The Computer Expression Recognition Toolbox (CERT)

Gwen Littlewort[1], Jacob Whitehill[1], Tingfan Wu[1], Ian Fasel[2],
Mark Frank[3], Javier Movellan[1], and Marian Bartlett[1]
{gwen, jake, ting, movellan}@mplab.ucsd.edu,
ianfasel@cs.arizona.edu, mfrank83@buffalo.edu, marni@salk.edu
1 Machine Perception Laboratory, University of California, San Diego
2 Department of Computer Science, University of Arizona
3 Department of Communication, University of Buffalo

***Abstract*— We present the Computer Expression Recognition Toolbox (CERT), a software tool for fully automatic real-time facial expression recognition, and officially release it for free academic use. CERT can automatically code the intensity of 19 different facial actions from the Facial Action Unit Coding System (FACS) and 6 different protoypical facial expressions. It also estimates the locations of 10 facial features as well as the 3-D orientation (yaw, pitch, roll) of the head. On a database of posed facial expressions, Extended Cohn-Kanade (CK+ [1]), CERT achieves an average recognition performance (probability of correctness on a two-alternative forced choice (2AFC) task between one positive and one negative example) of 90.1% when analyzing facial actions. On a spontaneous facial expression dataset, CERT achieves an accuracy of nearly 80%. In a standard dual core laptop, CERT can process 320×240 video images in real time at approximately 10 frames per second.**

I. INTRODUCTION

Facial expressions provide a wealth of information about a person's emotions, intentions, and other internal states [2]. The ability to recognize a person's facial expressions automatically and in real-time could give rise to a wide range of applications that we are only beginning to imagine.

The last decade has seen substantial progress in the field of automatic facial expression recognition systems (e.g., [3], [4], [1], [5], [6]). Such systems can operate reasonably accurately on novel subjects, exhibiting both spontaneous and posed facial expressions. This progress has been mainly enabled by the adoption of modern machine learning methods, and by the gathering of high-quality databases of facial expression necessary for using these methods (e.g., Cohn-Kanade [7], Extended Cohn-Kanade [8], MMI [9]). Systems for automatic expression recognition can interpret facial expression at the level of basic emotions [10] (happiness, sadness, anger, disgust, surprise, or fear), or they can analyze them at the level of individual muscle movements (facial "action units") of the face, in the manner of the Facial Action Coding System (FACS) [10].

To date, no fully automatic real-time system that recognizes FACS Action Units with state-of-the-art accuracy has been publicly available. In this paper, we present one such tool – the Computer Expression Recognition Toolbox (CERT). CERT is a fully automatic, real-time software tool that estimates facial expression both in terms of 19 FACS Action Units, as well as the 6 universal emotions. While the technology continues to advance, at this time CERT provides sufficiently accurate estimates of facial expression to enable real-world applications such as driver fatigue detection [11] and emotional reactivity such as pain reactions [12].

The objective of this paper is to announce the release of CERT to the research community, to provide a description of the technical components of CERT, and to provide benchmark performance data as a resource to accompany the Toolbox. The development of the various components of CERT has been published in previous papers. Here we provide a coherent description of CERT in a single paper with updated benchmarks.

Outline: We briefly describe the Facial Action Coding System in Section I-A, which defines the Action Units that CERT endeavors to recognize. We then present the software features offered by CERT in Section II and describe the system architecture. In Section IV-A we evaluate CERT's accuracy on several expression recognition datasets. In Section V we describe higher-level applications based on CERT that have recently emerged.

A. Facial Action Coding System (FACS)

In order to objectively capture the richness and complexity of facial expressions, behavioral scientists found it necessary to develop objective coding standards. The Facial Action Coding System (FACS) [10] is one of the most widely used expression coding system in the behavioral sciences. FACS was developed by Ekman and Friesen as a comprehensive method to objectively code facial expressions. 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 (AUs) 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, and the letters identify the level of activation. FACS provides an objective and comprehensive language for describing facial expressions and relating them back to what is known about their meaning from the behavioral science literature. Because it is comprehensive, FACS also allows for the discovery of new patterns related to emotional or situational states.

II. COMPUTER EXPRESSION RECOGNITION TOOLBOX (CERT)

The Computer Expression Recognition Toolbox (CERT) is a software tool for real-time fully automated coding of facial expression. It can process live video using a standard Web camera, video files, and individual images. CERT provides estimates of facial action unit intensities for 19 AUs, as well as probability estimates for the 6 prototypical emotions (happiness, sadness, surprise, anger, disgust, and fear). It also estimates the intensity of posed smiles, the 3-D head orientation (yaw, pitch, and roll), and the (x, y) locations of 10 facial feature points. All CERT outputs can be displayed within the GUI (see Figure 1) and can be written to a file (updated in real-time so as to enable secondary processing). For real time interactive applications CERT provides a sockets-based interface.

