Behav Res (2015) 47:1136–1147
DOI 10.3758/s13428-014-0536-1

Spontaneous facial expression in unscripted social interactions can be measured automatically

Jeffrey M. Girard · Jeffrey F. Cohn · Laszlo A. Jeni · Michael A. Sayette · Fernando De la Torre

Published online: 9 December 2014

Abstract Methods to assess individual facial actions have potential to shed light on important behavioral phenomena ranging from emotion and social interaction to psychological disorders and health. However, manual coding of such actions is labor intensive and requires extensive training. To date, establishing reliable automated coding of unscripted facial actions has been a daunting challenge impeding development of psychological theories and applications requiring facial expression assessment. It is therefore essential that automated coding systems be developed with enough precision and robustness to ease the burden of manual coding in challenging data involving variation in participant gender, ethnicity, head pose, speech, and occlusion. We report a major advance in automated coding of spontaneous facial actions during an unscripted social interaction involving three strangers. For each participant ($n = 80$, 47 % women, 15 % Nonwhite), 25 facial action units (AUs) were manually coded from video using the Facial Action Coding System. Twelve AUs occurred more than 3 % of the time and were processed using automated FACS coding. Automated coding showed very strong reliability for the proportion of time that each AU occurred (mean intraclass correlation = 0.89), and the more stringent criterion of frame-by-frame reliability was moderate to strong (mean Matthew's correlation = 0.61). With few exceptions, differences in AU detection related to gender, ethnicity, pose, and average pixel intensity were small. Fewer than 6 % of frames could be coded manually but not automatically. These findings suggest automated FACS coding has progressed sufficiently to be applied to observational research in emotion and related areas of study.

Keywords Facial expression · FACS · Affective computing · Automated coding

Introduction

During the past few decades, some of the most striking findings about affective disorders, schizophrenia, addiction, developmental psychopathology, and health have been based on sophisticated coding of facial expressions. For instance, it has been found that facial expression coding using the Facial Action Coding System (FACS), which is the most comprehensive system for coding facial behavior (Ekman, Friesen, & Hager, 2002), identifies which depressed patients are at greatest risk for reattempting suicide (Archinard, Haynal-Reymond, & Heller, 2000); constitutes an index of physical pain with desirable psychometric properties (Prkachin & Solomon, 2008); distinguishes different types of adolescent behavior problems (Keltner, Moffitt, & Stouthamer-Loeber 1995); and distinguishes between European-American, Japanese, and Chinese infants (Camras et al., 1998). These findings have offered glimpses into critical areas of human behavior that were not possible using existing methods of assessment, often generating considerable research excitement and media attention.

Electronic supplementary material The online version of this article (doi:10.3758/s13428-014-0536-1) contains supplementary material, which is available to authorized users.

J. M. Girard (✉) · J. F. Cohn · M. A. Sayette
Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260 USA
e-mail: jmg174@pitt.edu

J. F. Cohn · L. A. Jeni · F. De la Torre
The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA

As striking as these original findings were, it is just as striking how little follow-up work has occurred using these methods. The two primary reasons for this curious state of affairs are the intensive training required to learn facial expression coding and the extremely time-consuming nature of the coding itself. Paul Ekman, one of the creators of FACS, notes that certification in FACS requires about 6 months of training and that FACS coding a single minute of video can take over an hour (Ekman, 1982).

FACS (Ekman & Friesen, 1978; Ekman et al., 2002) is an anatomically based system for measuring nearly all visually discernible facial movement. FACS describes facial activities in terms of unique facial action units (AUs), which correspond to the contraction of one or more facial muscles. Any facial expression may be represented as a single AU or a combination of multiple AUs. For example, the Duchenne smile (i.e., enjoyment smile) is indicated by simultaneous contraction of the zygomatic major (AU 12) and orbicularis oculi pars lateralis (AU 6). Although there are alternative systems for characterizing facial expression (e.g., Izard, 1979; Abrantes & Pereira, 1999), FACS is recognized as the most comprehensive and objective means for measuring facial movement currently available, and it has become the standard for facial measurement in behavioral research (Cohn & Ekman, 2005; Ekman & Rosenberg, 2005).

Given the often-prohibitive time commitment of FACS coding, there has been great interest in developing computer vision methods for automating facial expression coding. If successful, these methods would greatly improve the efficiency and reliability of AU detection, and importantly make its use feasible in applied settings outside of research.

Although the advantages of automated facial expression coding are apparent, the challenges of developing such systems are considerable. While human observers easily accommodate variations in pose, scale, illumination, occlusion, and individual differences (e.g., gender and ethnicity), these and other sources of variation represent considerable challenges for a computer vision system. Further, there is the machine learning challenge of automatically detecting actions that require significant training and expertise even for human coders.

There has been significant effort to develop computer-vision-based approaches to automated facial expression analysis. Most of this work has focused on prototypic emotion expressions (e.g., joy and anger) in posed behavior. Zeng, Pantic, Roisman, and Huang (2009) have reviewed this literature through 2009. Within the past few years, studies have progressed to AU detection in actor portrayals of emotion (Valstar, Bihan, Mehu, Pantic, & Scherer 2011) and the more challenging task of AU detection during spontaneous facial behavior. Examples of the latter include AU detection in physical pain (Littlewort, Bartlett, & Lee, 2009; Lucey, Cohn, Howlett, Member, & Sridharan, 2011), interviews (Bartlett et al., 2006; Girard et al., 2013; Lucey, Matthews, Ambadar, De la Torre, & Cohn, 2006), and computer-mediated tasks such as watching a video clip or filling out a form (Hoque, McDuff, & Picard, 2012; Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2013; Littlewort et al., 2011; Mavadati, Mahoor, Bartlett, Trinh, & Cohn, 2013; McDuff, El Kaliouby, Kodra, & Picard, 2013).

While much progress has been made, the current state of the science is limited in several key respects. Stimuli to elicit spontaneous facial actions have been highly controlled (e.g., watching pre-selected video clips or replying to structured interviews) and camera orientation has been frontal with little or no variation in head pose. Non-frontal pose matters because the face looks different when viewed from different orientations and parts of the face may become self-occluded. Rapid head movement also may be difficult to automatically track through a video sequence. Head motion and orientation to the camera are important if AU detection is to be accomplished in social settings where facial expressions often co-occur with head motion. For example, the face and head pitch forward and laterally during social embarrassment (Keltner et al., 1995; Ambadar, Cohn, & Reed, 2009). Kraut and Johnston (1979) found that successful bowlers smile only as they turn away from the bowling lane and toward their friends.

