In Press, *Psychophysiology*

Jeffrey F. Cohn, Adena J. Zlochower, James Lien, and Takeo Kanade. *Automated Face Analysis by Feature Point Tracking Has High Concurrent Validity with Manual FACS Coding*.

Department of Psychology, University of Pittsburgh and Robotics Institute, Carnegie Mellon University.

Abstract

The face is a rich source of information about human behavior. Available methods for coding facial displays, however, are human-observer dependent, labor intensive, and difficult to standardize. To enable rigorous and efficient quantitative measurement of facial displays, we have developed an automated method of facial display analysis. In this report we compare the results with those of manual FACS (Facial Action Coding System, Ekman & Friesen, 1978a) coding. One hundred university students were videotaped while performing a series of facial displays. The image sequences were coded from videotape by certified FACS coders. Fifteen action units and action unit combinations that occurred a minimum of 25 times were selected for automated analysis. Facial features were automatically tracked in digitized image sequences using a hierarchical algorithm for estimating optical flow. The measurements were normalized for variation in position, orientation, and scale. The image sequences were randomly divided into a training set and a cross-validation set, and discriminant function analyses were conducted on the feature point measurements. In the training set, average agreement with manual FACS coding was 92% or higher for action units in the brow, eye, and mouth regions. In the cross-validation set, average agreement was 91%, 88%, and 81% for action units in the brow, eye, and mouth regions, respectively. Automated Face Analysis by feature point tracking demonstrated high concurrent validity with manual FACS coding.

Automated Face Analysis by Feature Point Tracking Has High Concurrent Validity with Manual FACS Coding

The face is a rich source of information about human behavior. Facial displays indicate emotion (Ekman, 1993; Russell, 1994) and pain (Craig, Hyde, & Patrick, 1991), regulate social behavior (Cohn & Elmore, 1988; DePaulo, 1992; Fridlund, 1994), reveal brain function (Ekman, Davidson, & Friesen, 1990; Fox & Davidson, 1988) and pathology (Katsikitis & Pilowsky, 1988; Rinn, 1984), and signal developmental transitions in infants (Campos, Bertenthal, & Kermoian, 1992; Emde, Gaensbauer, & Harmon, 1976). To make use of the information afforded by facial displays, reliable, valid, and efficient methods of measurement are critical.

Current methods, which require human observers, vary in their specificity, comprehensiveness, degree of objectivity, and ease of use. The anatomically based Facial Action Coding System (FACS: Ekman & Friesen, 1978a; Baby FACS: Oster & Rosenstein, 1993) is the most comprehensive method of coding facial displays. Using FACS and viewing videotaped facial behavior in slow motion, coders can manually code all possible facial displays, which are referred to as "action units" (AUs). More than 7,000 combinations have been observed (Ekman, 1982). Ekman and Friesen (1978b) proposed that specific combinations of FACS action units represent prototypic expressions of emotion (i.e. joy, sadness, anger, disgust, fear, and surprise). Emotion expressions, however, are not part of FACS; they are coded in a separate system known as EMFACS (Friesen & Ekman, 1984) or the more recent FACS Interpretive Dictionary (Friesen & Ekman, undated, cited in Oster Hegley, & Nagel, 1992). FACS itself is purely descriptive and uses no emotion or other inferential labels.

Another anatomically based objective system, which also requires slow motion viewing of videotape, is the Maximally Discriminative Facial Movement Coding System (MAX: Izard, 1983). Compared with FACS, MAX is less comprehensive and was intended to include only facial displays (referred to as "movements" in MAX) related to emotion. MAX does not discriminate among some anatomically distinct displays (e.g., inner- and outer-brow raises) and considers as autonomous some displays that are not anatomically distinct (Oster et al., 1992). Malatesta (Malatesta, Culver, Tesman, & Shephard, 1989) added some displays in an effort to make MAX more comprehensive. Unlike FACS, MAX makes explicit claims that specific combinations of displays are expressions of emotion, and the goal of MAX coding is to identify these MAX-specified emotion expressions.

Whereas FACS and MAX use objective, physically measurable criteria, other videotape viewing systems are based on subjective criteria for facial expressions of emotions (AFFEX: Izard, Dougherty, & Hembree, 1983) and other expressive modalities (e.g., Monadic Phases: Cohn & Tronick, 1988; Tronick, Als, & Brazelton, 1980). The expression codes in these systems are given emotion labels (e.g., "joy") based on the problematic assumption that facial expression and emotion have an exact correspondence (Camras, 1992; Fridlund, 1994; Russell, 1994). Like FACS and MAX, these systems also require slow motion viewing of videotaped facial behavior.

As used below, "emotion expression" refers to facial displays that have been given emotion labels. Note that emotion expressions with the same label do not necessarily refer to the same facial displays. Systems such as MAX, AFFEX, and EMFACS can and do use the same terms when referring to different phenomena. For instance, Oster et al. (1992) found that MAX and the FACS Interpretive Dictionary gave different emotion labels to the same displays.

The lack of standard meaning to specific “emotion expressions” as well as the implication that emotion expressions represent subjective experience of emotion, are concerns about the use of emotion labels in referring to facial displays. The descriptive power of FACS, by contrast, has made it well suited to a broad range of substantive applications, including nonverbal behavior, pain research, neuropsychology, and computer graphics, in addition to emotion science (Ekman & Rosenberg, 1997; Parke & Waters, 1996; Rinn, 1984; 1991).

