC for detection (1 vs. all) across 7 expressions of basic emotion. , and 93% correct for a 7 alternative forced choice. This is the highest performance reported to our knowledge on this benchmark dataset. Performance for recognizing facial actions from the Facial Action Coding System was .93 mean area under the ROC for posed facial actions. (This was mean detection performance across 20 facial action detectors). Recognition of spontaneous facial actions was tested on the RU-FACS dataset (Bartlett et al., 2006). This dataset is an interview setting containing speech. Performance was tested for 33 subjects with four minutes of continuous video each. Mean area under the ROC for detection of 11 facial actions was 0.84.

System outputs consist of the margin of the SVM (distance to the separating hyperplane between the two classes), for each frame of video. System outputs are significantly correlated with the intensity of the facial action, as measured by FACS expert intensity codes (Bartlett et al., 2006), and also as measured by naïve observers obtained by turning a dial while watching continuous video (Whitehill et al., in press). Thus the frame-by-frame intensities provide information on the dynamics of facial expression at temporal resolutions that were previously impractical via manual coding. There is also preliminary evidence from Jim Tanaka's laboratory of concurrent validity with EMG. CERT outputs significantly correlated with EMG measures of zygomatic and corrugator activity despite the visibility of the electrodes in the video processed by CERT.

A second-layer classifier to detect internal states

The overall CERT system gives a frame-by-frame output with N channels, consisting of N facial Action Units. This system can be applied to data mine human behavior. By applying CERT to face video while subjects experience spontaneous expressions of a given state, we can learn new things about the facial behaviors associated with that state. Also, by passing the N channel output to a machine learning system, we can directly train detectors for the specific state in question. See Figure 3.

[INSERT FIGURE 3 ABOUT HERE]

Real versus Faked Expressions of Pain

The ability to distinguish real pain from faked pain (malingering) is an important issue in medicine (Fishbain, 2006). Naïve human subjects are near chance for differentiating real from fake pain from observing facial expressions (e.g. Hadjistavropoulos et al., 1996). In the absence of direct training in facial expressions, clinicians are also poor at assessing pain from the face (e.g. Prkachin et al. 2001 and 2007; Grossman, 1991). However a number of studies using the Facial Action Coding System (FACS) (Ekman & Friesen, 1978) have shown that information exists in the face for differentiating real from posed pain (e.g. Hill and Craig, 2002; Craig et al., 1991; Prkachin 1992). In fact, if subjects receive corrective feedback, their performance improves substantially (Hill & Craig, 2004). Thus it appears that a signal is present, but that most people don't know what to look for.

This study explored the application of a system for automatically detecting facial actions to this problem. This section is based on work described in Littlewort, Bartlett & Lee (2009). The goal of this work was to 1) assess whether the automated measurements with CERT were consistent with expression measurements obtained by human experts, and 2) develop a classifier to automatically differentiate real from faked pain in a subject-independent manner from the automated measurements.

In this study, participants were videotaped under three experimental conditions: baseline, posed pain, and real pain. We employed a machine learning approach in a two-stage system. In the first stage, the video was passed through a system for detecting facial actions from the Facial Action Coding System (Bartlett et al., 2006). This data was then passed to a second machine learning stage, in which a classifier was trained to detect the difference between expressions of real pain and fake pain. Naïve human subjects were tested on the same videos to compare their ability to differentiate faked from real pain.

The ultimate goal of this work is not the detection of malingering per se, but rather to demonstrate the ability of the automated systems to detect facial behavior that the untrained eye might fail to interpret, and to differentiate types of neural control of the face.

It holds out the prospect of illuminating basic questions pertaining to the behavioral fingerprint of neural control systems, and thus opens many future lines of inquiry.

Video data collection

Video data was collected of 26 human subjects during real pain, faked pain, and baseline conditions. Human subjects were university students consisting of 6 men and 20 women. The pain condition consisted of cold pressor pain induced by immersing the arm in icewater at 5^0 Celsius. For the baseline and faked pain conditions, the water was 20^0 Celsius. Subjects were instructed to immerse their forearm into the water up to the elbow, and hold it there for 60 seconds in each of the three conditions. For the faked pain condition, subjects were asked to manipulate their facial expressions so that an "expert would be convinced they were in actual pain." In the baseline condition, subjects were instructed to display their expressions naturally. Participants facial expressions were recorded using a digital video camera during each condition. Examples are shown in Figure 4. For the 26 subjects analyzed here, the order of the conditions was baseline, faked pain, and then real pain. Another 22 subjects received the counterbalanced order: baseline, real pain, then faked pain. Because repeating facial movements that were experienced minutes before differs from faking facial expressions without immediate pain experience, the two conditions were analyzed separately. The following analysis focuses on the condition in which subjects fake first.

[INSERT FIGURE 4 ABOUT HERE]

After the videos were collected, a set of 170 naïve observers were shown the videos and asked to guess whether each video contained faked or real pain. Subjects were undergraduates with no explicit training in facial expression measurement. They were primarily Introductory Psychology students at UCSD. Mean accuracy of naïve human subjects for discriminating fake from real pain in these videos was at chance at 49.1% (standard deviation 13.7%).

Characterizing the Difference Between Real and Faked Expressions of Pain

The computer expression recognition toolbox was applied to the three one-minute videos of each subject. The following set of 20 facial actions[1] was detected for each frame [1 2 4 5 6 7 9 10 12 14 15 17 18 20 23 24 25 26 1+4 1+2+4]. This produced a 20 channel output stream, consisting of one real value for each learned AU, for each frame of the video. We first assessed which AU outputs contained information about genuine pain expressions, faked pain expressions, and show differences between genuine versus faked pain. The results were compared to studies that employed expert human coding.

*Real pain vs. baseline***.** We first examined which facial action detectors were elevated in real

[1] A version of CERT was employed that just detected these 20 AUs.

pain compared to the baseline condition. Z-scores for each subject and each AU detector were computed as $Z=(x-\mu)/\sigma$, where (μ,σ) are the mean and variance for the output of frames 100-1100 in the baseline condition (warm water, no faked expressions). The mean difference in Z-score between the baseline and pain conditions was computed across the 26 subjects. Table 1 shows the action detectors with the largest difference in Z-scores. We observed that the actions with the largest Z-scores for genuine pain were Mouth opening and jaw drop (25 and 26), lip corner puller by zygomatic (12), nose wrinkle (9), and to a lesser extent, lip raise (10) and cheek raise (6). These facial actions have been previously associated with cold pressor pain (e.g. Prkachin, 1992; Craig & Patrick 1985).

*Faked pain vs. baseline***.** The Z-score analysis was next repeated for faked versus baseline. We observed that in faked pain there was relatively more facial activity than in real pain. The facial action outputs with the highest z-scores for faked pain relative to baseline were brow lower (4), distress brow (1 or 1+4), inner brow raise (1), mouth open and jaw drop (25 and 26), cheek raise (6), lip raise (10), fear brow (1+2+4), nose wrinkle (9), mouth stretch (20), and lower lid raise (7).

*Real vs. Faked Pain***.** Differences between real and faked pain were examined by computing the difference of the two z-scores. Differences were observed primarily in the outputs of action unit 4 (brow lower), as well as distress brow (1 or 1+4) and inner brow raise (1 in any combination).

[INSERT TABLE 1 ABOUT HERE]

There was considerable variation among subjects in the difference between their faked and real pain expressions. However the most consistent finding is that 9 of the 26 subjects showed significantly more brow lowering activity (AU4) during the faked pain condition, whereas none of the subjects showed significantly more AU4 activity during the real pain condition. Also 7 subjects showed more cheek raise (AU6), and 6 subjects showed more inner brow raise (AU1), and the fear brow combination (1+2+4). The next most common differences were to show more 12, 15, 17, and distress brow (1 alone or 1+4) during faked pain.

