The motion in emotion – a CERT based approach to the FERA emotion challenge

Gwen Littlewort, Jacob Whitehill, Ting-Fan Wu, Nicholas Butko, Paul Ruvolo, Javier Movellan, Marian Bartlett
Machine Perception Laboratory, Institute for Neural Computation
University of California, San Diego
gwen@mplab.ucsd.edu

***Abstract*—This paper assesses the performance of measures of facial expression dynamics derived from the Computer Expression Recognition Toolbox (CERT) for classifying emotions in the Facial Expression Recognition and Analysis (FERA) Challenge. The CERT system automatically estimates facial action intensity and head position using learned appearance-based models on single frames of video. CERT outputs were used to derive a representation of the intensity and motion in each video, consisting of the extremes of displacement, velocity and acceleration. Using this representation, emotion detectors were trained on the FERA training examples. Experiments on the released portion of the FERA dataset are presented, as well as results on the blind test. No consideration of subject identity was taken into account in the blind test. The F1 scores were well above the baseline criterion for success.**

I. INTRODUCTION

Many approaches to automatic detection of facial muscle action and expression have been developed over the past two decades. One line of research resulted in the Computer Expression Recognition Toolbox (CERT), developed by the Machine Perception Laboratory at UCSD ([2] [3] [5] [10]). A synopsis of the system is presented in [11]. CERT is an appearance-based system that is fully automated and operates in real-time. It automatically detects 30 facial actions from the facial action coding system [8], including three dimensions of head pose (yaw, pitch, and roll) and blink, as well as 6 expressions of basic emotion.

CERT currently performs single frame analysis, and produces a real valued output for each Action Unit channel for each frame of video. Because this output is correlated with AU intensity [3], the frame-by-frame output provides information about expression dynamics. CERT is available for academic use and will be presented at FG2011.

In this paper, we test CERT on the emotion recognition subtest of the Facial Expression Recognition Analysis Challenge (FERA) [13]. Performance on the Facial Action recognition subtest is presented in a companion paper [14]. Specifically, we derived measures of facial expression dynamics from the frame-by-frame output of the Computer Expression Recognition Toolbox (CERT). These measures were used to train a system for emotion classification using multinomial logistic regression (MLR). Training was performed on the released portion of the FERA data, and then tested on the blind test of the FERA Challenge.

Thanks to the use of machine learning methods, the field of automated facial expression recognition is rapidly advancing. Technologies like smile detection have already become common place in electronic appliances such as digital cameras [15]. Yet generalizing expression recognition to new expression classes, and/or to new database contexts remains unsolved. One approach to recognizing arbitrary facial expressions focuses on automating the Facial Action Coding System (FACS) [8]. FACS is a system to taxonomize facial expressions as a combination of 57 elementary components including head pose and eye movements. These elementary expressions, known as Action Units (AUs), roughly correspond to the contraction of an individual facial muscle. They can be understood as the phonemes of facial expressions.

Machine learning approaches train expression detectors from image datasets. These datasets need to capture critical sources of variability, such as lightening conditions, image capture instruments, ethnicity, gender, age, and use of facial artifacts such as glasses. An additional challenge is the manner in which expressions are elicited: for example, the timing and morphology of facial expressions changes dramatically when they are produced spontaneously rather than posed. It can be a challenge to elicit examples of the emotional states

in question, for the very large numbers of subjects needed to train robust machine learning systems.

Here we explore performance of a FACS-based approach to facial expression measurement on a novel database, with novel expression characteristics due to the method for expression elicitation.

An open question in computer vision research for facial expression recognition is whether it is better to first detect facial actions, or to directly train emotion detectors from the low-level image features. An advantage of FACS based approaches is that they can take advantage of large datasets used to train AU detection, and apply these to datasets where there may be only tens of subject samples, compared to the thousands needed for robust machine learning. In contrast, direct training approaches may learn image features that are lost or attenuated in the transform to the FACS code.

II. Approach

A. The FERA challenge

To address a need for standardized evaluation procedures, as well as a need for blind testing, the Social Signal Processing Network (SSPNET) hosted the Facial Expression Recognition and Analysis (FERA) challenge. [13] The challenge consists of recognizing five expressions of emotion in three previously seen subjects and three previously unseen subjects.