Whether automated methods can detect spontaneous facial expressions in the presence of head pose variation is unknown, as too few studies have encountered or reported on it. Messinger, Mahoor, Chow, and Cohn (2009) encountered out-of-plane head motion in video of infants, but neglected to report whether it affected AU detection. Cohn and Sayette (2010) reported preliminary evidence that AU detection may be robust to pose variation up to 15 degrees from frontal. Similarly, we know little about the effects of gender and ethnicity on AU detection. Face shape and texture vary between men and women (Bruce & Young 1998), and may be further altered through the use of cosmetics. Skin color is an additional factor that may affect AU detection. Accordingly, little is known about the operational parameters of automated AU detection. For these reasons, automated FACS coding must prove robust to these challenges.

The current study evaluates automated FACS coding using a database that is well suited to testing just how far automated methods have progressed, and how close we are to using them to study naturally occurring facial expressions. This investigation focuses on spontaneous facial expression in a far larger database (over 400,000 video frames from 80 people) than ever attempted; it includes men and women, Whites and Nonwhites, and a wide range of facial AUs that vary in intensity and head orientation. Because this database contains variation in head pose and participant gender, as well as moderate variation in

illumination and participant ethnicity, we can examine their effect on AU detection. To demonstrate automated AU detection in such a challenging database would mark a crucial step toward the goal of establishing fully automated systems capable of use in varied research and applied settings.

Methods

Participants

The current study used digital video from 80 participants (53 % male, 85 % white, average age 22.2 years) who were participating in a larger study on the impact of alcohol on group formation processes (for elaboration, see Sayette et al., 2012). They were randomly assigned to groups of three unacquainted participants. Whenever possible, all three participants in a group were analyzed. Some participants were not analyzable due to excessive occlusion from hair or head wear ($n = 6$) or gum chewing ($n = 1$). Participants were randomly assigned to drink isovolumic alcoholic beverages ($n = 31$), placebo beverages ($n = 21$), or nonalcoholic control beverages ($n = 28$); all participants in a group drank the same type of beverage. The majority of participants were from groups with a mixed gender composition of two males and one female ($n = 32$) or two females and one male ($n = 26$), although some were from all male ($n = 12$) or all female ($n = 10$) groups. All participants reported that they had not consumed alcohol or psychoactive drugs (except nicotine or caffeine) during the 24 hour period leading up to the observations.

Setting and equipment

All participants were previously unacquainted. They first met only after entering the observation room where they were seated approximately equidistantly from each other around a circular (75 cm diameter) table. They were asked to consume a beverage consisting of cranberry juice or cranberry juice and vodka (a 0.82 g/kg dose of alcohol for males and a 0.74 g/kg dose of alcohol for females) before engaging in a variety of cognitive tasks. We focus on a portion of the 36-min unstructured observation period in which participants became acquainted with each other (mean duration 2.69 min). Separate wall-mounted cameras faced each person. It was initially explained that the cameras were focused on their drinks and would be used to monitor their consumption rate from the adjoining room, although participants later were told of our interest in observing their behavior and a second consent form was signed if participants were willing. All participants consented to this use of their data.

The laboratory included a custom-designed video control system that permitted synchronized video output for each participant, as well as an overhead shot of the group (Fig. 1). The individual view for each participant was used in this report. The video data collected by each camera had a standard frame rate of 29.97 frames per second and a resolution of 640×480 pixels. Audio was recorded from a single microphone. The automated FACS coding system was processed on a Dell T5600 workstation with 128GB of RAM and dual Xeon E5 processors. The system also runs on standard desktop computers.

Manual FACS coding

The FACS manual (Ekman et al., 2002) defines 32 distinct facial action units. All but 7 were manually coded. Omitted were three "optional" AUs related to eye closure (AUs 43, 45, and 46), three AUs related to mouth opening or closure (AUs 8, 25, and 26), and one AU that occurs on the neck rather than the face (AU 21). The remaining 25 AUs were manually coded from onset (start) to offset (stop) by one of two certified and highly experienced FACS coders using Observer XT software (Noldus Information Technology, 2013). AU onsets were annotated when they reached slight or B level intensity according to FACS; the corresponding offsets were annotated when they fell below B level intensity. AU of lower intensity (i.e., A level intensity) are ambiguous and difficult to detect for both manual and automated coders. The original FACS manual (Ekman & Friesen, 1978) did not code A level intensity (referred to there as "trace."). All AUs were annotated during speech.

Because highly skewed class distributions severely attenuate measures of classifier performance (Jeni, Cohn, & De la Torre, 2013), AUs that occurred less than about 3 % of the time were excluded from analysis. Thirteen AUs were omitted on this account. Five of them either never occurred or occurred less than 1 % of the time. Manual coding of

Fig. 1 Examples of video frames with facial landmark tracking

these five AUs was suspended after the first 56 subjects. Visual inspection of Fig. 2 reveals that there was a large gap between the AUs that occurred approximately 10 % or more of the time and those that occurred approximately 3 % or less of the time. The class distributions of the excluded AUs were at least three times more skewed than those of the included AUs. In all, 12 AUs met base-rate criteria and were included for automatic FACS coding.

To assess inter-observer reliability, video from 17 participants was annotated by both coders. Mean frame-level reliability was quantified with the Matthews Correlation Coefficient (MCC), which is robust to agreement due to chance as described below. The average MCC was 0.80, ranging from 0.69 for AU 24 to 0.88 for AU 12; according to convention, these numbers can be considered strong to very strong reliability (Chung, 2007). This high degree of inter-observer reliability is likely due to extensive training and supervision of the coders.

Automatic FACS coding

Figure 3 shows an overview of the AU detection pipeline. The face is detected automatically and facial landmarks are detected and tracked. The face images and landmarks are normalized to control for variation in size and orientation, and appearance features are extracted. The features then are input to classification algorithms, as described below. Please note that the mentioned procedures do not provide incremental results; all the procedures are required to perform classification and calculate an inter-system reliability score.

Landmark registration

The first step in automatically detecting AUs was to locate the face and facial landmarks. Landmarks refer to points that define the shape of permanent facial features, such as the eyes and lips. This step was accomplished using the LiveDriver SDK (Image Metrics, 2013), which is a generic tracker that requires no individualized training to track facial landmarks of persons it has never seen before. It locates the two-dimensional coordinates of 64 facial landmarks in each image. These landmarks correspond to important facial points such as the eye and mouth corners, the tip of the nose, and the eyebrows (Fig. 1). LiveDriver SDK also tracks head pose in three dimensions for each video frame: pitch (i.e., vertical motion such as nodding), yaw (i.e., horizontal motion such as shaking the head), and roll (i.e., lateral motion such as tipping the head sideways).