CERT's processing pipeline, from video to expression intensity estimates, is given in Figure 2. In the subsections below we describe each stage.

A. Face Detection

The CERT face detector was trained using an extension of the Viola-Jones approach [13], [14]. It employs GentleBoost [15] as the boosting algorithm and WaldBoost [16] for automatic cascade threshold selection. On the CMU+MIT dataset, CERT's face detector achieves a hit rate of 80.6% with 58 false alarms. At run-time, the face detector is applied to each video frame, and only the largest found face is segmented for further processing. The output of the face detector is shown in blue in Figure 1.

B. Facial Feature Detection

After the initial face segmentation, a set of 10 facial features, consisting of inner and outer eye corners, eye centers, tip of the nose, inner and outer mouth corners, and center of the mouth, are detected within the face region using feature-specific detectors (see [17]). Each facial feature detector, trained using GentleBoost, outputs the log-likelihood ratio of that feature being present at a location (x, y) within the face, to being not present at that location. This likelihood term is combined with a feature-specific prior over (x, y) locations within the face to estimate the posterior probability of each feature being present at (x, y) given the image pixels.

Given the initial constellation of the (x, y) locations of the 10 facial features, the location estimates are refined using linear regression. The regressor was trained on the GENKI dataset [18], which was labeled by human coders for the positions of all facial features. The outputs of the facial feature detectors are shown in small red boxes (except the eye centers, which are blue) within the face in Figure 1.

C. Face Registration

Given the set of 10 facial feature positions, the face patch is re-estimated at a canonical size of 96x96 pixels using an affine warp. The warp parameters are computed to minimize the L2 norm between the warped facial feature positions of the input face and a set of canonical feature point positions computed over the GENKI datset. The pixels of this face patch are then extracted into a 2-D array and are used for further processing. In Figure 1 the re-estimated face box is shown in green.

D. Feature Extraction

The cropped 96x96-pixel face patch is then convolved (using a Fast Fourier Transform) with a filter bank of 72 complex-valued Gabor filters of 8 orientations and 9 spatial frequencies (2:32 pixels per cycle at 1/2 octave steps). The magnitudes of the complex filter outputs are concatenated into a single feature vector.

E. Action Unit Recognition

The feature vector computed in the previous stage is input to a separate linear support vector machine (SVM) for each AU. The SVM outputs can be interpreted as estimates of the AU intensities (see Section II-F).

The action unit SVMs were trained from a compilation of several databases: Cohn-Kanade [7], Ekman-Hager [19], M3 [20], Man-Machine Interaction (MMI) [9], and two non-public datasets collected by the United States government which are similar in nature to M3. Cohn-Kanade and Ekman-Hager are databases of posed facial expression, whereas the M3 and the two government datasets contained spontaneous expressions. From the MMI dataset, only posed expressions were used for training. For AUs 1, 2, 4, 5, 9, 10, 12, 14, 15, 17, and 20, all of the databases listed above were used for training. For AUs 6, 7, 18, 23, 24, 25, and 26, only Cohn-Kanade, Ekman-Hager, and M3 were used. The number of positive training examples for each AU is given by the column "N_p train" in Table I.

F. Expression Intensity and Dynamics

For each AU, CERT outputs a continuous value for each frame of video, consisting of the distance of the input feature vector to the SVM's separating hyperplane for that action unit. Empirically it was found that CERT outputs are significantly correlated with the intensities of the facial actions, as measured by FACS expert intensity codes [5]. 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 of concurrent validity with EMG.