Paired t-tests were conducted for each AU to assess whether it was a reliable indicator of genuine versus faked pain in a within-subjects design. Of the 20 actions tested, the difference was statistically significant for three actions. It was highly significant for AU 4 ($p < .001$), and marginally significant for AU 7 and distress brow ($p < .05$).

Comparison with human expert coding

The findings from the automated system were first compared to previous studies that used manual FACS coding by human experts. Overall, the outputs of the automated system

showed similar patterns to previous studies of real and faked pain.

Real pain. In previous studies using manual FACS coding by human experts, at least 12 facial actions showed significant relationships with pain across multiple studies and pain modalities. Of these, the ones specifically associated with cold pressor pain were 4, 6, 7, 9, 10, 12, 25, 26 (Craig & Patrick, 1985; Prkachin, 1992). Agreement of the automated system with the human coding studies was computed as follows: first a superset of the AU's tested in the two cold pressor pain studies was created. AU's that were significantly elevated in either study were assigned a 1, and otherwise a 0. This vector was then correlated against the findings for the 20 AUs measured by the automated system. AUs with the highest z-scores, shown in Table 1A, were assigned a 1, and the others a 0. Only AU's that were measured both by the automated system and by at least one of the human coding studies were included in the correlation analysis. Agreement computed in this manner was 90% for AU's associated with real pain.

Faked pain. A study of faked pain in adults showed elevation of the following AU's: 4, 6, 7, 10, 12, 25. (Craig, Hyde & Patrick, 1991). A study of faked pain in children ages 8-12, (LaRochette et al., 2006) observed significant elevation in the following AU's for fake pain relative to baseline: 1 4 6 7 10 12 20 23 25 26. These findings again match the AU's with the highest z-scores in the automated system output of the present study, as shown in Table 1B. (The two human coding studies did not measure AU 9 or the brow combinations). Agreement of the computer vision findings with these two studies was 85%.

Real vs. Faked pain. Exaggerated activity of the brow lower (AU 4) during faked pain is consistent with previous studies in which the real pain condition was exacerbated lower back pain (Craig et al. 1991, Hill & Craig, 2002). Only one other study looked at real versus faked pain in which the real pain condition was cold pressor pain. This was a study with children ages 8-12 (LaRochette et al., 2006). When faked pain expressions were compared with real cold pressor pain in children, LaRochette et al found significant differences in AU's 1 4 7 10. Again the findings of the present study using the automated system are similar, as the AU channels with the highest z-scores were 1, 4, and 1+4 (Table 1C), and the t-tests were significant for 4, 1+4 and 7. Agreement of the automated system with the human coding findings was 90%.

Human FACS coding of video in this study. In order to further assess the validity of the automated system findings, we obtained FACS codes for a portion of the video data employed in this study. FACS codes were obtained by an expert coder certified in the Facial Action Coding System. For each subject, the last 500 frames of the fake pain and real pain conditions were FACS coded (about 15 seconds each). It took 60 man hours to collect the human codes, over the course of more than 3 months, since human coders can only code up to 2 hours per day before having negative repercussions in accuracy and coder burn-out.

The sum of the frames containing each action unit were collected for each subject

condition, as well as a weighted sum, multiplied by the intensity of the action on a 1-5 scale. To investigate whether any action units successfully differentiated real from faked pain, paired t-tests were computed on each individual action unit. (Tests on specific brow region combinations 1+2+4 and 1,1+4 have not yet been conducted.) The one action unit that significantly differentiated the two conditions was AU 4, brow lower, ($p<.01$) for both the sum and weighted sum measures. This finding is consistent with the analysis of the automated system, which also found action unit 4 most discriminative.

Automatic Discrimination of Real from Fake Pain

We next turned to the problem of automatically discriminating genuine from faked pain expressions from the facial action output stream. This section describes the second machine learning stage, in which a classifier was trained to discriminate genuine from faked pain from the output of the 20 facial action detectors. The task was to perform subject-independent classification. If the task were to simply detect the presence of a red-flag set of facial actions, then differentiating fake from real pain expressions would be relatively simple. However, it should be noted that subjects display actions such as AU 4, for example, in both real and fake pain, and the distinction is in the magnitude and duration of AU 4. Also, there is inter-subject variation in expressions of both real and fake pain, there may be combinatorial differences in the sets of actions displayed during real and fake pain, and the subjects may cluster. We therefore applied machine learning to the task of discriminating real from faked pain expressions.

A classifier was trained to discriminate genuine pain from faked pain based on the CERT output. The input to this classifier consisted of the 20 facial action detector outputs from the full minute of video in each condition. Before training the classifier we developed a representation that summarized aspects of the dynamics of the facial behavior. The frame-by-frame AU detection data was integrated into temporal 'events'. We applied temporal filters at 8 different fundamental frequencies to the AU output stream using difference of Gaussians convolution kernels. The half-width of the positive Gaussian (σ) ranged from 8 to 90.5 frames at half-octave intervals (0.25 to 1.5 seconds), and the size of the negative Gaussian was fixed at 2σ. Zero-crossings of the filter outputs were used to segment the output into positive and negative regions. The integral of each region was then computed. See Figure 5. The distributions of the positive events was then characterized using histograms. These histograms of 'events' comprised the input representation to train a nonlinear SVM with Gaussian kernel. The input therefore consisted of 20 histograms, one for each AU.

[INSERT FIGURE 5 ABOUT HERE]

The system was trained using cross-validation on the 26 subject videos. In the cross-validation approach, the system was trained and tested 26 times, each time using data from

25 subjects for parameter estimation and reserving a different subject for testing. This provided an estimate of subject-independent detection performance. The percent correct 2-alternative forced choice of fake versus real pain on new subjects was 88 percent. This was significantly higher than performance of naïve human subjects, who obtained a mean accuracy of 49% correct for discriminating faked from real pain on the same set of videos. Performance using the integrated event representation was also considerably higher than an alternative representation that did not take temporal dynamics into account. (See Littlewort et al., in press, for details.) This integrated event representation contains useful dynamic information allowing more accurate behavioral analysis. This suggests that this decision task depends not only on which subset of AU's are present at which intensity, but also on the duration and number of AU events.

Discussion of pain study

The field of automatic facial expression analysis has advanced to the point that we can begin to apply it to address research questions in behavioral science. Here we describe a pioneering effort to apply fully automated facial action coding to the problem of differentiating fake from real expressions of pain. While naïve human subjects were only at 49% accuracy for distinguishing fake from real pain, the automated system obtained .88 area under the ROC, which is equivalent to 88% correct on a 2-alternative forced choice. Moreover, the pattern of results in terms of which facial actions may be involved in real pain, fake pain, and differentiating real from fake pain is similar to previous findings in the psychology literature using manual FACS coding.

Here we applied machine learning on a 20-channel output stream of facial action detectors. The machine learning was applied to samples of spontaneous expressions during the subject state in question. Here the state in question was fake versus real pain. The same approach can be applied to learn about other subject states, given a set of spontaneous expression samples. Section 4 develops another example in which this approach is applied to the detection of driver drowsiness from facial expression.

Automatic Detection of Driver Fatigue[2]

Vural, E., Cetin, M., Ercil, A., Littlewort, G., Bartlett, M., and Movellan, J. (2007). Drowsy driver detection through facial movement analysis. ICCV Workshop on Human Computer Interaction. © 2007 IEEE.

[2] Based on Vural, E., Cetin, M., Ercil, A., Littlewort, G., Bartlett, M., and Movellan, J. (2007). Drowsy driver detection through facial movement analysis. ICCV Workshop on Human Computer Interaction. © 2007 IEEE.

The US National Highway Traffic Safety Administration estimates that in the US alone approximately 100,000 crashes each year are caused primarily by driver drowsiness or fatigue (Department of Transportation, 2001). In fact, the NHTSA has concluded that drowsy driving is just as dangerous as drunk driving. Thus methods to automatically detect drowsiness may help save many lives.