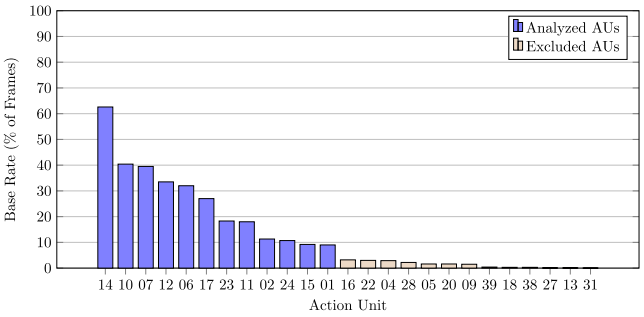

Fig. 2 Base rates of all the coded facial action units from a subset of the data ($n = 56$)

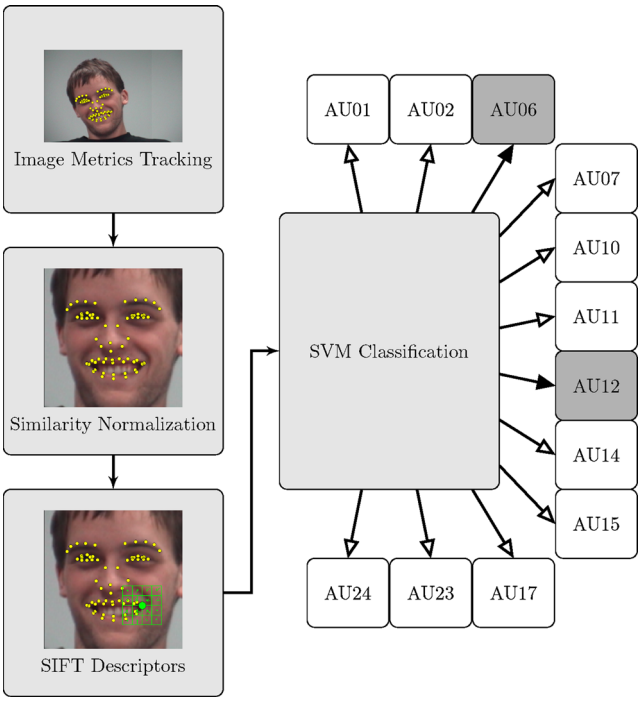

Fig. 3 Automated FACS Coding Pipeline. Example shown is for AU 6+12

Shape and texture information can only be used to identify facial expressions if the confounding influence of head motion is controlled (De la Torre & Cohn, 2011). Because participants exhibited a great deal of rigid head motion during the group formation task, the second step was to remove the influence of such motion on each image. Many techniques for alignment and registration are possible (Zeng et al., 2009); we chose the widely used similarity transformation (Szeliski, 2011) to warp the facial images to the average pose and a size of 128×128 pixels, thereby creating a common space in which to compare them. In this way, variation in head size and orientation would not confound the measurement of facial actions.

Feature extraction

Once the facial landmarks had been located and normalized, the third step was to measure the deformation of the face

caused by expression. This was accomplished by extracting scale-invariant feature transform (SIFT) descriptors (Lowe, 1999) in localized regions surrounding each facial landmark. SIFT applies a geometric descriptor to an image region and measures features that correspond to changes in facial texture and orientation (e.g., facial wrinkles, folds, and bulges). It is robust to changes in illumination and shares properties with neurons responsible for object recognition in primate vision (Serre et al., 2005). SIFT feature extraction was implemented using the VLFeat open-source library (Vedali & Fulkerson, 2008). The diameter of the SIFT descriptor was set to 24 pixels, as illustrated above the left lip corner in Fig. 3.

Classifier training

The final step in automatically detecting AUs was to train a classifier to detect each AU using SIFT features. By providing each classifier multiple examples of an AU's presence and absence, it was able to learn a mapping of SIFT features to that AU. The classifier then extrapolated from the examples to predict whether the AU was present in new images. This process is called supervised learning and was accomplished using support vector machine (SVM) classifiers (Vapnik, 1995). SVM classifiers extrapolate from examples by fitting a hyperplane of maximum margin into the transformed, high dimensional feature space. SVM classification was implemented using the LIBLINEAR open-source library (Fan, Wang, & Lin, 2008).

The performance of a classifier is evaluated by testing the accuracy of its predictions. To ensure generalizability of the classifiers, they must be tested on examples from people they have not seen previously. This is accomplished by cross-validation, which involves multiple rounds of training and testing on separate data. Stratified k-fold cross-validation (Geisser, 1993) was used to partition participants into 10 folds with roughly equal AU base rates. On each round of cross-validation, a classifier was trained using data (i.e., features and labels) from eight of the ten folds. The classifier's cost parameter was optimized using one of the two remaining folds through a "grid-search" procedure (Hsu, Chang, & Lin, 2003). The predictions of the optimized classifier were then tested through extrapolation to the final fold. This process was repeated so that each fold was used once for testing and parameter optimization; classifier performance was averaged over these 10 iterations. In this way, training and testing of the classifiers was independent.

Inter-system reliability

The performance of the automated FACS coding system was measured in two ways. Following the example of Girard et al. (2013), we measured both session-level and frame-level reliability. Session-level reliability asks whether the expert coder and the automated system are consistent in their estimates of the proportion of frames that include a given AU. Frame-level reliability represents the extent to which the expert coder and the automated system make the same judgments on a frame-by-frame basis. That is, for any given frame, do both detect the same AU? For many purposes, such as comparing the proportion of positive and negative expressions in relation to severity of depression, session-level reliability of measurement is what matters. Session-level reliability was assessed using intraclass correlation (ICC) (Shrout & Fleiss, 1979). Frame-level reliability was quantified using the Matthews Correlation Coefficient (MCC) (Powers, 2007).

$$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1)WMS} \tag{1}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2}$$

The intraclass correlation coefficient (ICC) is a measure of how much the units in a group resemble one another (Shrout & Fleiss, 1979). It is similar to the Pearson correlation coefficient, except that for ICC the data are centered and scaled using a pooled mean and standard deviation rather than each variable being centered and scaled using its own mean and standard deviation. This is appropriate when the same measure is being applied to two sources of data (e.g., two manual coders or a manual coder and an automated AU detector), and prevents an undesired handicap from being introduced by invariance to linear transformation. For example, an automated system that always detected a base rate twice as large as that of the human coder would have a perfect Pearson correlation coefficient, but a poor ICC. For this reason, the behavior of ICC is more rigorous than that of the Pearson correlation coefficient when applied to continuous values. We used the one-way, random effects model ICC described in Eq. 1.