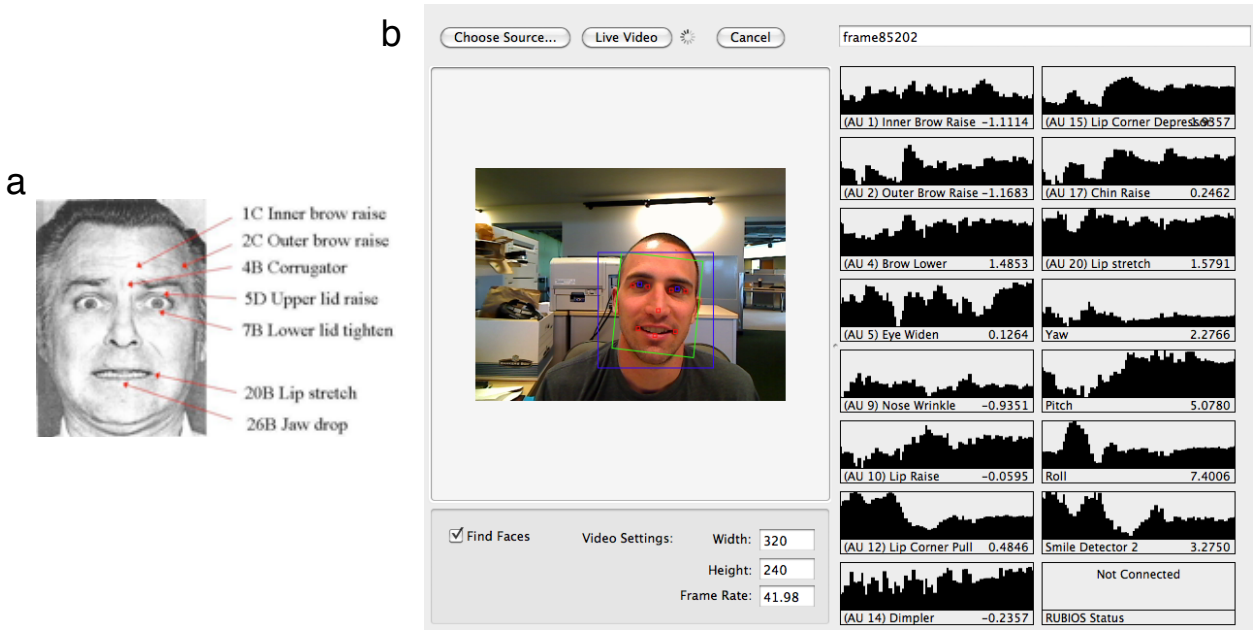

Fig. 1. (a) Example of comprehensive Facial Action Coding System (FACS) coding of a facial expression. The numbers identify the action unit, which approximately corresponds to one facial muscle; the letter identifies the level of activation. (b) Screenshot of CERT.

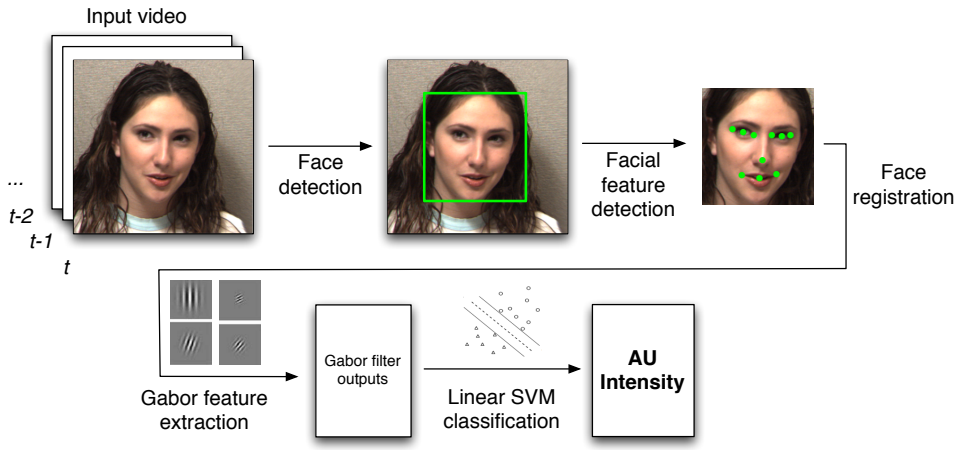

Fig. 2. Processing pipeline of the Computer Expression Recognition Toolbox (CERT) from video to expression intensity estimates.

CERT outputs significantly correlated with EMG measures of zygomatic and corrugator activity despite the visibility of the electrodes in the video [21].

III. EXTENSION MODULES

The CERT architecture allows for extension modules that can intercept the processing pipeline at several possible points, including just after the face registration stage, and after all AUs have been recognized (the endpoint). This allows for the implementation of three particular modules that are part of CERT – a detector of posed smiles, a 3-D head pose estimator, and a basic emotion recognizer. These are described below. Other secondary processing applications of CERT's AU outputs will be discussed in Section V.