The FERA challenge was conducted on the GEMEP-FERA dataset. This dataset consists of recordings of 10 actors displaying a range of expressions, while uttering a meaningless phrase, or the word 'Aaah'. Videos were approximately 2-5 seconds in length. There were 5 emotion categories: Anger, Fear, Joy, Relief and Sadness. Emotions were labeled per video and the task was to provide one prediction per video for the test data. The performance measure for the challenge was the F1 statistic.

The training partition consisted of 155 videos of 7 subjects. The FERA challenge guidelines required that the training of the emotion detectors was performed on the GEMEP-FERA dataset only, and did not include examples from other datasets. The blind test data for emotion detection consisted of 134 videos of 6 subjects. Three of the subjects in the test data also appeared in the training data, while the three others did not, providing a subject-dependent and a subject-independent test. Performance numbers for the blind test set were provided by the administrators of the FERA challenge. Challenge participants did not have access to the labels of the blind test set.

Although it would have been possible to do so, the authors of this paper did not attempt to label the blind set themselves, as that can lead to inadvertent over-fitting even if the labels were only used to assess performance.

B. The Computer Expression Recognition Toolbox

The Computer Expression Recognition Toolbox (CERT) is a fully automated system that analyzes facial expressions from video in real-time. Detection of approximately frontal faces [9] is followed detection of six internal facial features which are used for 2D alignment [6] [11]. The aligned face image is then passed through a bank of 72 Gabor filters. Normalized filter output magnitudes are passed to facial action classifiers, which were trained using linear support vector machines on a training set of over 10000 images including 5000 examples of spontaneous expressions [3].These images were labeled by certified human experts in the Facial Action Coding System.

The output of each facial action detector consists of a value indicating the distance to the hyper-plane that separates the two classes (the margin) for each frame of video. System outputs are significantly correlated with the intensity of the facial action, as measured by FACS expert intensity codes [3]. In addition to the facial actions, other channels include the head pose, (3 angular displacement measures of yaw, pitch and roll), a smile detector that was trained on over 20,000 images from the web [15], detectors for combinations of brow movements, detectors for basic emotions [11], and positions of facial landmarks (in pixels) in the original image,

The results presented in this paper employ CERT version 4.4. The following set of CERT output measures were used for this particular emotion detection challenge: 20 AU detectors (1 2 4 5 6 7 9 10 12 14 15 17 18 20 23 24 25 26 28 45), the three head pose measures (yaw, pitch, and roll), the smile

Support for this work was provided by NSF grants SBE-0542013, IIS-0905622, CNS-0454233 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.

detector and inter-ocular distance. The 20 actions were chosen based on there being sufficient numbers of training examples available. The dynamics of inter-ocular distance capture head movements towards and away from the camera. The following were NOT used: unusual actions such as AU11 AU13 AU19 AU27 AU32, basic emotion detectors, unilateral actions and brow action combinations.

Figure 1. Sample facial expressions from the FERA database. Top row left to right shows anger, fear, joy. Bottom row shows relief and sadness.

C. *Measurements of expression dynamics*

From the frame-by-frame CERT output, a set of measures of expression dynamics was extracted from each video. We refer to this set of measures as the EDVA representation (extremes of displacement velocity and acceleration). This representation was defined as follows. Displacement is simply the intensity level and velocity is the slope thereof. Consider the vector of CERT outputs for one action, for all frames of a video. There were 26 such vectors for each video example. For each vector, sliding windows were used to measure the mean intensity, best-fit slope, and rate-of-change of slope within each window. Four different window sizes were used, consisting of W=4, 8, 16 frames, as well as W=full video length. The final EVDA representation consisted of the following for each CERT output vector: The mean over the whole video, the slope over the whole video, the maximum of mean, slope, and acceleration for w=4, 8, and 16 frames, plus the minimum acceleration for w=4, 8 and 16. The acceleration minima add measures of the sharpness of the peaks. Note that other ways of measuring slope and acceleration, such as using different amounts of smoothing before taking standard differences, yielded very similar results.

D. *Emotion Classifiers*

The EDVA measures were used to represent the FERA training videos. Multinomial Logistic Regression was used to train 5 emotion classifiers, that is a set of 26 weights plus one threshold per emotion class. Per the competition guidelines, the training data for the classifiers consisted of the training partition of the GEMEP-FERA database only. These classifiers were then applied to the FERA blind test data.