The Matthews correlation coefficient (MCC), also known as the phi coefficient, can be used as a measure of the quality of a binary classifier (Powers, 2007). It is equivalent to a Pearson correlation coefficient computed for two binary measures and can be interpreted in the same way: an MCC of 1 indicates perfect correlation between methods, while an MCC of 0 indicates no correlation (or chance agreement). MCC is related to the chi-squared statistic for a 2×2 contingency table, and is the geometric mean of Informedness

(DeltaP) and Markedness (DeltaP'). Using Eq. 2, MCC can be calculated directly from a confusion matrix. Although there is no perfect way to represent a confusion matrix in a single number, MCC is preferable to alternatives (e.g., the F-measure or Kappa) because it makes fewer assumptions about the distributions of the data set and the underlying populations (Powers 2012).

Because ICC and MCC are both correlation coefficients, they can be evaluated using the same heuristic, such as the one proposed by Chung (2007): that coefficients between 0.0 and 0.2 represent very weak reliability, coefficients between 0.2 and 0.4 represent weak reliability, coefficients between 0.4 and 0.6 represent moderate reliability, coefficients between 0.6 and 0.8 represent strong reliability, and coefficients between 0.8 and 1.0 represent very strong reliability.

Error analysis

We considered a variety of factors that could potentially influence automatic AU detection. These were participant gender, ethnicity, mean pixel intensity of the face, seating location, and variation in head pose. Mean pixel intensity is a composite of several factors that include skin color, orientation to overhead lighting, and head pose. Orientation to overhead lighting could differ depending on participants' location at the table. Because faces look different when viewed from different angles, pose for each frame was considered.

The influence of ethnicity, sex, average pixel intensity, seating position, and pose on classification performance was evaluated using hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002). HLM is a powerful statistical tool for modeling data with a "nested" or interdependent structure. In the current study, repeated observations were nested within participants. By creating sub-models (i.e., partitioning the variance and covariance) for each level, HLM accounted for the fact that observations from the same participant are likely to be more similar than observations from different participants.

Classifier predictions for each video frame were assigned a value of 1 if they matched the manual coder's annotation and a value of 0 otherwise. These values were entered into a two-level HLM model as its outcome variable; a logit-link function was used to transform the binomial values into continuous log-odds. Four frame-level predictor variables were added to the first level of the HLM: z-scores of each frame's head pose (yaw, pitch, and roll) and mean pixel intensity. Two participant-level predictor variables were added to the second level of the HLM: dummy codes for participant gender (0 = male, 1 = female) and ethnicity (0 = White, 1 = Nonwhite). A sigmoid function was used to transform log-odds to probabilities for ease of interpretation.

Results

Descriptive statistics

Using manual FACS coding, the mean base rate for AUs was 27.3 % with a relatively wide range. AU 1 and AU 15 were least frequent, with each occurring in only 9.2 % of frames; AU 12 and AU 14 occurred most often, in 34.3 % and 63.9 % of frames, respectively (Table 1). Occlusion, defined as partial obstruction of the view of the face, occurred in 18.8 % of all video frames.

Base rates for two AUs differed between men and women. Women displayed significantly more AU 10 than men, $t(78) = 2.79$, $p < .01$, and significantly more AU 15 than men, $t(78) = 3.05$, $p < .01$. No other significant differences between men and women emerged, and no significant differences in base rates between Whites and Nonwhites emerged.

Table 1 Action unit base rates from manual FACS coding (% of frames)

AU	Overall	Male	Female	White	Other
1	9.2	7.8	9.3	8.5	9.8
2	11.7	10.0	11.7	10.4	14.3
6	33.4	28.5	38.1	33.2	34.8
7	41.5	37.9	43.7	41.1	38.9
10	40.3	33.0	46.6	38.9	46.3
11	16.9	11.7	21.5	18.1	7.3
12	34.1	30.8	36.4	33.0	38.8
14	63.9	59.7	69.6	63.5	73.1
15	9.2	5.8	11.7	8.3	12.9
17	28.3	30.2	23.4	26.8	25.6
23	20.4	20.7	17.8	19.6	16.5
24	18.8	11.2	12.6	11.9	12.1

Note: Shaded cells indicate significant differences between groups ($p < .05$).

Approximately 5.6 % of total frames could be coded manually but not automatically. 9.7 % of total frames could be coded neither automatically nor manually. Occlusion was responsible for manual coding failures. Tracking failure most likely due to occlusion was responsible for automatic coding failures.

Head pose was variable, with most of that variation occurring within the interval of 0 to 20° from frontal view. (Here and following, absolute values are reported for head pose.) Mean pose was 7.6° for pitch, 6.9° for yaw, and 6.1° for roll. The 95th percentiles were 20.1° for pitch, 15.7° for yaw, and 15.7° for roll.

Although illumination was relatively consistent in the observation room, the average pixel intensity of faces did vary. Mean pixel intensity was 40.3 % with a standard deviation of 9.0 %. Three potential sources of variation were considered: ethnicity, seating location, and head pose. Mean pixel intensity was lower for Nonwhites than for Whites, $t(78) = 4.87$, $p < 0.001$. Effects of seating location were also significant, with participants sitting in one of the chairs showing significantly lower mean pixel intensity than participants sitting in the other chairs, $F(79) = 5.71$, $p < .01$. Head pose was uncorrelated with pixel intensity: for yaw, pitch, and roll, $r = -0.09$, -0.07, and -0.04, respectively.

Inter-system reliability

The mean session-level reliability (i.e., ICC) for AUs was very strong at 0.89, ranging from 0.80 for AU 17 to 0.95 for AU 12 and AU 7 (Fig. 4). The mean ICC was 0.91 for male participants and 0.79 for female participants. The mean ICC was 0.86 for participants self-identifying as White and 0.91 for participants self-identifying as Nonwhite.

The mean frame-level reliability (i.e., MCC) for AUs was strong at 0.60, ranging from 0.44 for AU 15 to 0.79 for AU 12 (Fig. 4). The mean MCC was 0.61 for male participants and 0.59 for female participants. The mean MCC was 0.59 for participants self-identifying as White and 0.63 for participants self-identifying as Nonwhite.