A. Smile Detection

Since smiles play such an important role in social interaction, CERT provides multiple ways of encoding them. In addition to AU 12 (lip corner puller, present in all smiles), CERT is also equipped with a smile detector that was trained on a subset of 20,000 images from the GENKI dataset [18]. These were images of faces obtained from the Web representing a wide variety of imaging conditions and geographical locations. The smile detector utilizes the same processing pipeline as the AU detectors up through the face registration stage. Instead of using Gabor filters (as for action unit recognition), the smile detector extracts Haar-like box filter features, and then uses GentleBoost to classify the resulting feature vector into $\{\text{Smile}, \text{NonSmile}\}$. Smile detection accuracy (2AFC) on a subset of GENKI not used for training was 97.9%. In addition, the smile detector outputs were found to be significantly correlated with human judgments of smile intensity (Pearson $r = 0.894$) [22]. Comparisons of Haar+GentleBoost versus Gabor+SVMs showed that the former approach is faster and yields slightly higher accuracy for the smile detection problem [22].

B. Pose Estimation

CERT also outputs estimates of the 3-D head orientation. After the face-registration stage, the patch of face pixels are passed through an array of pose range classifiers that are trained to distinguish between different ranges of yaw, pitch, and roll (see [23]). Two types of such classifiers are used: 1-versus-1 classifiers that distinguish between two disjoint pose ranges (e.g., $[6, 18)^\circ$, $[18, 30)^\circ$); and 1-versus-all classifiers that distinguish between one pose range and all other pose ranges. The pose range discriminators were trained using GentleBoost on Haar-like box features and output the log probability ratio of the face belonging to one pose range class compared to another. These detectors' outputs are combined with the (x, y) coordinates of all 10 facial feature detectors (Section II-D) and then passed through a linear regressor to estimate the real-valued angle of each of the yaw, pitch, and roll parameters.

Accuracy of the pose detectors was measured on the GENKI 4K dataset (not used for training) [24]; see Figure 3 for Root Mean Square Error (RMSE) of pose estimation as a function of human-labeled pose.

C. Basic Emotion Recognition

Since CERT exports a real-time stream of estimated AU intensities, these values can then be utilized by second-layer recognition systems in a variety of application domains. One such application is the recognition of basic emotions. CERT implements a set of 6 basic emotion detectors, plus neutral expression, by feeding the final AU estimates into a multivariate logistic regression (MLR) classifier. The classifier was trained on the AU intensities, as estimated by CERT, on the Cohn-Kanade dataset and its corresponding ground-truth emotion labels. MLR outputs the posterior probability of each emotion given the AU intensities as inputs. Performance of the basic emotion detectors is discussed in Section IV-A.

IV. EXPERIMENTAL EVALUATION

We evaluated CERT's AU recognition performance on two high-quality databases of facial expression: the Extended Cohn-Kanade Dataset, containing posed facial expressions, and the M3 Dataset, containing spontaneous facial expressions. We measure accuracy as the probability of correctness in discriminating between a randomly drawn positive example (in which a particular AU is present) and a random negative example (in which the AU is not present) based on the real-valued classifier output. We call this accuracy statistic the 2AFC Score (two alternative forced choice). Under mild conditions it is mathematically equivalent to the area under the Receiver Operating Characteristics curve, which is sometimes called the the A' statistic (e.g., [8]). An estimate of the standard error associated with estimating the 2AFC value can be computed as

$$\widehat{se} = \sqrt{\frac{p(1-p)}{\min\{N_p, N_n\}}}$$

where p is the 2AFC value and N_p and N_n are the number of positive and negative examples, respectively, for each particular AU [22].

A. Extended Cohn-Kanade Dataset (CK+)

We evaluated CERT on the Extended Cohn-Kanade Dataset (CK+) [8]. Since CK+ is a superset of the original Cohn-Kanade Dataset (CK) [7], and since CERT was trained partially on CK, we restricted our performance evaluation to only those subjects of CK+ not included in CK. These were subject numbers: 5, 28, 29, 90, 126, 128, 129, 139, 147, 148, 149, 151, 154, 155, 156, 157, 158, 160, 501, 502, 503, 504, 505, 506, 895, and 999.