III. RESULTS

A. *Cross-validation results*

We first present an analysis using leave-one-subject-out cross-validation on the training partition. Cross-validation estimates person-independent test performance. The percent correct score for 5-alternative forced choice on 155 cases was 63% correct.

TABLE I. TABLE 1 MEASURES OF PERFORMANCE ON LEAVE-ONE-SUBJECT OUT VALIDATION ON THE TRAINING PARTITION.

Emotion	F1 gen	F1 opt	F1 bin	ROC
anger	.50	.66	.48	.77
fear	.62	.78	.61	.87
joy	.72	.86	.70	.91
relief	.67	.90	.72	.86
sad	.65	.80	.69	.90
average	.61	.79	.64	.86

The performance measure for the FERA challenge is the F1 statistic. Thresholds were assumed to be zero unless otherwise specified. Three F1 measures and ROC scores are shown in Table I. For column 1, the thresholds were determined by optimizing F1 for the subjects in the training set and applying this to the validation subject (true generalization). For column 2, the same threshold was used for all subjects, so F1 was optimized using all subjects, despite the classifier being trained on only different subjects. For column 3, the outputs were converted to binary with only the maximum in each row being 1, before calculating F1. This is the measure used by the judging team. This method requires no threshold since it's a choice across the classes not across examples. The area under the ROC measure is also threshold independent since it

is calculated using 2-alternative forced choice. Note ROC varies from .5 to 1 for random to perfect, whereas F1 varies from 0 to 1, so ROC scores are typically higher.

We next examined the weights learned by the MLR. This revealed that the weights for the head pose and head motion information were the largest. We note that head pose measurements are also facial actions as defined in the Facial Action Coding System. The MLR weights also revealed that mean intensity was weighted more than slope, and that acceleration was weighted least.

To measure how much information was carried in the head motion measures alone, we ran an MLR on only head information and compared it to the internal facial action channels and smile detector. See Table II. Performance based on the head measures alone was above chance, and it is evident that there is important information in head dynamics for the expression recognition task in the FERA database. This is particularly true for sadness, which can be detected equally well using only head signals as with only internal facial expression. Adding smiles made a considerable improvement to anger, joy and fear detection, while adding internal facial actions did not improve performance for joy and anger. Fear and relief are the two classes which benefited from internal facial actions. This applies specifically to these data and this particular classifier, but in general animation researchers have proposed that joy may be expressed largely in the smile and sadness largely in posture, particularly during speech [4].

TABLE II. PERFORMANCE OF THREE COMBINATIONS OF CERT CHANNELS, MEASURED FOUR WAYS. PERFORMANCE IS LEAVE-ONE-OUT CROSS VALIDATION ON THE TRAINING PARITION, AVERAGED OVER EMOTIONS. BINARY F1 FOR EACH EMOTION, FOR FOUR DIFFERENT COMBINATIONS OF CERT CHANNELS. LEAVE-ONE-SUBJECT OUT VALIDATION ON THE TRAINING PARTITION.

F1 bin	Ang	Fear	Joy	Rel	Sad	Ave
Head	.33	.25	.32	.61	.69	.44
FACS+smile	.39	.55	.70	.42	.68	.55
Head+smile	.53	.47	.76	.63	.70	.62
Head+smile+ FACS	.48	.61	.70	.72	.69	.64

B. Blind Test results

Emotion classification performance was then measured for the blind test of the FERA challenge. Table III shows the blind test results. The confusion matrix for the test set, table IV, suggests that the threshold for fear is low when compared with anger and joy, whereas the confusion between sadness and relief goes in both directions. Relief is positive and sadness is negative, but the gestures for both may be a relaxation, a drooping and a slowing down. Since these factors play a large role in the classifier, it is not surprising that such confusions occur. The full set of confusion matrices for subject-independent and subject-dependent performance could not fit in this paper..

Comparison data for the baseline FERA competition system is provided in Table V. Comparing the system presented in this paper to the LBP baseline system (Table III with Table V) shows that the EVDA representation of CERT performs worse than LBP on anger and better than LBP on the remaining emotions, particularly fear, leading to an average F1 score of .76 compared with LBP's .56.