Error analysis

HLM found that a number of participant- and frame-level factors affected the likelihood that the automated system would make classification errors for specific AUs (Table 2). For several AUs, participant gender and self-reported ethnicity affected performance. Errors were 3.45 % more likely in female than male participants for AU 6 ($p < .05$), 2.91 % more likely in female than male participants for AU 15 ($p < .01$), and 5.15 % more likely in White than Nonwhite participants for AU 17 ($p < .05$). For many AUs, frame-level head pose and mean pixel intensity affected performance. For every one standard deviation increase in the absolute value of head yaw, the probability of making an error increased by 0.79 % for AU 2 ($p < .05$), by 0.15 % for AU 11 ($p < .05$), by 1.24 % for AU 12 ($p < .01$), by 1.39 % for AU 23 ($p < .05$), and by 0.77 % for AU 24 ($p < .05$). For every one standard deviation increase in the absolute value of head pitch, the probability of making an error increased by 1.24 % for AU 15 ($p < .05$). No significant effects were found for deviations in head roll. Finally, for every one standard deviation increase in mean

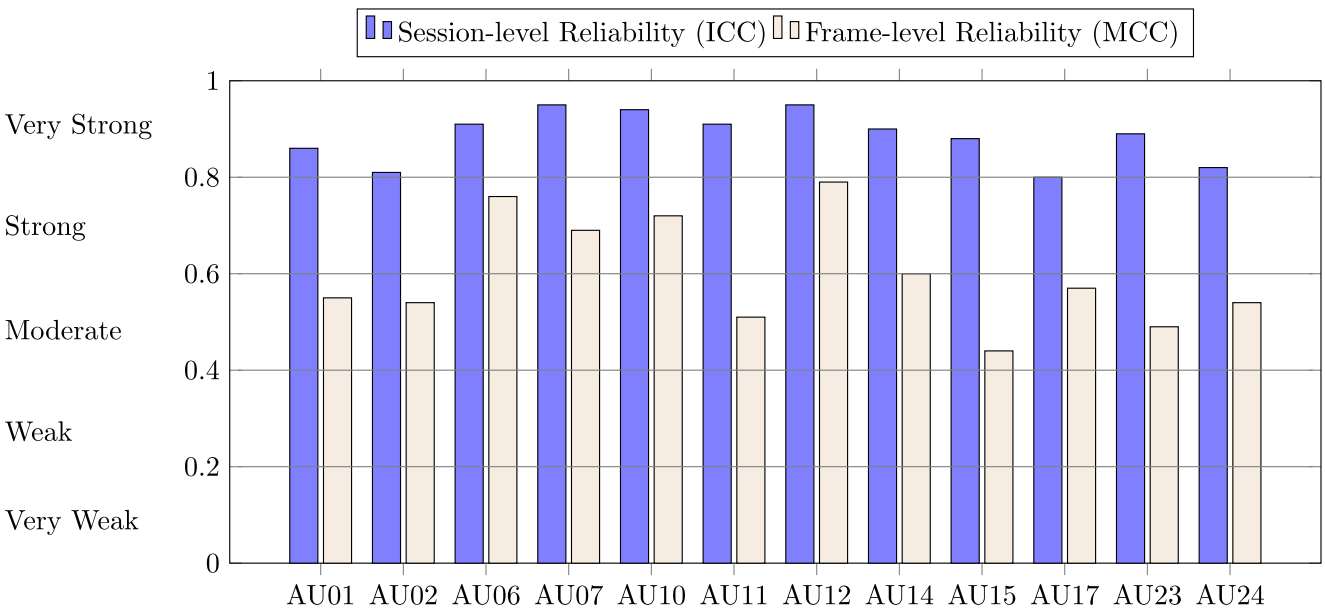

Fig. 4 Mean inter-system reliability for twelve FACS action units

pixel intensity, the probability of making an error increased by 2.21 % for AU 14 ($p < .05$).

Discussion

The major finding of the present study was that spontaneous facial expression during a three person, unscripted social interaction can be reliably coded using automated methods. This represents a significant breakthrough in the field of affective computing and offers exciting new opportunities for both basic and applied psychological research.

We evaluated the readiness of automated FACS coding for research use in two ways. One was to assess session-level reliability: whether manual and automated measurement yield consistent estimates of the proportion of time that different AUs occur. The other, more-demanding metric was frame-level reliability: whether manual and automated measurement agree on a frame-by-frame basis. When average rates of actions are of interest, session-level reliability is the critical measure (e.g., Sayette & Hufford, 1995, Girard et al., 2013). When it is important to know when particular actions occur in the stream of behavior, for instance to define particular combinations of AUs, frame-level reliability is what matters (e.g., Ekman & Heider, 1988; Reed, Sayette, & Cohn, 2007). For AUs that occurred as little as 3 % of the time, we found evidence of very strong session-level reliability and moderate to strong frame-level reliability. AUs occurring less than 3 % of the time were not analyzed.

Session-level reliability (i.e., ICC) averaged 0.89, which can be considered very strong. The individual coefficients were especially strong for AUs associated with positive affect (AU 6 and AU 12), which is of particular interest in studies of group formation (Fairbairn, Sayette, Levine, Cohn, & Creswell, 2013; Sayette et al., 2012) as well as emotion and social interaction more broadly (Ekman & Rosenberg, 2005). Session-level reliability for AUs related to brow actions and smile controls, which counteract the upward pull of the zygomatic major (Ambadar et al., 2009; Keltner, 1995), were only somewhat lower. Smile controls have been related to embarrassment, efforts to down-regulate positive affect, deception, and social distancing (Ekman & Heider, 1988; Girard, Cohn, Mahoor, Mavadati, & Rosenwald, 2013; Keltner & Buswell, 1997; Reed et al., 2007).

The more demanding frame-level reliability (i.e., MCC) averaged 0.60, which can be considered strong. Similar to the session-level reliability results, actions associated with positive affect had the highest frame-level reliability (0.76 for AU 6 and 0.79 for AU 12). MCC for smile controls was more variable. For AU 14 (i.e., dimpler), which is associated with contempt and anxiety (Fairbairn et al., 2013), and AU 10, which is associated with disgust (Ekman, 2003), reliability was strong (MCC = 0.60 and 0.72, respectively). MCC for some others was lower (e.g., 0.44 for AU 15). When frame-by-frame detection is required, reliability is strong for some AUs but only moderate for others. Further research is indicated to improve detection of the more difficult AUs (e.g., AU 11 and AU 15).