Our evaluation procedure was as follows: For each video session of each of the 26 subjects listed above, we used CERT to estimate the AU intensity for the first frame (containing a neutral expression) and the last frame (containing the expression at peak intensity). The first frames constituted negative examples for all AUs, while the last frame constituted positive examples for those AUs labeled in CK+ as present and negative examples for all other AUs. From the real-valued AU intensity estimates output by CERT, we then calculated for each AU the 2AFC statistic and

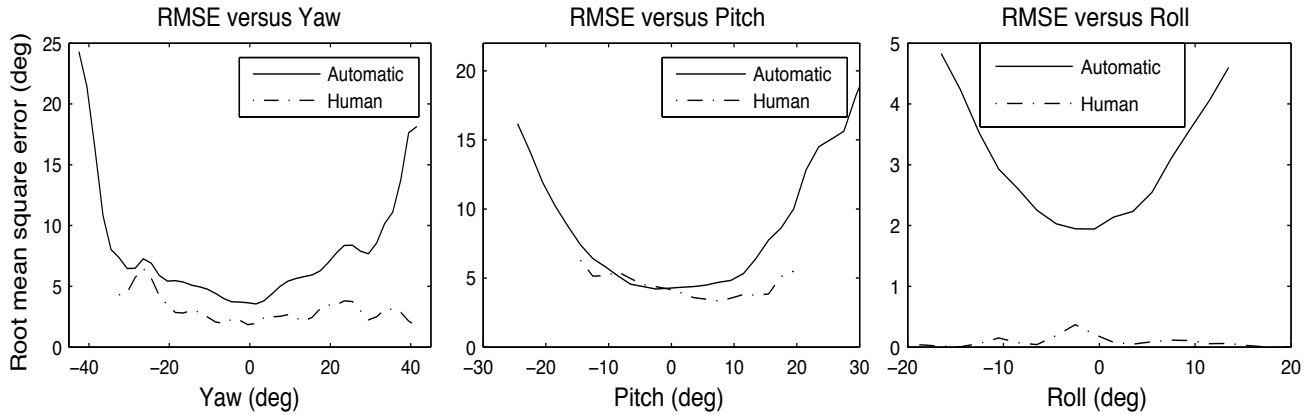

Fig. 3. Smoothed root-mean-square errors (RMSE), as a function of human-labeled pose, for both the automatic pose tracker and the individual human labelers. RMSE for the automatic pose tracker was estimated over GENKI-4K using the average human labeler's pose as ground-truth. RMSE for humans was measured on a different subset of GENKI comprising 671 images on which at least 4 different humans had labeled pose.

Performance on CK+

AU	N_p **train**	N_p **test**	$2AFC(\%) \pm \widehat{se}$
1	2186	14	$\mathbf{97.5} \pm 4.1$
2	1848	9	$\mathbf{87.1} \pm 11.2$
4	1032	23	$\mathbf{97.4} \pm 3.3$
5	436	14	$\mathbf{87.0} \pm 9.0$
6	278	6	$\mathbf{80.2} \pm 16.3$
7	403	9	$\mathbf{89.1} \pm 10.4$
9	116	5	$\mathbf{100.0} \pm 0.0$
10	541	2	$\mathbf{86.8} \pm 23.9$
12	1794	8	$\mathbf{92.4} \pm 9.4$
14	909	22	$\mathbf{91.0} \pm 6.1$
15	505	14	$\mathbf{91.0} \pm 7.6$
17	1370	31	$\mathbf{89.0} \pm 5.6$
18	121	1	$\mathbf{93.0} \pm 25.4$
20	275	6	$\mathbf{91.1} \pm 11.6$
23	57	9	$\mathbf{81.3} \pm 13.0$
24	49	3	$\mathbf{96.8} \pm 10.2$
25	376	11	$\mathbf{90.7} \pm 8.7$
26	86	7	$\mathbf{69.5} \pm 17.4$
Avg			$\mathbf{90.1}$

TABLE I

CERT'S AU RECOGNITION ACCURACY ON THE 26 SUBJECTS OF THE EXTENDED COHN-KANADE DATASET (CK+) NOT INCLUDED IN THE ORIGINAL COHN-KANADE DATASET (CK).

standard error. An average 2AFC over all AUs, weighted by the number of positive examples for each AU, was also calculated. Results are shown in Table I.

We also assessed the accuracy of CERT's prototypical emotion recognition module (Section III-C) on the same 26 subjects in CK+ not in CK. We measured accuracy in two different ways: (a) using the 2AFC statistic when discriminating images of each emotion i from images of all other emotions $\{1, \ldots, 7\} \backslash \{i\}$, and (b) as the percent-correct classification of each image on a seven-alternative forced choice (among all 7 emotions). The test set consisted of 86 frames – all the first (neutral) and last (apex) frames from each of the 26 subjects whose emotion was one of happiness, sadness, anger, fear, surprise, disgust, or neutral. For (a), the individual 2AFC scores were 93.5, 100.0, 100.0, 100.0, 100.0, 100.0, and 97.94 for the emotions as listed above; the average 2AFC was 98.8%. For (b), a confusion table is given in Table II. The row labels are ground-truth, and the column labels are the automated classification results. The seven-alternative forced choice performance was 87.21%.