There are a number of techniques for analyzing driver drowsiness. One set of techniques places sensors on standard vehicle components, e.g., steering wheel, gas pedal, and analyzes the signals sent by these sensors to detect drowsiness (Takei & Furukawa, 2005). It is important for such techniques to be adapted to the driver, since Abut and his colleagues note that there are noticeable differences among drivers in the way they use the gas pedal (Igarashi, et al., 2005). A second set of techniques focuses on measurement of physiological signals such as heart rate, pulse rate, and Electroencephalography (EEG) (e.g. Cobb, 1983). It has been reported by researchers that as the alertness level decreases, EEG power of the alpha and theta bands increases (Hung & Chung, 2005), and thereby provides indicators of drowsiness. However this method has drawbacks in terms of practicality since it requires a person to wear an EEG cap while driving.

A third set of solutions focuses on computer vision systems that can detect and recognize the facial motion and appearance changes occurring during drowsiness (Gu & Ji, 2004; Gu, Zhang & Ji, 2005; Zhang & Zhang, 2006). The advantage of computer vision techniques is that they are non-invasive, and thus are more amenable to use by the general public. Most of the previous research on computer vision approaches to detection of fatigue primarily make pre-assumptions about the relevant behavior, focusing on blink rate, eye closure, and yawning. Here we employ machine learning methods to data mine actual human behavior during drowsiness episodes. The objective of this study was to discover what facial configurations are predictors of fatigue. In this study, facial motion was analyzed automatically from video using the computer expression recognition toolbox (CERT) to automaticaloy code facial actions from the Facial Action Coding System.

Driving task

Subjects played a driving video game on a windows machine using a steering wheel[3] and an open source multi-platform video game[4] (See Figure 6). Subjects were instructed to keep the car as close to the center of the road as possible as they drove a simulation of a winding road. At random times, a wind effect was applied that dragged the car to the right or left, forcing the subject to correct the position of the car. This type of manipulation had been found in the past to increase fatigue (Orden, Jung & Makeig, 2000). Driving speed was held

[3]Thrustmaster[R]Ferrari Racing Wheel
[4]The Open Racing Car Simulator (TORCS)

constant. Four subjects performed the driving task over a three hour period beginning at midnight. During this time subjects fell asleep multiple times thus crashing their vehicles. Episodes in which the car left the road (crash) were recorded. Video of the subjects face was recorded using a DV camera for the entire 3 hour session.

In addition to measuring facial expressions with CERT, head movement was measured using an accelerometer placed on a headband, as well as steering wheel movement data. The accelerometer had 3 degrees of freedom, consisting of three one dimensional accelerometers mounted at right angles measuring accelerations in the range of 5g to +5g where g represents earth gravitational force.

[INSERT FIGURE 6 ABOUT HERE]

Facial Actions Associated with Driver Fatigue

Subject data was partitioned into drowsy (non-alert) and alert states as follows. The one minute preceding a sleep episode or a crash was identified as a non-alert state. There was a mean of 24 non-alert episodes per subject, with each subject contributing between 9 and 35 non-alert samples. Fourteen alert segments for each subject were collected from the first 20 minutes of the driving task.

[INSERT FIGURE 7 ABOUT HERE]

The output of the facial action detector (CERT) consisted of a continuous value for each facial action and each video frame which was the distance to the separating hyperplane, i.e., the margin. Histograms for two of the action units in alert and non-alert states are shown in Figure 7. The area under the ROC (A') was computed for the outputs of each facial action detector to see to what degree the alert and non-alert output distributions were separated.

In order to understand how each action unit is associated with drowsiness across different subjects, Multinomial Logistic Ridge Regression (MLR) was trained on each facial action individually. Examination of the A' for each action unit reveals the degree to which each facial movement was able to predict drowsiness in this study. The A's for the drowsy and alert states are shown in Table 2. The five facial actions that were the most predictive of drowsiness by increasing in drowsy states were 45, 2 (outer brow raise), 15 (frown), 17 (chin raise), and 9 (nose wrinkle). The five actions that were the most predictive of drowsiness by decreasing in drowsy states were 12 (smile), 7 (lid tighten), 39 (nostril compress), 4 (brow lower), and 26 (jaw drop). The high predictive ability of the blink/eye closure measure was expected. However the predictability of the outer brow raise (AU 2) was previously unknown.

We observed during this study that many subjects raised their eyebrows in an attempt to keep their eyes open, and the strong association of the AU 2 detector is consistent with that observation. Also of note is that action 26, jaw drop, which occurs

during yawning, actually occurred less often in the critical 60 seconds prior to a crash. This is consistent with the prediction that yawning does not tend to occur in the final moments before falling asleep.

[INSERT TABLE 2 ABOUT HERE]

Automatic Detection of Driver Fatigue

The ability to predict drowsiness in novel subjects from the facial action code was then tested by running MLR on the full set of facial action outputs. Prediction performance was tested by using a leave-one-out cross validation procedure, in which one subjects' data was withheld from the MLR training and retained for testing, and the test was repeated for each subject. The data for each subject by facial action was first normalized to zero-mean and unit standard deviation. The MLR output for each AU feature was summed over a temporal window of 12 seconds (360 frames) before computing A'.

MLR trained on all AU features obtained an A' of .90 for predicting drowsiness in novel subjects. Because prediction accuracy may be enhanced by feature selection, in which only the AU's with the most information for discriminating drowsiness are included in the regression, a second MLR was trained by contingent feature selection, starting with the most discriminative feature (AU 45), and then iteratively adding the next most discriminative feature given the features already selected. These features are shown on Table 3. Best performance of .98 was obtained with five features: 45, 2, 19 (tongue show), 26 (jaw drop), and 15. This five feature model outperformed the MLR trained on all features.

*Effect of Temporal Window Length***.** The performances shown in Table 3 employed a temporal window of 12 seconds, meaning that the MLR output was summed over 12 seconds (360 frames) for making the classification decision (drowsy/not-drowsy). We next examined the effect of the size of the temporal window on performance. Here, the MLR output was summed over windows of N seconds, where N ranged from 0.5 to 60 seconds. The five feature model was again employed for this analysis. Figure 8 shows the area under the ROC for drowsiness detection in novel subjects over a range of temporal window sizes. Performance saturates at about 0.99 as the window size exceeds 30 seconds. In other words, given a 30 second video segment the system can discriminate sleepy versus non-sleepy segments with 0.99 accuracy across subjects.

[INSERT TABLE 3 ABOUT HERE]

[INSERT FIGURE 8 ABOUT HERE]

Coupling of behaviors

Coupling of steering and head motion. Observation of the subjects during drowsy and nondrowsy states indicated that the subjects head motion differed substantially when alert versus when the driver was about to fall asleep. Surprisingly, head motion increased as the driver became drowsy, with large roll motion coupled with the steering motion as the driver became drowsy. Just before falling asleep, the head would become still.

We also investigated the coupling of the head and arm motions. Correlations between head motion as measured by the roll dimension of the accelerometer output and the steering wheel motion are shown in Figure 9. For this subject (subject 2), the correlation between head motion and steering increased from 0.33 in the alert state to 0.71 in the non-alert state. For subject 1, the correlation between head motion and steering similarly increased from 0.24 in the alert state to 0.43 in the non-alert state. The other two subjects showed a smaller coupling effect. Future work includes combining the head motion measures and steering correlations with the facial movement measures in the predictive model for detecting driver drowsiness.

Coupling of eye openness and eyebrow raise. We observed that for some of the subjects coupling between eye brow up's and eye openness increased in the drowsy state. In other words subjects tried to open their eyes using their eyebrows in an attempt to keep awake. See Figure 10.

[INSERT FIGURE 9 ABOUT HERE]

[INSERT FIGURE 10 ABOUT HERE]

Conclusions of Driver Fatigue Study

This section presented a system for automatic detection of driver drowsiness from video. Previous approaches focused on assumptions about behaviors that might be predictive of drowsiness. Here, a system for automatically measuring facial expressions was employed to data mine spontaneous behavior during real drowsiness episodes. This is the first work to our knowledge to reveal significant associations between facial expression and fatigue beyond eyeblinks. The project also revealed a potential association between head roll and driver drowsiness, and the coupling of head roll with steering motion during drowsiness. Of note is that a behavior that is often assumed to be predictive of drowsiness, yawn, was in fact a negative predictor of the 60-second window prior to a crash. It appears that in the moments before falling asleep, drivers yawn less, not more, often. This highlights the importance of using examples of fatigue and drowsiness conditions in which subjects actually fall sleep.