Our findings from a demanding group formation task with frequent changes in head pose, speech, and intensity are highly consistent with what has been found previously in more constrained settings. In psychiatric interview, for instance, we found that automated coding was highly consistent with manual coding and revealed the same pattern of state-related changes in depression severity over time (Girard et al., 2013).

Results from error analysis revealed that several participant-level factors influenced the probability of misclassification. Errors were more common for female than male participants for AU 6 and AU 15, which may be

Table 2 Standardized regression coefficients predicting the likelihood of correct automated annotation

AU	Participant Variables		Video Frame Variables			
	Female	Nonwhite	Yaw	Pitch	Roll	Pixel
1	0.01	-0.84	-0.01	0.05	0.09	-0.35
2	-0.18	-0.53	-0.22∗	0.02	-0.03	-0.27
6	-0.53∗	0.19	-0.08	-0.03	0.05	0.05
7	-0.26	0.07	-0.11	-0.00	-0.01	-0.13
10	-0.23	0.39	-0.13	-0.01	0.01	0.02
11	-1.29	0.66	-0.23∗	-0.03	0.11	-0.23
12	-0.29	0.16	-0.23∗∗	0.04	0.00	0.14
14	0.17	-0.08	-0.01	-0.02	0.06	-0.19∗
15	-0.73∗∗	-0.57	0.06	-0.11∗	0.06	-0.18
17	-0.14	0.59∗	-0.03	-0.02	0.04	0.24
23	-0.24	0.15	-0.16∗∗	0.02	0.04	0.16
24	-0.50	0.27	-0.19∗	0.09	-0.00	0.24

Note: Standardized regression coefficients are in log-odds form. $* = p < .05$ and $** = p < .01$

due to gender differences in facial shape, texture, or cosmetics-usage. AU 15 was also more than twice as frequent in female than male participants, which may have led to false negatives for females. With this caveat in mind, the overall findings strongly support use of automated FACS coding in samples with both genders. Regarding participant ethnicity, errors were more common in White than Nonwhite participants for AU 17. This finding may suggest that the facial texture changes caused by AU 17 are easier to detect on darker skin. Replication of this finding, however, would be important as the number of Nonwhite participants was small relative to the number of White participants (i.e., 12 Nonwhite vs. 68 White).

Several frame-level factors also influenced the probability of misclassification. In the group formation task, most head pose variation was within plus or minus 20° of frontal and illumination was relatively consistent. Five AUs showed sensitivity to horizontal change in head pose (i.e., yaw): the probability of errors increased for AU 2, AU 11, AU 12, AU 23, and AU 24 as participants turned left or right and away from frontal. Only one AU showed sensitivity to vertical change in head pose (i.e., pitch): the probability of errors increased for AU 15 as participants turned up or down and away from frontal. No AUs showed sensitivity to rotational change in head pose (i.e., roll). Finally, only one AU showed sensitivity to change in illumination: the probability of errors increased for AU 14 as mean pixel intensity increased. These findings suggest that horizontal motion is more of a concern than vertical or rotational motion. However, the overall reliability results suggest that automated FACS coding is suitable for use in databases with the amount of head motion that can be expected in the context of a spontaneous social interaction. For contexts in which larger pose variation is likely, pose-dependent training may be needed (Guney, Arar, Fischer, & Ekenel, 2013). Although the effects of mean pixel intensity were modest, further research is needed in databases with more variation in illumination.

Using only a few minutes of manual FACS coding each from 80 participants, we were able to train classifiers that repeatedly generalized (during iterative cross-validation) to unseen portions of the data set, including unseen participants. This suggests that the un-coded portions of the data set—over 30 min of video from 720 participants—could be automatically coded via extrapolation with no additional manual coding. Given that it can take over an hour to manually code a single minute of video, this represents a substantial savings of time and opens new frontiers in facial expression research.

A variety of approaches to AU detection using appearance features have been pursued in the literature. One is static modeling; another is temporal modeling. In static modeling, each video frame is evaluated independently. For this reason, it is invariant to head motion. Static modeling is the approach we used. Early work used neural networks for static modeling (Tian, Kanade, & Cohn, 2001). More recently, support vector machine classifiers such as we used have predominated (De la Torre & Cohn, 2011). Boosting, an iterative approach, has been used to a lesser extent for classification as well as for feature selection (Littlewort, Bartlett, Fasel, Susskind, & Movellan, 2006; Zhu, De la Torre, Cohn, & Zhang, 2011). Others have explored rule-based systems (Pantic & Rothkrantz, 2000) for static modeling. In all, static modeling has been the most prominent approach.

In temporal modeling, recent work has focused on incorporating motion features to improve performance. A popular strategy is to use hidden Markov models (HMM) to temporally segment actions by establishing a correspondence between AU onset, peak, and offset and an underlying latent state. Valstar and Pantic (2007) used a combination of SVM and HMM to temporally segment and recognize AUs. In several papers, Qiang and his colleagues (Li, Chen, Zhao, & Ji, 2013; Tong, Chen, & Ji, 2010; Tong, Liao, & Ji, 2007) used what are referred to as dynamic Bayesian networks (DBN) to detect facial action units. DBN exploits the known correlation between AU. For instance, some AUs are mutually exclusive. AU 26 (mouth open) cannot co-occur with AU 24 (lips pressed). Others are mutually "excitatory." AU 6 and AU 12 frequently co-occur during social interaction with friends. These "dependencies" can be used to reduce uncertainty about whether an AU is present. While they risk false positives (e.g., detecting a Duchenne smile when only AU 12 is present), they are a promising approach that may become more common (Valstar & Pantic, 2007).

The current study is, to our knowledge, the first to perform a detailed and statistically controlled error analysis of an automated FACS coding system. Future research would benefit from evaluating additional factors that might influence classification, such as speech and AU intensity. The specific influence of speech could not be evaluated because audio was recorded using a single microphone and it was not feasible to code speech and non-speech separately for each participant. The current study also focused on AU detection and ignored AU intensity.

Action units can vary in intensity across a wide range from subtle, or trace, to very intense. The intensity of facial expressions is linked to both the intensity of emotional experience and social context (Ekman, Friesen, & Ancoli, 1980; Hess, Banse, & Kappas, 1995; Fridlund, 1991), and is essential to the modeling of expression dynamics over time.