Emotion Classification Confusion Matrix

	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**	**Ne**
An	**36.4**	9.1	0.0	0.0	0.0	0.0	54.5
Di	0.0	**100.0**	0.0	0.0	0.0	0.0	0.0
Fe	0.0	0.0	**60.0**	0.0	0.0	40.0	0.0
Ha	0.0	0.0	0.0	**100.0**	0.0	0.0	0.0
Sa	0.0	0.0	0.0	0.0	**60.0**	0.0	40.0
Su	0.0	0.0	0.0	0.0	0.0	**100.0**	0.0
Ne	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**

TABLE II

SEVEN-ALTERNATIVE FORCED CHOICE EMOTION CLASSIFICATION OF THE 26 SUBJECTS OF THE CK+ DATASET NOT IN CK.

B. M3 Dataset

The M3 [20] is a database of spontaneous facial behavior that was FACS coded by certified FACS experts. The dataset consists of 100 subjects participating in a "false opinion" paradigm. In this paradigm, subjects first fill out

a questionnaire regarding their opinions about a social or political issue. Subjects are then asked to either tell the truth or take the opposite opinion on an issue on which they rated strong feelings, and convince an interviewer they are telling the truth. This paradigm has been shown to elicit a wide range of emotional expressions as well as speech-related facial expressions [25]. The dataset was collected from four synchronized Dragonfly video cameras from Point Grey. M3 can be considered a particularly challenging dataset due to the typically lower intensity of spontaneous compared to posed expressions, the presence of speech-related mouth movements, and the out-of-plane head rotations that tend to be present during discourse.

In earlier work [5], we trained a FACS recognition system on databases of posed expressions and measured its accuracy on the frontal video stream of M3. In contrast, here we present results based on training data with both posed and spontaneous facial expressions. The evaluation procedure was as follows: M3 subjects were divided into three disjoint validation folds. When testing on each fold i, the corresponding subjects from fold i were *removed* from the CERT training set described in Section II-E. The re-trained CERT was then evaluated on each video frame on all subjects of fold i. 2AFC statistics and corresponding standard errors for each AU, along with the total number of positive examples (defined as the number of onset-apex-offset action unit events in video) of each AU occurring in the entire M3 dataset (over all folds), are shown in Table III. The average over all AUs, weighted by the number of positive examples for each AU (as in [8]), was also calculated.

Performance on M3

AU	N_p test	$2AFC(\%) \pm \widehat{se}$
1	169	$\mathbf{82.3} \pm 0.8$
2	153	$\mathbf{81.2} \pm 2.8$
4	32	$\mathbf{75.6} \pm 3.9$
5	36	$\mathbf{82.8} \pm 2.8$
6	50	$\mathbf{95.5} \pm 1.4$
7	46	$\mathbf{77.3} \pm 3.3$
9	2	$\mathbf{86.5} \pm 6.1$
10	38	$\mathbf{73.1} \pm 3.6$
12	3	$\mathbf{90.1} \pm 1.8$
14	119	$\mathbf{74.4} \pm 0.5$
15	87	$\mathbf{83.1} \pm 4.1$
17	77	$\mathbf{84.0} \pm 2.4$
18	121	$\mathbf{78.0} \pm 4.9$
20	12	$\mathbf{64.5} \pm 5.0$
23	24	$\mathbf{74.0} \pm 5.2$
24	68	$\mathbf{83.0} \pm 2.0$
25	200	$\mathbf{76.8} \pm 5.3$
26	144	$\mathbf{80.1} \pm 6.9$
Avg		**79.9**

TABLE III
CERT'S AU RECOGNITION ACCURACY ON THE M3 DATASET OF SPONTANEOUS FACIAL EXPRESSIONS, USING 3-FOLD CROSS-VALIDATION (SEE SECTION IV-B). N_p REFERS TO NUMBER OF AU EVENTS IN THE VIDEO, NOT NUMBER OF VIDEO FRAMES.

V. APPLICATIONS

The adoption of and continued improvement to real-time expression recognition systems such as CERT will make possible a broad range of applications whose scope we are only beginning to imagine. As described in Section II-F CERT's real-time outputs enable the study of facial expression *dynamics*. Below we describe two example projects utilizing CERT as the back-end system for two different application domains.