Automated Feedback for Intelligent Tutoring Systems

Whitehill, Bartlett and Movellan (2008) investigated the utility of integrating automatic facial expression recognition into an automated teaching system. We briefly describe that study here. There has been a growing thrust to develop tutoring systems and agents that respond to the students' emotional and cognitive state and interact with them in a social manner (e.g. Kapoor et al. 2007,D'Mello et al., 2007). Whitehill's work used expression to estimate the student's preferred viewing speed of the videos, and the level of difficulty, as perceived by the individual student, of the lecture at each moment of time. This study took first steps towards developing methods for closed loop teaching policies, i.e., systems that have access to real time estimates of cognitive and emotional states of the students and act accordingly. The goal of this study was to assess whether automated facial expression measurements using CERT, collected in real-time, could predict factors such as the perceived difficulty of a video lecture or the preferred viewing speed.

In this study, 8 subjects separately watched a video lecture composed of several short clips on mathematics, physics, psychology, and other topics. The playback speed of the video was controllable by the student using the keyboard - the student could speed up, slow down, or rewind the video by pressing a different key. The subjects were instructed to watch the video as quickly as possible (so as to be efficient with their time) while still retaining accurate knowledge of the video's content, since they would be quizzed afterwards.

While watching the lecture, the student's facial expressions were measured in real-time by the CERT system (Bartlett et al., 2006). CERT is fully automatic, and requires no manual initialization or calibration, which enables real-time applications. The version of CERT used in the study contained detectors for 12 different Facial Action Units as well as a Smile detector trained to recognize social smiles. Each detector output a real-valued estimate of the expression's intensity at each moment in time. After watching the video and taking the quiz, each subject then watched the lecture video again at a fixed speed of 1.0 (real-time). During this second viewing, subjects specified how easy or difficult they found the lecture to be at each moment in time using the keyboard.

In total, the experiment resulted in three data series for each student: (1) Expression, consisting of 13 facial expression channels (12 different facial actions, plus smile); (2) Speed, consisting of the video speed at each moment in time as controlled by the user during the first viewing of the lecture; and (3) Difficulty, as reported by the student him/herself during the second viewing of the video.

For each subject, the three data series were time-aligned and smoothed. Then, a regression analysis was performed using the Expression data (both the 13 intensities themselves, as well as the first temporal derivatives) as independent variables to predict both Difficulty and Expression using standard linear regression. An example of such predictions is shown in Figure 11c for one subject.

[INSERT FIGURE 11 ABOUT HERE]

The results of the correlation study suggested that facial expression, as estimated by a fully automatic facial expression recognition system, is significantly predictive of both preferred viewing Speed and perceived Difficulty. Across the 8 subjects, and evaluated on a separate 'validation' set taken from the three data series that was not used for training, the average correlations with Difficulty and Speed were 0.42 and 0.29, respectively. The specific facial expressions which were correlated varied highly from subject to subject. No individual facial expression was consistently correlated across all subjects, but the most consistently correlated expression (taking parity into account) was AU 45 (``blink''). The blink action was negatively correlated with perceived Difficulty, meaning that subjects blinked less during the more difficult sections of video. This is consistent with previous work associating decreases in blink rate with increases in cognitive load (Holland & Tarlow, 1972; Tada 1986).

Overall, this pilot study provided proof of principal, that fully automated facial expression recognition at the present state of the art can be used to provide real-time feedback in automated tutoring systems. The validation correlations were small, but statistically significant (Wilcoxon sign rank test $p < .05$), showing that a signal exists in the face and that the automated system can detect it in real time. Since conducting the pilot study, we have added 25 additional facial expression dimensions, and made a number of improvements to CERT. Moreover, the correlation analysis was quite simple. Future work includes exploration of more powerful dynamical models which may give stronger prediction accuracy.

Discussion

The computer vision field has advanced to the point that we are now able to begin to apply automatic facial expression recognition systems to important research questions in behavioral science. This chapter explored three such applications, in which the automated measurement system revealed information about facial expression that was previously unknown. While the accuracy of individual facial action detectors is still below that of human experts, automated systems can be applied to large quantities of video data. Statistical pattern recognition on this large quantity of data can reveal emergent behavioral patterns that previously would have required hundreds of coding hours by human experts, and would be unattainable by the non-expert. Moreover, automated facial expression analysis will enable investigations into facial expression dynamics that were previously intractable by human coding because of the time required to code intensity changes.

Acknowledgements

Support for this work was provided in part by NSF grants CNS-0454233, SBE-0542013, and NSF ADVANCE award 0340851, and by a grant from Turkish State Planning Organization. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Portions of the research in this paper use the MMI Facial Expression Database collected by M. Pantic & M.F. Valstar.

References

Bartlett, M. Littlewort, G. Vural, E., Lee, K., Cetin, M., Ercil, A., and Movellan, M. (2008). Data mining spontaneous facial behavior with automatic expression coding. Lecture Notes in Computer Science 5042 p. 1-21.

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

Bartlett, M.S., Hager, J.C., Ekman, P., and Sejnowski, T.J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, *36*, 253-263.

Bartlett, M. Stewart, Viola, P.A., Sejnowski, T.J., Golomb, B.A., Larsen, J., Hager, J.C., and Ekman, P. (1996). Classifying facial action, *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge, MA. p. 823-829.

Cockburn, J., Bartlett, M., Tanaka, J., Movellan, J., Pierce, M., and Schultz, R. (2008). SmileMaze: A Tutoring System in Real-Time Facial Expression Perception and Production for Children with Autism Spectrum Disorder. Intl Conference on Automatic Face and Gesture Recognition, Workshop on Facial and Bodily expressions for Control and Adaptation of Games.

Cobb. W. (1983). Recommendations for the practice of clinical neurophysiology. Elsevier.

Cohn, J.F. & Schmidt, K.L. (2004). The timing of facial motion in posed and spontaneous smiles. *J. Wavelets, Multi-resolution & Information Processing*, Vol. 2, No. 2, pp. 121-132.

Craig KD, Hyde S, Patrick CJ. (1991). Genuine, supressed, and faked facial behaviour during exacerbation of chronic low back pain. Pain 46:161– 172.

Craig KD, Patrick CJ. (1985). Facial expression during induced pain. J Pers Soc Psychol. 48(4):1080-91.

D'Mello, S., Picard, R., & Graesser, A.(2007). Towards an affect-sensitive autotutor. IEEE Intelli- gent Systems, Special issue on Intelligent Educational Systems, 22 (4), 53-61.

DOT (2001). Saving lives through advanced vehicle safety technology. USA Department of Transportation. http://www.its.dot.gov/ivi/docs/AR2001.pdf.

Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P. & Sejnowski, T.J. (1999). Classifying facial actions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 21, No. 10, pp. 974-989.

Ekman, P. (2001). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. W.W. Norton, New York, USA.

Ekman, P., Levenson, R.W. and Friesen, W.V. (1983). Autonomic nervous system activity distinguishes between emotions. *Science*, *221,* 1208–1210.

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

Ekman, P. & Rosenberg, E.L., (Eds.), (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the FACS*, Oxford University Press, Oxford, UK.

Fasel I., Fortenberry B., Movellan J.R. "A generative framework for real-time object detection and classification.," *Computer Vision and Image Understanding* 98, 2005.

Fishbain DA, Cutler R, Rosomoff HL, Rosomoff RS. (2006). Accuracy of deception judgments. Pers Soc Psychol Rev. 10(3):214-34.