TABLE III. CLASSIFICATION RATES (F1 BINARY) – BLIND TEST

Emotion	Person independent	Person specific	Overall
anger	0.786	0.769	0.778
fear	0.867	0.900	0.880
Joy	0.750	0.818	0.774
relief	0.500	0.900	0.654
sadness	0.667	0.800	0.720
average	0.714	0.837	0.761

TABLE IV. CONFUSION MATRIX –OVERALL BLIND TEST

Emotion	Ang	Fear	Joy	Relf	Sad
Anger	21	1	1	0	1
Fear	5	22	3	0	0
Joy	1	0	24	4	0
Relief	0	0	2	17	6
Sadness	0	2	1	5	18

TABLE V. LBP BASELINE COMPARISON DATA PROVIDED BY THE FERA CHALLENGE

Emotion	Person independent	Person specific	Overall
anger	.86	.92	.89
fear	.07	.40	.20

Joy	.70	.73	.71
relief	.31	.70	.46
sadness	.27	.90	.52
average	.44	.73	.56

IV. Comparison Data on Emotion Recognition with CERT

A. *CERT basic emotion detectors on FERA data*

The CERT toolbox includes a set of 7 basic emotion detectors, plus neutral expression, which were implemented 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, from the Cohn-Kanade dataset (CK+) [12] and Pictures of Facial Affect (POFA) [7], and their corresponding ground-truth emotion labels.

Although it was not part of the competition design, we also looked at performance of CERT's basic emotion detectors on the FERA data. The FERA competition indicated that emotion detectors are to be trained on the GEMEP-FERA data only, whereas CERT's emotion detectors were trained on CK+ and POFA. We nevertheless thought it would be instructive to look at performance of CERT's frame-by-frame emotion detectors directly on the FERA training data. The 4 emotions common to both datasets were tested: anger, fear, joy and sadness. Using the maximum emotion channel response per video as output, the performance was F1=0.43. With additional MLR training using the CERT emotion outputs from FERA, along with the EDVA representation of these, the on leave-one-out validation performance was F1=0.59. This was considerably higher and comparable to the mean of these 4 emotions for FACS+smile trained on the FERA data (F1=.58) with EDVA and MLR. In other words, the CERT emotion detectors trained on CK+ were well above chance when applied directly to this data without further training. The CERT emotion detectors don't take head pose information as input. When the head motion signals are included with the CERT emotion detectors, classification on FERA improves by 5% and is equivalent to performance of Head+smile+FACS in Table II.

B. *CERT basic emotion detectors on a benchmark dataset*

Performance of CERT emotion detectors on a benchmark dataset, CK+ [12], is shown here for comparison. For this assessment, the emotion detectors were re-trained using CK+ only, without POFA. The first frame (neutral) and last frame (apex) of each video were used for training and testing. CERT 4.4 FACS outputs were used to train nonlinear support vector machines with Gaussian kernel, one for each of 8 emotion classes (7 emotions + neutral). The same set of FACS channels as listed in Section II was used, but excluding smiles, blinks, head pose and inter-ocular distance, for a total of 19 channels. (In this dataset the subjects' pose remained frontal to the camera.) Subject-independent performance was estimated using leave-subject-out cross validation. Because the CERT 4.4 FACS detectors themselves were trained on an older release of the CK database from 2000, cross-validation testing was performed only for the 26 subjects in CK+ that were not in the 2000 release, such that they were novel subjects both for the FACS recognition step as well as for the emotion recognition step. Results are shown Table VI, line 1.

TABLE VI. Performance for leave-subject-out cross-validation on CK+ . Comparing Direct Training on Gabors to Emotion detectors on top of CERT.. 8AFC: Percent correct for an 8-alternative forced choice.

Train	Test	Direct/Indirect	ROC	F1	8AFC
CK+	CK+	Cert SVM	.99	.86	95%
CK+	CK+	Gabor SVM	.98	.79	92%

C. *Facial Actions or direct training?*

A longstanding question in computer vision research for facial expression analysis is whether expression recognition performance is improved by using facial actions, or if it is more advantageous to directly train emotion detectors from the low-level image features such as Gabor features.

Using the CK+ database, we performed a comparison of direct training versus training on top of facial action detectors. To perform this comparison, the nonlinear SVM's were re-trained to detect each of the 8 emotions in CK+, this time taking a bank of Gabor filters as input instead of the CERT facial action measurements. The Gabor filter parameters were identical to the ones used in CERT 4.4. The results are shown in line 2 of Table VI. There was a modest advantage for first detecting

FACS and then training emotion detectors on top of FACS, compared to directly training on the image features. There was also an advantage of complexity, as the classifiers were trained on 19 inputs compared to 72x96x96 Gabor filters.