In an earlier study using automated tracking of facial landmarks, we found marked differences between posed and spontaneous facial actions. In the former, amplitude and velocity of smile onsets were strongly correlated consistent with ballistic timing (Cohn & Schmidt, 2004). For posed smiles, the two were uncorrelated. In related work, Messinger et al. (2009) found strong covariation in the timing of mother and infant smile intensity. While the present data provide compelling evidence that automated coding systems now can code the occurrence of spontaneous facial actions, future research is necessary to test the ability to automatically code change in AU intensity.

Some investigators have sought to measure AU intensity using a probability or distance estimate from a binary classifier. Recall that for an SVM, each video frame can be located with respect to its distance from a hyper-plane that separates positive and null instances of AU. When the value exceeds a threshold, a binary classifier declares the AU is present. When the value falls short of the threshold, the binary classifier rules otherwise. As a proxy for intensity, Bartlett and others have proposed using either the distance measure or a pseudo-probability based on that distance measure. This method worked well for posed facial actions but not for spontaneous ones (Bartlett et al., 2006; Girard, 2014; Yang, Qingshan, & Metaxas, 2009). To automatically measure intensity of spontaneous facial actions, we found that it is necessary to train classifiers on manually coded AU intensity (Girard, 2014). In two separate data sets, we found that classifiers trained in this way consistently out-performed those that relied on distance measures. Behavioral researchers are cautioned to be wary of approaches that use distance measures in such a way.

Because classifier models may be sensitive to differences in appearance, behavior, context, and recording environment (e.g., cameras and lighting), generalizability of AU detection systems from one data set to another cannot be assumed. A promising approach is to personalize classifiers by exploiting similarities between test and training subjects (Chu, De la Torre, & Cohn, 2013; Chen, Liu, Tu, & Aragones, 2013; Sebe, 2014). For instance, some subjects in the test set may have similar face shape, texture, or lighting to subsets of subjects in the training. These similarities could be used to optimize classifier generalizability between data sets. Preliminary work of this type has been encouraging. Using an approach referred to as a selective transfer machine, Chu et al. (2013) achieved improved generalizability between different data sets of spontaneous facial behavior.

In summary, we found that automated AU detection can be achieved in an unscripted social context involving spontaneous expression, speech, variation in head pose, and individual differences. Overall, we found very strong session-level reliability and moderate to strong frame-level reliability. The system was able to detect AUs in participants it had never seen previously. We conclude that automated FACS coding is ready for use in research and applied settings, where it can alleviate the burden of manual coding and enable more ambitious coding endeavors than ever before possible. Such a system could replicate and extend the exciting findings of seminal facial expression analysis studies as well as open up entirely new avenues of research.

Acknowledgments This work was supported in part by US National Institutes of Health grants MH096951 and AA015773.

References

Abrantes, G. A., & Pereira, F. (1999). MPEG-4 facial animation technology: Survey, implementation, and results. *IEEE Transactions on Circuits and Systems for Video Technology*, *9*(2), 290-305.

Ambadar, Z., Cohn, J. F., Reed, L. I. (2009). All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior*, *33*(1), 17–34.

Archinard, M., Haynal-Reymond, V., Heller, M. (2000). Doctor's and patients' facial expressions and suicide reattempt risk assessment. *Journal of Psychiatric Research*, *34*(3), 261–262.

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

Bruce, V., & Young, A. (1998). *In the eye of the beholder: The science of face perception*. New York: Oxford University Press.

Camras, L. A., Oster, H., Campos, J., Campos, R., Ujiie, T., Miyake, K. (1998). Production of emotional facial expressions in European American, Japanese, and Chinese infants. *Developmental Psychology*, *34*(4), 616–628.

Chen, J., Liu, X., Tu, P., Aragones, A. (2013). Learning person-specific models for facial expression and action unit recognition. *Pattern Recognition Letters*, *34*(15), 1964–1970.

Chu, W.-S., De la Torre, F., Cohn, J.F. (2013). Selective transfer machine for personalized facial action unit detection. *IEEE International Conference on Computer Vision and Pattern Recognition*, 3515–3522.

Chung, M. (2007). Correlation coefficient. In N. J. Salkin (Ed.) *Encyclopedia of measurement and statistics*, (pp. 189–201).

Cohn, J. F., & Ekman, P. (2005). Measuring facial action by manual coding, facial EMG, and automatic facial image analysis. In J. A. Harrigan, R. Rosenthal, K. R. Scherer (Eds.) *The new handbook of nonverbal behavior research*, (pp. 9–64). New York: Oxford University Press.

Cohn, J. F., & Sayette, M. A. (2010). Spontaneous facial expression in a small group can be automatically measured: An initial demonstration. *Behavior Research Methods*, *42*(4), 1079–1086.

Cohn, J. F., & Schmidt, K. L. (2004). The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets Multiresolution and Information Processing*, *2*(2), 57–72.

De la Torre, F., & Cohn, J.F. (2011). Facial expression analysis. In T. B. Moeslund, A. Hilton, A. U. Volker Krüger, L. Sigal (Eds.) *Visual analysis of humans*, (pp. 377–410). New York: Springer.

Ekman, P. (1982). Methods for measuring facial action. In K. R. Scherer, & P. Ekman (Eds.) *Handbook of methods in nonverbal behavior research*, (pp. 45–90). Cambridge: Cambridge University Press.

Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, *1000*(1), 205–221.

Ekman, P., & Friesen, W.V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., Ancoli, S. (1980). Facial signs of emotional experience. *Journal of Personality and Social Psychology*, *39*(6), 1125–1134.

Ekman, P., Friesen, W. V., Hager, J. (2002). *Facial action coding system: A technique for the measurement of facial movement*. Salt Lake City, UT: Research Nexus.

Ekman, P., & Heider, K. G. (1988). The universality of a contempt expression: A replication. *Motivation and Emotion*, *12*(3), 303–308.

Ekman, P., & Rosenberg, E. L. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*, 2nd edn. New York: Oxford University Press.

Fairbairn, C. E., Sayette, M. A., Levine, J. M., Cohn, J. F., Creswell, K. G. (2013). The effects of alcohol on the emotional displays of whites in interracial groups. *Emotion*, *13*(3), 468–477.

Fan, R.-e., Wang, X.-r., Lin, C.-j. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, *9*, 1871–1874.

Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology*, *60*(2), 12.

Geisser, S. (1993). *Predictive inference*. New York: Chapman and Hall.