Frank MG, Ekman P, Friesen WV. (1993). Behavioral markers and recognizability of the smile of enjoyment. J Pers Soc Psychol. 64(1):83-93.

Grossman, S., Shielder, V., Swedeen, K., Mucenski, J. (1991). Correlation of patient and caregiver ratings of cancer pain. Journal of Pain and Symptom Management 6(2), p. 53-57.

Gu, H., Ji, Q. (2004). An automated face reader for fatigue detection. In: FGR. (2004) 111–116.

Gu, H., Zhang, Y., Ji, Q. (2005). Task oriented facial behavior recognition with selective sensing. Comput. Vis. Image Underst. 100(3) p. 385–415.

Hadjistavropoulos HD, Craig KD, Hadjistavropoulos T, Poole GD. (1996). Subjective judgments of deception in pain expression: accuracy and errors. Pain. 65(2-3):251-8.

Hill, M.L., & Craig, K.D. (2004). Detecting deception in facial expressions of pain: Accuracy and training. *Clinical Journal of Pain, 20*, 415-422.

Hill ML, Craig KD (2002) Detecting deception in pain expressions: the structure of genuine and deceptive facial displays. Pain. 98(1-2):135-44.

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

Hong, Chung, K. (2005). Electroencephalographic study of drowsiness in simulated driving with sleep deprivation. International Journal of Industrial Ergonomics. 35(4), Pages 307-320.

Igarashi, K., Takeda, K., Itakura, F., Abut, H.: (2005). DSP for In-Vehicle and Mobile Systems. Springer US

Kapoor, A., Burleson, W., & Picard, R. (2007). Automatic prediction of frustration. International Journal of Human-Computer Studies, 65 (8), 724-736.

Larochette AC, Chambers CT, Craig KD (2006). Genuine, suppressed and faked facial expressions of pain in children. Pain. 2006 Dec 15;126(1-3):64-71.

Littlewort, Bartlett, & Lee (in press). Automatic coding of Facial Expressions displayed during Posed and Genuine Pain. Image and Vision Computing, 2009.

Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J. & Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *J. Image & Vision Computing,* Vol. 24, No. 6, pp. 615-625.

Morecraft RJ, Louie JL, Herrick JL, Stilwell-Morecraft KS. (2001). Cortical innervation of the facial nucleus in the non-human primate: a new interpretation of the effects of stroke and related subtotal brain trauma on the muscles of facial expression. Brain 124(Pt 1):176-208.

Orden, K.F.V., Jung, T.P., Makeig, S. (2000). Combined eye activity measures accurately estimate changes in sustained visual task performance. Biological Psychology 52(3):221-40.

Prkachin KM., Solomon, P.A. & Ross, A.J. (2007). The underestimation of pain among health-care providers. *Canadian Journal of Nursing Research*, 39, 88-106.

Prkachin KM, Schultz I, Berkowitz J, Hughes E, Hunt D. Assessing pain behaviour of low-back pain patients in real time: concurrent validity and examiner sensitivity. Behav Res Ther. 40(5):595-607.

Prkachin KM. (1992). The consistency of facial expressions of pain: a comparison across modalities. Pain. 51(3):297-306.

Rinn WE. The neuropsyhology of facial expression: a review of the neurological and psychological mechanisms for producing facial expression. Psychol Bull 95:52–77.

Schmidt KL, Cohn JF, Tian Y. (2003). Signal characteristics of spontaneous facial expressions: automatic movement in solitary and social smiles. Biol Psychol. 65(1):49-66.

Tada, H. (1986). Eyeblink rates as a function of the interest value of video stimuli. Tohoku Psychologica Folica, 45.

Takei, Y. Furukawa, Y. (2005): Estimate of driver's fatigue through steering motion. In: Man and Cybernetics, 2005 IEEE International Conference. Volume: 2, pg. 1765- 1770.

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

Vural, E., Cetin, M., Ercil, A., Littlewort, G., Bartlett, M., and Movellan, J. (2007). Drowsy driver detection through facial movement analysis. ICCV Workshop on Human Computer Interaction. © 2007 IEEE.

Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., & Movellan, J. (in press). Towards practical smile detection. Transactions on Pattern Analysis and Machine Intelligence, 2009.

Whitehill, J., Bartlett, M., and Movellan, J. (2008). Automated teacher feedback using facial expression recognition. Workshop on CVPR for Human Communicative Behavior Analysis, IEEE Conference on Computer Vision and Pattern Recognition.

Zhang, Z., shu Zhang, J. (2006). Driver fatigue detection based intelligent vehicle control. In: Proceedings of the 18th International Conference on Pattern Recognition, Washington, DC, USA, IEEE Computer Society p. 1262–1265.

Table 1. Z-score differences of the three pain conditions, averaged across subjects. FB: Fear brow 1+2+4. DB: Distress brow (1,1+4).

A. Real Pain vs baseline:

Action Unit	25	12	9	26	10	6
Z-score	1.4	1.4	1.3	1.2	0.9	0.9

B. Faked Pain vs Baseline:

Action Unit	4	DB	1	25	12	6	26	10	FB	9	20	7
Z-score	2.7	2.1	1.7	1.5	1.4	1.4	1.3	1.3	1.2	1.1	1.0	0.9

C. Real Pain vs Faked Pain:

Action Unit	4	DB	1
Z-score difference	1.8	1.7	1.0

Table 2. MLR model for predicting drowsiness across subjects. Predictive performance of each facial action individually is shown.

More when critically drowsy

AU	Name	A'
45	Blink/Eye Closure	0.94
2	Outer Brow Raise	0.81
15	Lip Corner	0.80
17	Depressor	0.79
9	Chin Raiser	0.78
30	Nose Wrinkle	0.76
20	Jaw Sideways	0.74
11	Lip stretch	0.71
14	Nasolabial Furrow	0.71
1	Dimpler	0.68
10	Inner Brow Raise	0.67
27	Upper Lip Raise	0.66
18	Mouth Stretch	0.66
22	Lip Pucker	0.64
24	Lip funneler	0.64
19	Lip presser	0.61
	Tongue show	

Less when critically drowsy

AU	Name	A'
12	Smile	0.87
7	Lid tighten	0.86
39	Nostril Compress	0.79
4	Brow lower	0.79
26	Jaw Drop	0.77
6	Cheek Raise	0.73
38	Nostril Dilate	0.72
23	Lip tighten	0.67
8	Lips toward	0.67
5	Upper lid raise	0.65
16	Upper lip depress	0.64
32	Bite	0.63

Table 3. Drowsiness detection performance for novel subjects, using a MLR classifier with different feature combinations. The weighted features are summed over 12 seconds before computing A'.