A. Automated Detection of Driver Fatigue

It is estimated that driver drowsiness causes more fatal crashes in the United States than drunk driving [26]. Hence an automated system that could detect drowsiness and alert the driver or truck dispatcher could potentially save many lives. Previous approaches to drowsiness detection by computer make assumptions about the relevant behavior, focusing on blink rate, eye closure, yawning, and head nods [27]. While there is considerable empirical evidence that blink rate can predict falling asleep, it was unknown whether there were other facial behaviors that could predict sleep episodes. Vural, et. al [11] employ a machine learning architecture to recognizing drowsiness in real human behavior.

In this study, four subjects participated in a driving simulation task over a 3 hour period between midnight and 3AM. Videos of the subjects faces, accelerometer readings of the head, and crash events were recorded in synchrony. The subjects' data were partitioned into drowsy and alert states as follows: The one minute preceding a crash was labeled as a drowsy state. A set of "alert" video segments was identified from the first 20 minutes of the task in which there were no crashes by any subject. This resulted in a mean of 14 alert segments and 24 crash segments per subject. The subjects' videos were analyzed frame-by-frame for AU intensity using CERT.

In order to understand how each action unit is associated with drowsiness across different subjects, a Multinomial Logistic Ridge Regressor (MLR) was trained on each facial action individually. The five most predictive facial actions whose intensities increased in drowsy states were blink, outer brow raise, frown, chin raise, and nose wrinkle. The five most predictive actions that decreased in intensity in drowsy states were smile, lid tighten, nostril compress, brow lower,

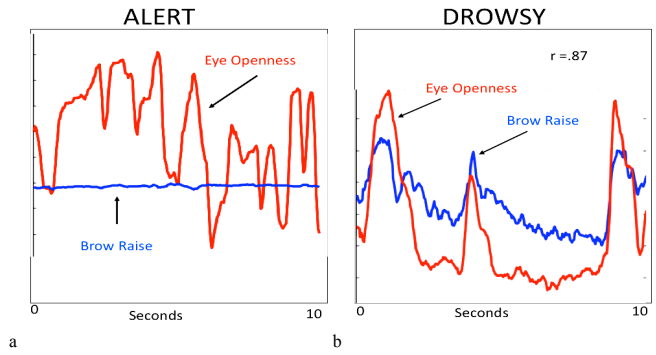

Fig. 4. Changes in movement coupling with drowsiniess. a,b: Eye Openness (red) and Eye Brow Raise (AU2) (Blue) for 10 seconds in an alert state (a) and 10 seconds prior to a crash (b), for one subject.

and jaw drop. The high predictive ability of the blink/eye closure measure was expected. However the predictability of the outer brow raise was previously unknown. It was observed during this study that many subjects raised their eyebrows in an attempt to keep their eyes open. Also of note is that AU 26, jaw drop, which occurs during yawning, actually occurred less often in the critical 60 seconds prior to a crash.

A fatigue detector that combines multiple AUs was then developed. An MLR classifier was trained using contingent feature selection, starting with the most discriminative feature (blink), and then iteratively adding the next most discriminative feature given the features already selected. MLR outputs were then temporally integrated over a 12 second window. Best performance of 98% (2AFC) was obtained with five features.

Changes were also observed in the coupling of behaviours with drowsiness. For some of the subjects, coupling between brow raise and eye openness increased in the drowsy state (Figure 4 a,b). Subjects appear to have pulled up their eyebrows in an attempt to keep their eyes open. This is the first work to our knowledge to reveal significant associations between facial expression and fatigue beyond eyeblinks. 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 just before falling asleep, drivers may yawn less often, not more often. This highlights the importance of designing a system around real, not posed, examples of examples of fatigue and drowsiness.

B. Automated Teaching Systems

There has been a growing thrust to develop tutoring systems and agents that respond to students' emotional and cognitive state and interact with them in a social manner (e.g., [28], [29]). Whitehill, et al. [30] conducted a pilot experiment in which expression was used 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.

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 controlled by the subject using a keypress. 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 CERT. After watching the video and taking the quiz, each subject then watched the lecture video again at a fixed speed of 1.0x. During this second viewing, subjects specified how easy or difficult they found the lecture to be at each moment in time using the keyboard.

For each subject, a regression analysis was performed to predict perceived difficulty and preferred viewing speed from the facial expression measures. The expression intensities, as well as their first temporal derivatives (measuring the instantaneous change in intensity), were the independent variables in a standard linear regression. The facial expression measures were significantly predictive of both perceived difficulty ($r = .75$) and preferred viewing speed ($r = .51$). The correlations on validation data were 0.42 and 0.29, respectively. The specific facial expressions that were correlated with difficulty and speed varied highly from subject to subject. The most consistently correlated expression was AU 45 ("blink"), where 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 [31].