Girard, J.M. (2014). *Automatic detection and intensity estimation of spontaneous smiles* (Master's thesis). Retrieved from http://d-scholarship.pitt.edu/19274/

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., Rosenwald, D. P. (2013). Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and Vision Computing*. Retrieved from doi:10.1016/j.imavis.2013.12.007

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Rosenwald, D. P. (2013). Social risk and depression: Evidence from manual and automatic facial expression analysis. *IEEE International Conference on Automatic Face & Gesture Recognition*, 1–8.

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., Lester, J. C. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. *International Conference on Educational Data Mining*.

Guney, F., Arar, N. M., Fischer, M., Ekenel, H. K. (2013). Cross-pose facial expression recognition. *IEEE International Conference and Workshops on Automatic Face & Gesture Recognition*, 1–6.

Hess, U., Banse, R., Kappas, A. (1995). The intensity of facial expression is determined by underlying affective state and social situation. *Journal of Personality and Social Psychology*, *69*(2), 280–288.

Hoque, M. E., McDuff, D. J., Picard, R. W. (2012). Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing*, *3*(3), 323–334.

Hsu, C.-W., Chang, C.-C., Lin, C.-J. (2003). *A practical guide to support vector classification* (Tech. Rep.)

Image Metrics. (2013). *LiveDriver SDK*. Manchester: Image Metrics.

Izard, C. E. (1979). *The maximally discriminative facial movement coding system (Max)*. Newark: University of Delaware, Instructional Resources Center.

Jeni, L. A., Cohn, J. F., De la Torre, F. (2013). Facing imbalanced data: Recommendations for the use of performance metrics. In *International conference on affective computing and intelligent interaction*.

Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, *68*(3), 441.

Keltner, D., & Buswell, B. N. (1997). Embarrassment: Its distinct form and appeasement functions. *Psychological Bulletin*, *122*(3), 250.

Keltner, D., Moffitt, T. E., Stouthamer-Loeber, M. (1995). Facial expressions of emotion and psychopathology in adolescent boys. *Journal of Abnormal Psychology*, *104*(4), 644–52.

Kraut, R. E., & Johnston, R. E. (1979). Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology*, *37*(9), 1539.

Li, Y., Chen, J., Zhao, Y., Ji, Q. (2013). Data-free prior model for facial action unit recognition. *IEEE Transactions on Affective Computing*, *4*(2), 127–141.

Littlewort, G., Bartlett, M. S., Fasel, I. R., Susskind, J., Movellan, J. R. (2006). Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, *24*(6), 615–625.

Littlewort, G., Whitehill, J., Tingfan, W., Fasel, I. R., Frank, M. G., Movellan, J. R., Bartlett, M. S. (2011). The computer expression recognition toolbox (CERT). *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 298–305.

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

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *IEEE International Conference on Computer Vision*, 1150–1157.

Lucey, P., Cohn, J. F., Howlett, J., Member, S. L., Sridharan, S. (2011). Recognizing emotion with head pose variation: Identifying pain segments in video. *IEEE Transactions on Systems, Man, and Cybernetics*.

Lucey, S., Matthews, I., Ambadar, Z., De la Torre, F., Cohn, J. F. (2006). AAM derived face representations for robust facial action recognition. *IEEE International Conference on Automatic Face & Gesture Recognition*, 155–162.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., Cohn, J. F. (2013). DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*.

McDuff, D., El Kaliouby, R., Kodra, E., Picard, R. (2013). Measuring voter's candidate preference based on affective responses to election debates. *HUMAINE Association Conference on Affective Computing and Intelligent Interaction*, 369–374.

Messinger, D. S., Mahoor, M. H., Chow, S.-M., Cohn, J. F. (2009). Automated measurement of facial expression in infant-mother interaction: A pilot study. *Infancy*, *14*(3), 285–305.

Noldus Information Technology. (2013). *The Observer XT*. Wageningen: The Netherlands.

Pantic, M., & Rothkrantz, L. J. M. (2000). Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, *18*(11), 881–905.

Powers, D. M. (2007). *Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation* (Tech. Rep.). Adelaide, Australia.

Powers, D. M. W. (2012). The problem with kappa. *Conference of the European Chapter of the Association for Computational Linguistics*, 345–355.

Prkachin, K. M., & Solomon, P. E. (2008). The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, *139*(2), 267–274.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, 2nd edn. Thousand Oaks: Sage.

Reed, L. I., Sayette, M. A., Cohn, J. F. (2007). Impact of depression on response to comedy: A dynamic facial coding analysis. *Journal of Abnormal Psychology*, *116*(4), 804–809.

Sayette, M. A., Creswell, K. G., Dimoff, J. D., Fairbairn, C. E., Cohn, J. F., Heckman, B. W., Moreland, R. L. (2012). Alcohol and group formation: A multimodal investigation of the effects of alcohol on emotion and social bonding. *Psychological Science*, *23*(8), 869–878.

Sayette, M. A., & Hufford, M. R. (1995). Urge and affect: A facial coding analysis of smokers. *Experimental and Clinical Psychopharmacology*, *3*(4), 417–423.

Sebe, N. (2014). We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *Proceedings of the ACM international conference on multimedia*. Orlando, FL.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., Poggio, T. (2005). A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *Artificial Intelligence*, 1–130.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86* (2), 420.

Szeliski, R. (2011). *Computer vision: Algorithms and applications*. London: Springer London.

Tian, Y.-l., Kanade, T., Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(2), 97–115.

Tong, Y., Chen, J., Ji, Q. (2010). A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(2), 258–273.

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

Valstar, M.F., Bihan, J., Mehu, M., Pantic, M., Scherer, K.R. (2011). The first facial expression recognition and analysis challenge. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 921–926.

Valstar, M. F., & Pantic, M. (2007). Combined support vector machines and hidden Markov models for modeling facial action temporal dynamics. In *IEEE international workshop on human-computer interaction* (pp. 118–127). Rio de Janeiro, Brazil: Springer-Verlag.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York, NY: Springer.

Vedali, A., & Fulkerson, B. (2008). *VLFeat: An open and portable library of computer vision algorithms*.

Yang, P., Qingshan, L., Metaxas, D. N. (2009). RankBoost with l1 regularization for facial expression recognition and intensity estimation. *IEEE International Conference on Computer Vision*, 1018–1025.

Zeng, Z., Pantic, M., Roisman, G. I., Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58.

Zhu, Y., De la Torre, F., Cohn, J. F., Zhang, Y.-J. (2011). Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior. *IEEE Transactions on Affective Computing*, *2*(2), 79–91.