Feature	A'
AU45,AU2,AU19,AU26,AU15	.9792
All AU features	.8954

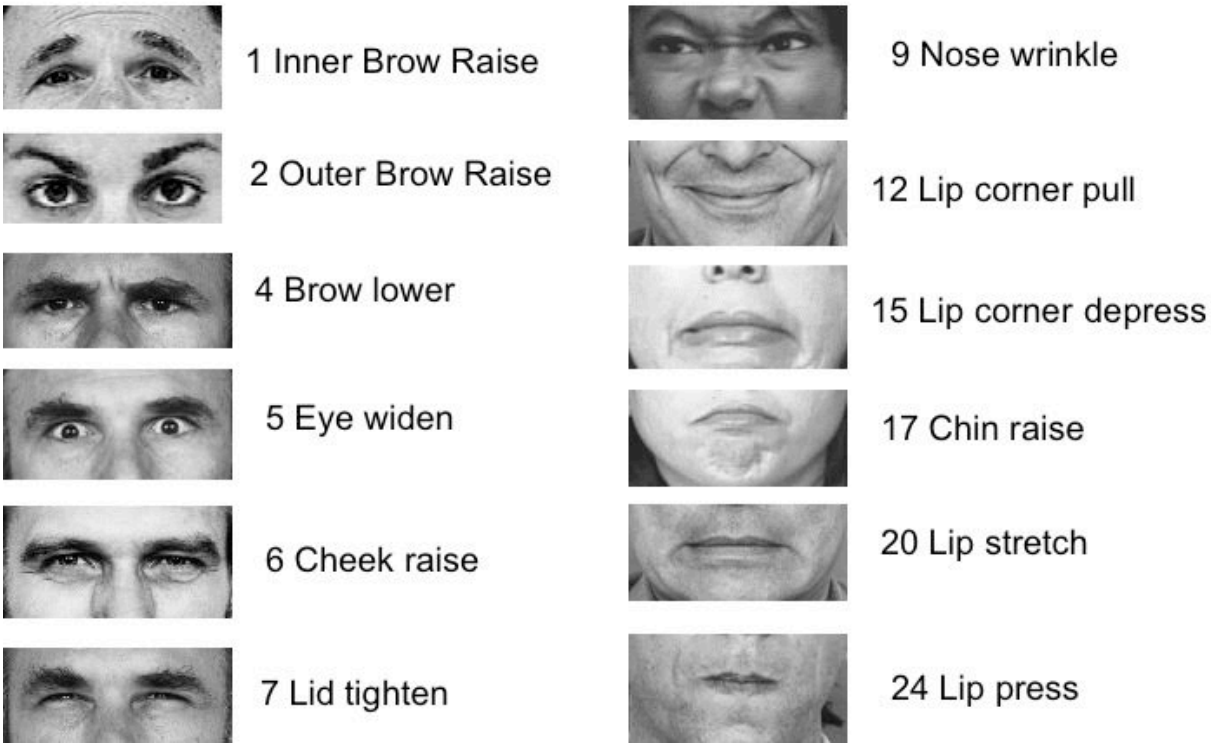
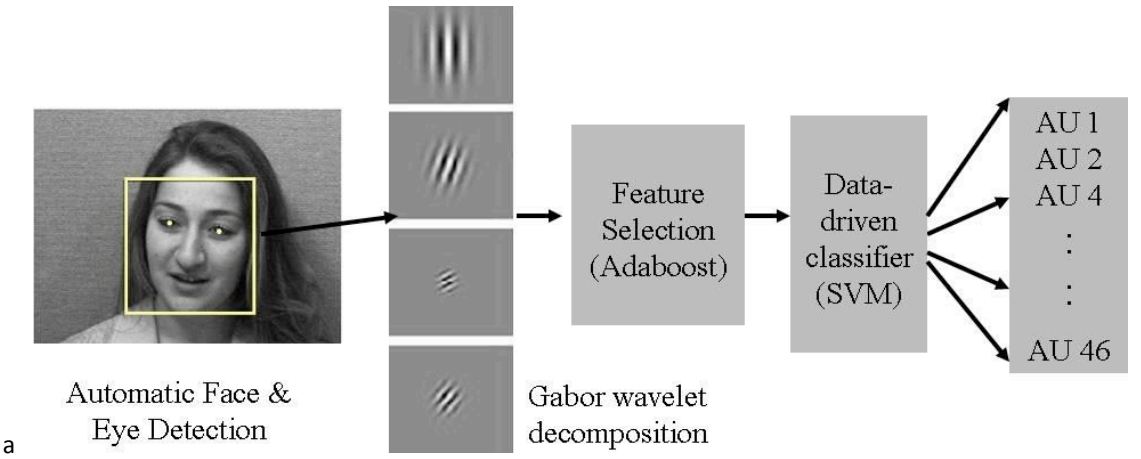


Figure 1. Sample facial actions from the Facial Action Coding System incorporated in CERT. CERT includes 30 total.



a

b

Figure 2. Computer Expression Recognition Toolbox (CERT). a. Overview of system design b. Example of CERT running on live video. Each subplot has time in the horizontal axis and the vertical axis indicates intensity of a particular facial movement.

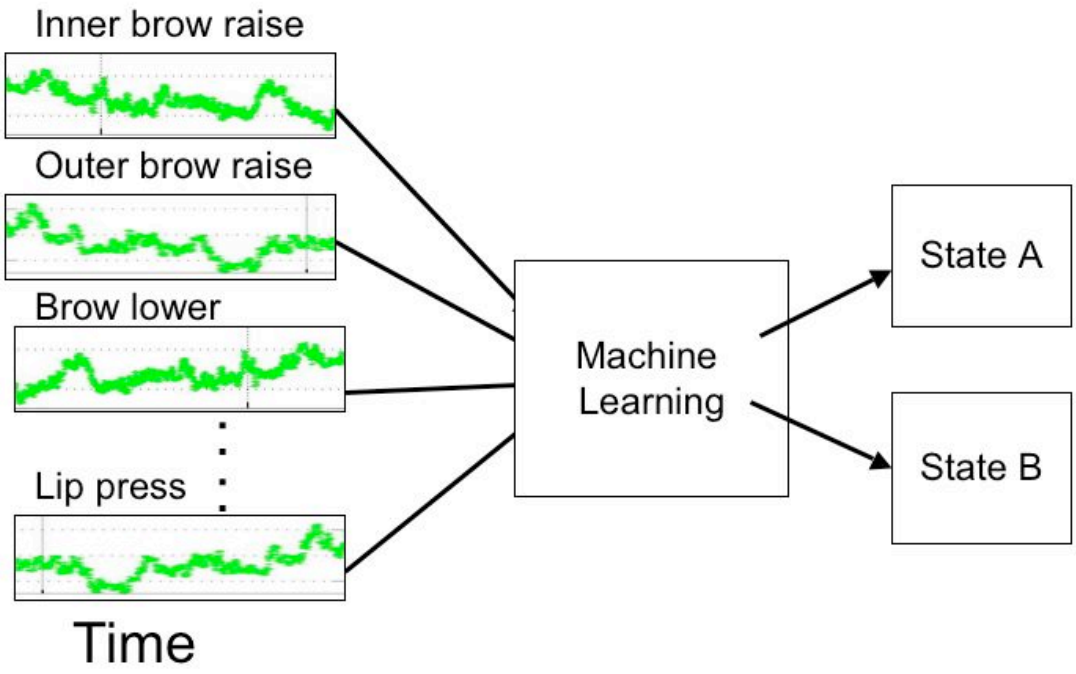


Figure 3. Data mining human behavior. CERT is applied to face videos containing spontaneous expressions of states in question. Machine learning is applied to the outputs of CERT to learn a classifier to automatically discriminate state A from State B.

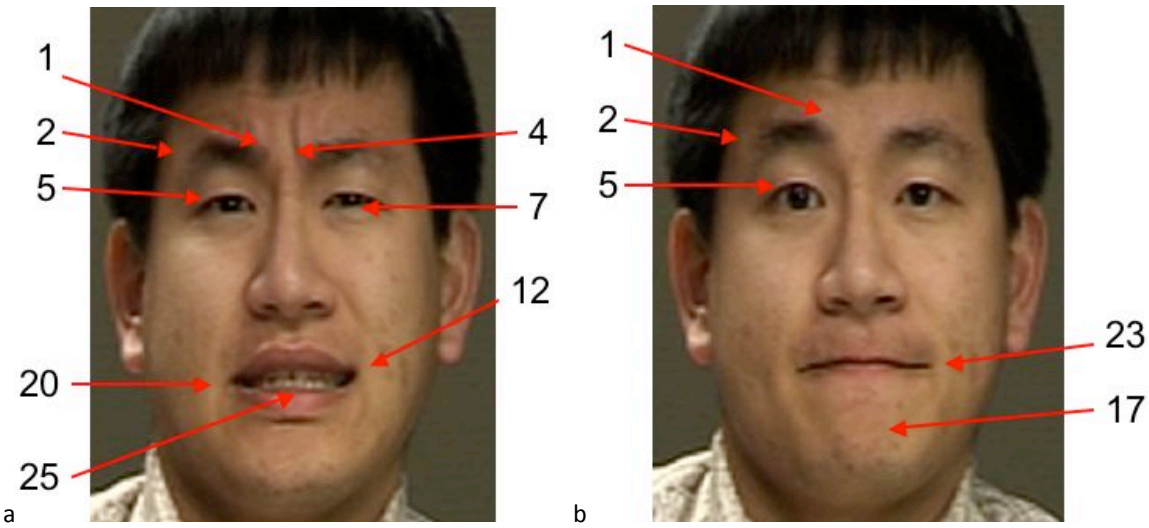


Figure 4. Sample facial behavior and facial action codes from the faked (a) and real pain (b) conditions.