This comparison indicates that there is sufficient information in the 19 CERT actions to classify the basic emotions. We speculate that the slight advantage of training on CERT features in some cases may be due to training set size. CERT was trained on over 8000 independent labeled images, whereas in many specific applications, there may be only tens or hundreds of images labeled for the emotion or cognitive state in question. Passing the images through CERT takes advantage of the large set of training data used to develop the facial action detectors.

DISCUSSION

This paper tested facial behavior measurements from the Computer Expression Recognition Toolbox (CERT) [11] on the emotion recognition subtest of the Facial Expression Recognition Analysis Challenge (FERA) [13]. Specifically, we derived measures of facial expression dynamics from the frame-by-frame output of CERT. These measures were used to train a system for emotion classification using multinomial logistic regression (MLR). Training was performed on the released portion of the FERA data, and then tested on the blind test of the FERA Challenge.

The FERA emotion dataset represents an unusually difficult classification task. Naive human observers were unable to agree on the intended emotion in many of these videos. The subjects had out-of-plane head rotations and there were only 7 training subjects. Despite these challenges, the system described here achieved some success at classifying expressions of emotion. Performance on the blind test outperformed the LBP baseline. The approach presented here did not take into account the identity of the subjects. Performance could be further improved by taking the identity of subjects into account for the subject-dependent test data.

The analysis presented here showed that there was substantial information about emotion from head pose and head motion in these videos. This is consistent with findings from the animation field [4].

A FACS–based approach, in which emotion detectors were trained on top of facial action signals, was compared to direct training, in which emotion detectors were trained from the low-level image features. The results provide some support for the notion that a FACS-based system trained with machine learning on a large database can provide an advantage when applied to a novel context with significantly fewer training samples.

REFERENCES

[1] T. Ba nziger and K. R. Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. In K. R. Scherer, T. Ba nziger, and E. B. Roesch, editors, *Blueprint for Affective Computing: A Sourcebook*, Series in affective science, chapter 6.1, pages 271–294. Oxford University Press, Oxford, 2010.

[2] Bartlett, M.S., 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.

[3] 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.

[4] Cassell, J. and K.R. Thórisson, The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. International Journal of Applied Artificial Intelligence, 1999. 13(4-5): p. 519-538.

[5] Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P. & Sejnowski, T.J. (1999). "Classifying facial actions." IEEE Trans. Pattern Analysis and Machine Intelligence 21(10): 974-989.

[6] Eckhardt, M., Fasel, I.R. and Movellan, J.R. (2009). Towards Practical Facial Feature Detection. International Journal of Pattern Recognition and Artificial Intelligence 23(3) pp. 379- 400.

[7] P. Ekman and W. Friesen. Pictures of facial affect. Photographs,1976. Available from Human Interaction Laboratory,UCSF, HIL-0984, San Francisco, CA 94143.

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

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

[10] Littlewort G.C., Bartlett M.S., Fasel I, Susskind J, Movellan "An automatic system for measuring facial expression in video." Computer Vision and Image Understanding: Special Issue on Face Processing in Video 24(6) p615-625,2006.

[11] Littlewort G, Whitehill J, Wu T, Fasel I, Frank M, Movellan J, and Bartlett M (2011) The Computer Expression Recognition Toolbox (CERT). Proc. IEEE International Conference on Automatic Face and Gesture Recognition.

[12] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kande Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression. Paper presented at the Third IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB).

[13] Michel F. Valstar, Bihan Jiang, Marc Méhu, Maja Pantic, and Klaus Scherer, "The First Facial Expression Recognition and Analysis Challenge", in Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, 2011, in print.

[14] Wu T, Ruvolo P, Butko N, Whitehill J, Bartlett M, Movellan J. (2011). Action Unit Recognition Transfer Across Datasets. Proc. IEEE Conference on Automatic Face and Gesture Recognition, Workshop on the Facial Expression Recognition and Analysis Challenge (FERA).

[15] Whitehill, J., Littlewort,G.C., Fasel, I.R., Bartlett, M.S., Movellan, J (2009). Towards Practical Smile Detection. Transactions on Pattern Analysis and Machine Intelligence