Overall, this study provided proof of principle 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 recognition system was able to extract a signal from the face video in real-time that provided information about internal states relevant to teaching and learning.

VI. DIRECTIONS FOR FURTHER RESEARCH

While state-of-the-art expression classifiers such as CERT are already finding practical applications, as described above, much room for improvement remains. Some of the most pressing issues are generalizing to non-frontal head poses, providing good performance across a broader range of ethnicities, and the development of learning algorithms that can benefit from unlabeled or weakly labeled datasets.

A. Obtaining a Free Academic License

CERT is available to the research community. Distribution is being managed by Machine Perception Technologies, Inc. CERT is being released under the name AFECT (Automatic Facial Expression Coding Tool). The software is available for free for academic use. Information about obtaining a copy is available at `http://mpt4u.com/AFECT`.

ACKNOWLEDGEMENT

Support for this work was provided by NSF grants SBE-0542013, IIS-0905622, CNS-0454233, NSF IIS INT2-Large 0808767, and NSF ADVANCE award 0340851. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, and J.F. Cohn. AAM derived face representations for robust facial action recognition. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 155–160, 2006.

[2] D. Keltner and P. Ekman. Facial expression of emotion. In M. Lewis and J. Haviland-Jones, editors, *Handbook of emotions*. Guilford Publications, Inc., New York, 2000.

[3] Sander Koelstra, Maja Pantic, and Ioannis Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *Pattern Analysis and Machine Intelligence*, 2010.

[4] Peng Yang, Qingshan Liu, and Dimitris N. Metaxas. Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters*, 30:132–139, 2009.

[5] 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.

[6] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007.

[7] Takeo Kanade, Jeffrey Cohn, and Ying Li Tian. Comprehensive database for facial expression analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

[8] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshop on Human-Communicative Behavior*, 2010.

[9] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *International Conference on Multimedia and Expo*, 2005.

[10] 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.

[11] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. R. Movellan. Drowsy driver detection through facial movement analysis. *ICCV*, 2007.

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

[13] Ian Fasel, Bret Fortenberry, and J. R. Movellan. A generative framework for real-time object detection and classification. *Computer Vision and Image Understanding*, 98(1):182–210, 2005.

[14] Paul Viola and Michael Jones. Robust real-time face detection. *International Journal of Computer Vision*, 2004.

[15] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28(2), 2000.

[16] Jan Sochman and Jiti Matas. Waldboost: Learning for time constrained sequential detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:150–156, 2005.

[17] M. Eckhardt, I. Fasel, and J. Movellan. Towards practical facial feature detection. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(3):379–400, 2009.

[18] `http://mplab.ucsd.edu`. The MPLab GENKI Database.

[19] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.

[20] M.G. Frank, M.S. Bartlett, and J.R. Movellan. The M3 database of spontaneous emotion expression (University of Buffalo). *In prep*, 2010.

[21] M. Pierce, J. Cockburn, I. Gordon, S. Butler, L. Dison, and J. Tanaka. Perceptual and motor learning in the recognition and production of dynamic facial expressions. In *All Hands Meeting of the Temporal Dynamics of Learning Center, UCSD*, 2009.

[22] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier R. Movellan. Toward practical smile detection. *Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[23] Jacob Whitehill and Javier R. Movellan. A discriminative approach to frame-by-frame head pose estimation. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[24] `http://mplab.ucsd.edu`. The MPLab GENKI Database, GENKI-4K Subset.

[25] M.G. Frank and P. Ekman. The ability to detect deceit generalizes across different types of high stake lies. *Journal of personality and social psychology*, 27:1429–1439.

[26] Department of Transportation. Saving lives through advanced vehicle safety technology, 2001.

[27] H. Gu and Q. Ji. An automated face reader for fatigue detection. In *Proc. Int. Conference on Automated Face and Gesture Recognition*, pages 111–116, 2004.

[28] A. Kapoor, W. Burleson, and R. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736.

[29] 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.

[30] Jacob Whitehill, Marian Bartlett, and Javier Movellan. Automatic facial expression recognition for intelligent tutoring systems. *Computer Vision and Pattern Recognition Workshop on Human-Communicative Behavior*, 2008.

[31] M.K. Holland and G. Tarlow. Blinking and thinking. *Perceptual and Motor Skills*, 41, 1975.