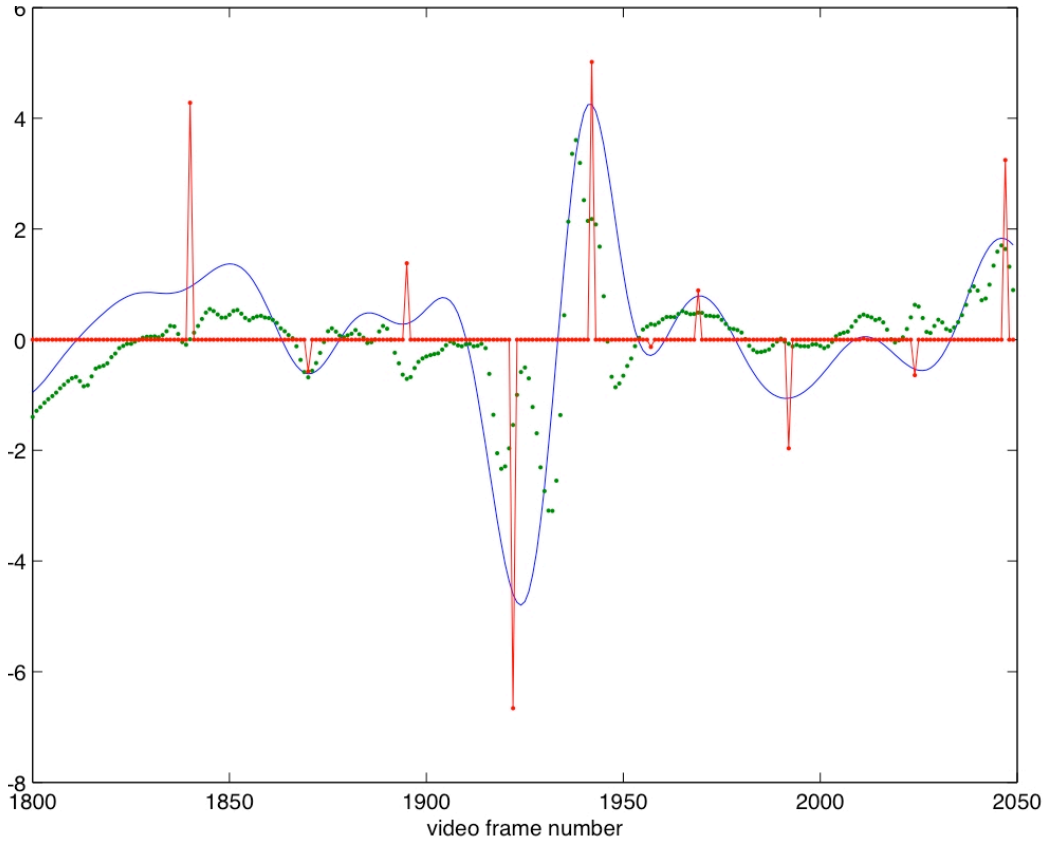


Figure 5. Example of Integrated Event Representation, for one subject, and one frequency band for AU 4. Green (.): Raw CERT output for AU4. Blue (-): DOG filtered signal for one frequency band. Red (-.): area under each curve.

Figure 6. Driving Simulation Task. Top: Screenshot of driving game. Center: Steering wheel. Bottom: Image of a subject in a drowsy state (left), and falling asleep (right).

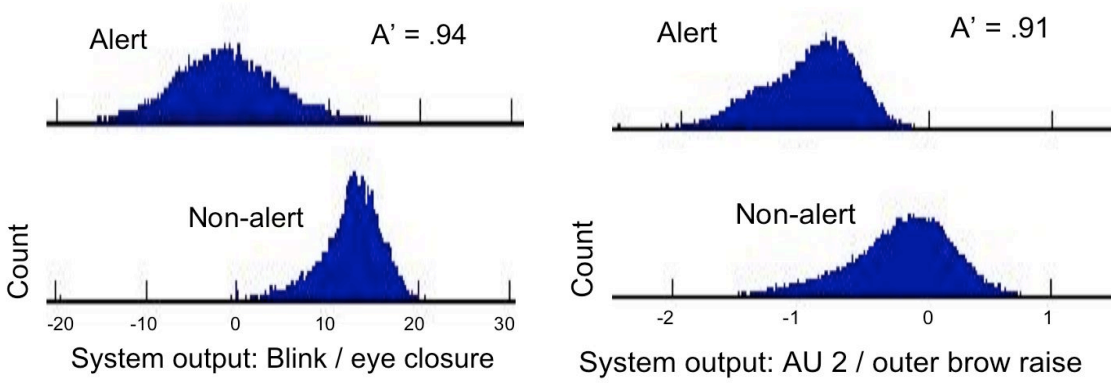


Figure 7. Example histograms for blink (left) and Action Unit 2 (right) in alert and non-alert states for one subject. A' is area under the ROC. The x-axis is CERT output.

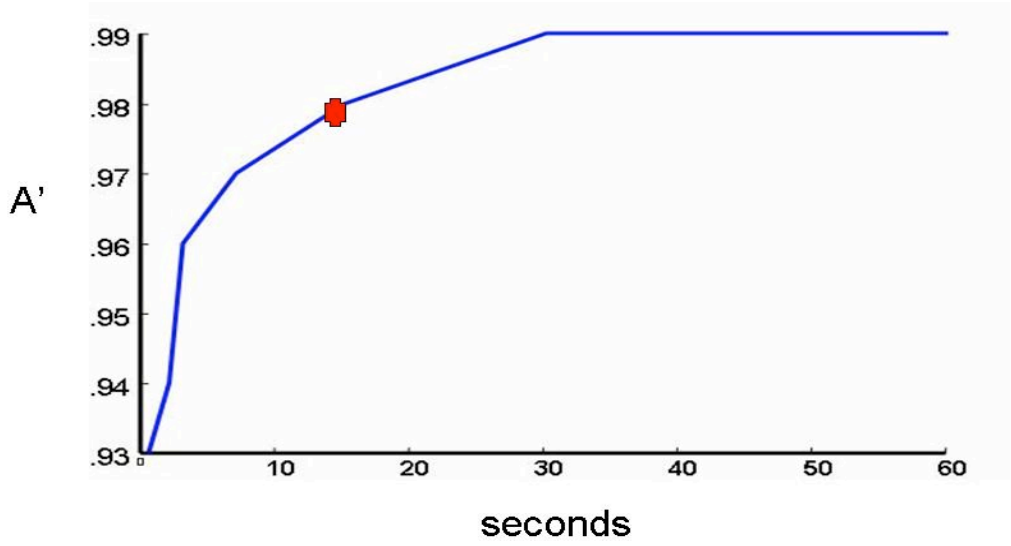


Figure 8. Performance for drowsiness detection in novel subjects over temporal window sizes.

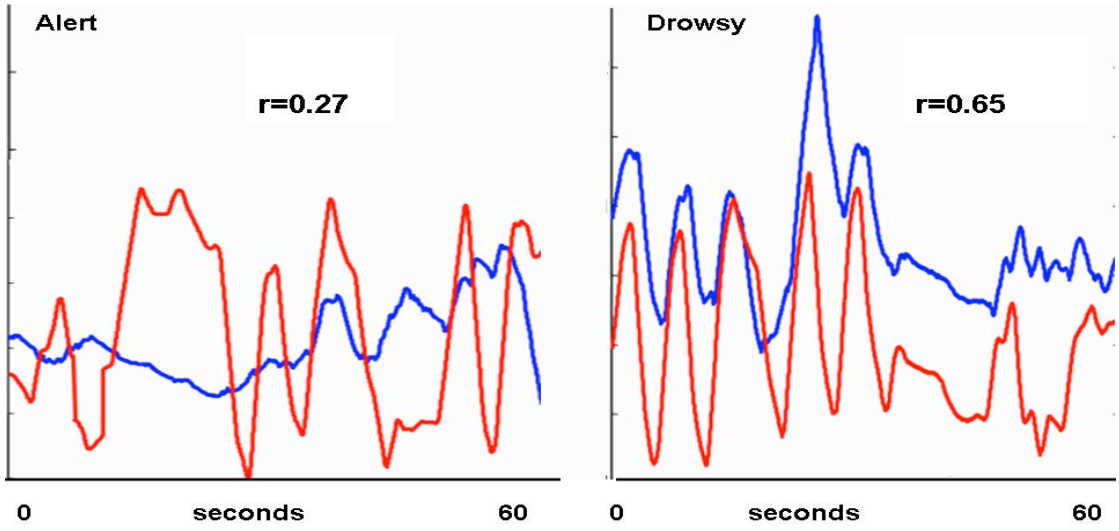


Figure 9. Head motion (blue/gray) and steering position (red/black) for 60 seconds in an alert state (left) and 60 seconds prior to a crash (right). Head motion is the output of the roll dimension of the accelerometer.

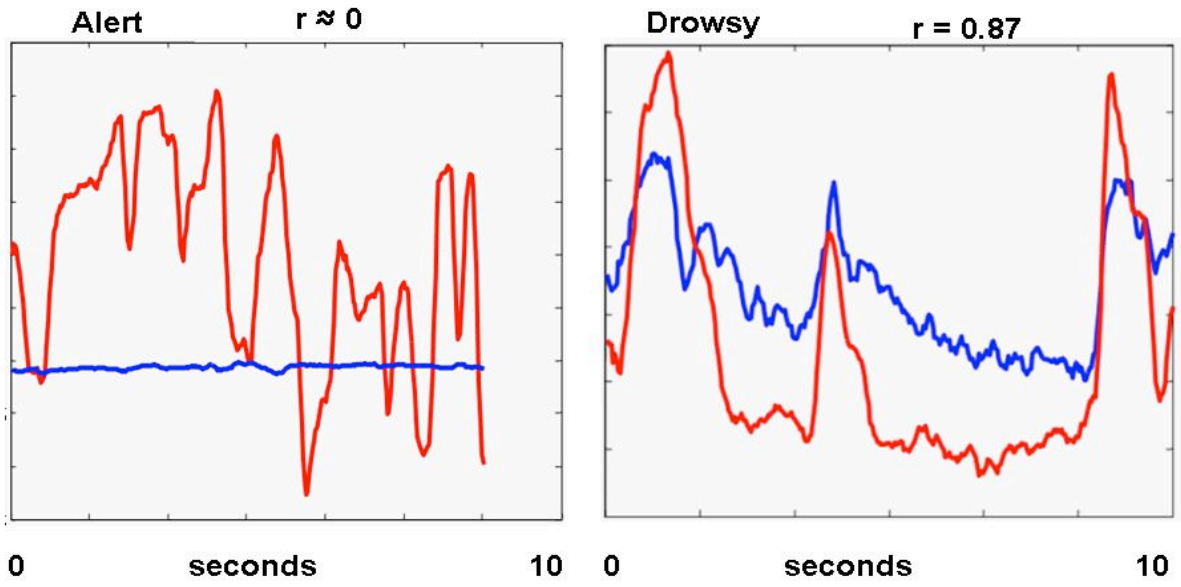
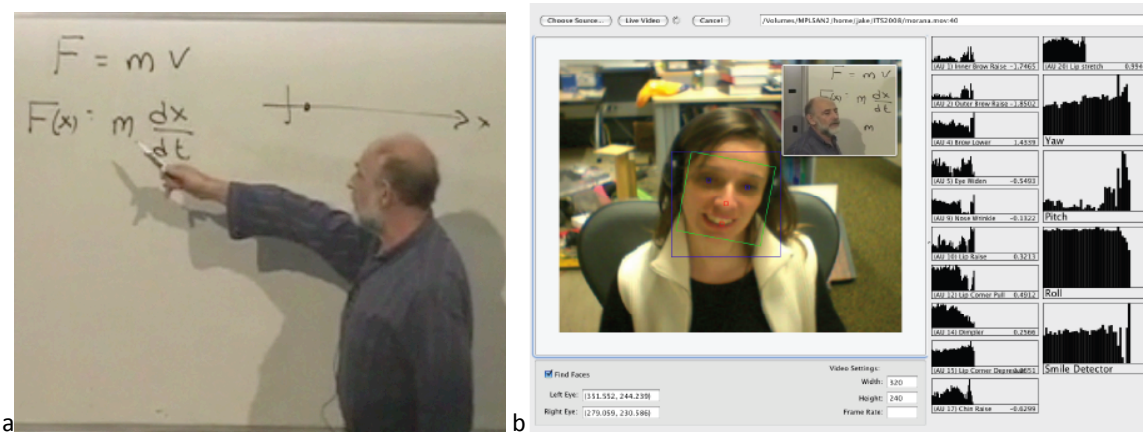
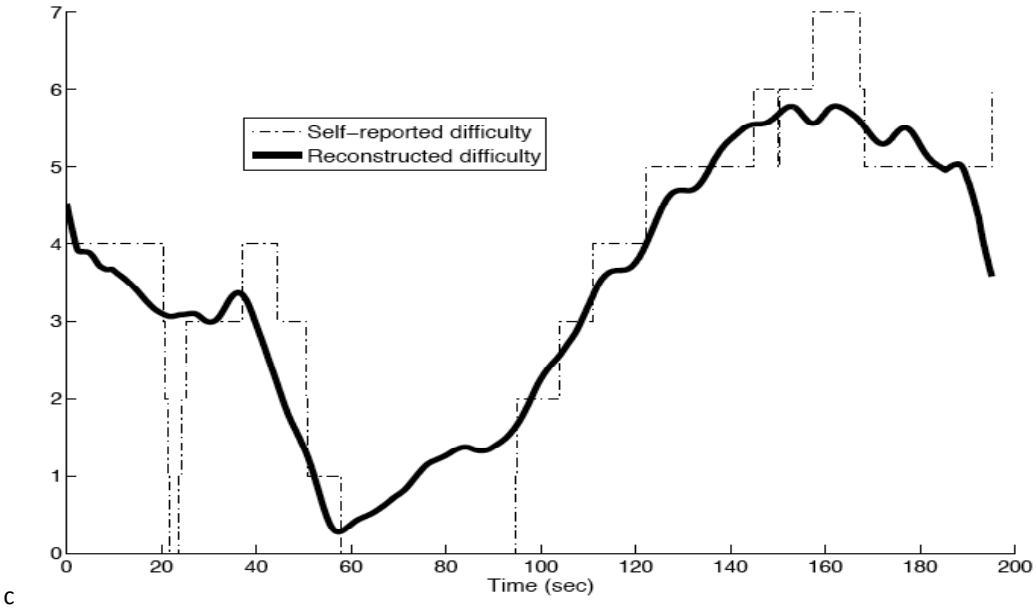


Figure 10. Eye Openness (red/black) and Eye Brow Raises (AU2) (blue/gray) for 10 seconds in an alert state (left) and 10 seconds prior to a crash (right).





c

Figure 11 a. Sample video lecture. b. Automated facial expression recognition is performed on subjects face as she watches the lecture. c. Self-reported difficulty values (dashed), and the reconstructed difficulty values (solid) computed using linear regression over the facial expression outputs for one subject (12 action units and smile detector). Reprinted with permission from Whitehill, Bartlett, & Movellan (2008). © 2008 IEEE.