In everyday life, expressions of emotion, whether defined by objective criteria (e.g., combinations of FACS action units or MAX movement codes) or by subjective criteria, occur infrequently. More often, emotion is communicated by small changes in facial features, such as furrowing of the brows to convey negative affect. Consequently, a system that describes only emotion expressions is of limited use. Only FACS, and to a lesser extent MAX, can produce the detailed descriptions of facial displays that are required to reveal components of emotion expressions (e.g., Carroll & Russell, 1997; Gosselin, Kirouac & Dore, 1995). FACS action units are the smallest visibly discriminable changes in facial display, and combinations of FACS action units can be used to describe emotion expressions (Ekman & Friesen, 1978b; Ekman, 1993) and global distinctions between positive and negative expression (e.g., Moore, Cohn, & Campbell, 1997).

With extensive training, human observers can achieve acceptable levels of inter-observer reliability in coding facial displays. Human-observer-based (i.e., manual) methods, however, are labor intensive, semi-quantitative, and, with the possible exception of FACS, difficult to standardize across laboratories or over time. Training is time consuming (approximately 100 hours with the most objective methods), and coding criteria may drift with time (Bakeman & Gottman, 1986; Martin & Bateson, 1986). Implementing comprehensive systems is reported to take up to 10 hours of coding time per minute of behavior depending upon the comprehensiveness of the system and the density of behavior changes (Ekman, 1982). Such extensive effort discourages standardized measurement and may encourage the use of less specific coding systems with unknown convergent validity (Matias, Cohn, & Ross, 1989). These problems tend to promote the use of smaller sample sizes (of subjects and behavior samples), prolong study completion times, and thus limit the generalizability of study findings.

To enable rigorous, efficient, and quantitative measurement of facial displays, we have used computer vision to develop an automated method of facial display analysis. Computer vision has been an active area of research for some 30 years (Duda & Hart, 1973); early work included attempts at automated face recognition (Kanade, 1973, 1977). More recently, there is significant interest in automated facial display analysis by computer vision. One approach, initially developed for face recognition, uses a combination of principal components analysis (PCA) of digitized face images and artificial neural networks. High dimensional face images (e.g., 640 by 480 gray scale pixel arrays) are reduced to a lower dimensional set of eigenvectors, or “eigenfaces” (Turk & Pentland, 1991). The eigenfaces then are used as input to an artificial neural network or other classifier. A classifier developed by Padgett, Cottrell, and Adolphs (1996) discriminated 86% of six prototypic emotion expressions as defined by Ekman (i.e., joy, sadness, anger, disgust, fear, and surprise). Another classifier, developed by Bartlett and colleagues (Bartlett, Viola, Sejnowski, Golomb, Larsen, Hager, & Ekman, 1996), discriminated 89% of six upper face FACS action units.

Although promising, these systems have some limitations. First, because Padgett et al. (1996) and Bartlett et al. (1996) perform PCA on gray scale values, information about individual identity is encoded along with information about expression, which may impair discrimination.

Some robust lower-level image processing may be required to produce more robust discrimination of facial displays. Second, it is reported that eigenfaces are highly sensitive to minor variation in image alignment for the task of face recognition (Phillips, 1996). It is expected that similar or even better precision in image alignment is required when eigenfaces are used to discriminate facial displays. The image alignment used by Padgett et al. (1996) and Bartlett et al. (1996) was limited to translation and scaling, which is insufficient to align face images across subjects with face rotation. Third, these methods have been tested only on rather limited image data sets. Padgett et al. (1996) analyzed photographs from Ekman's and Friesen's Pictures of Facial Affect, which are considered prototypic expressions of emotion. Prototypic expressions differ from each other in many ways, which facilitates automated discrimination. Bartlett et al. (1996) analyzed images of subjects many of whom were experts in recognizing and performing FACS action units, and target action units occurred individually rather than being embedded within other facial displays. Fourth, Bartlett et al. performed manual time warping to produce a standard set of six pre-selected frames for each subject. Manual time warping is of variable reliability and is time consuming. Moreover, in many applications behavior samples are variable in duration, and therefore standardizing duration may omit critical information.

More recent research has taken optical-flow-based approaches to discriminate facial displays. Such approaches are based on the assumption that muscle contraction causes deformation of overlying skin. In a digitized image sequence, algorithms for optical flow extract motion from the subtle texture changes in skin, and the pattern of such movement may be used to discriminate facial displays. Specifically, the velocity and direction of pixel movement across the entire face or within windows selected to cover certain facial regions are computed between successive frames. Using measures of optical flow, Essa, Pentland, and Mase (Essa and Pentland, 1994; Mase, 1991; Mase & Pentland, 1990), and Yacoob and Davis (1994) discriminated among emotion-specified displays (e.g., joy, surprise, fear). This level of analysis is comparable to the objective of manual methods that are based on prototypic emotion expressions (e.g., AFFEX: Izard et al., 1983).

The work of Mase (1991), Mase and Pentland (1991) and Essa and Pentland (1994) suggested that more subtle changes in facial displays, as represented by FACS action units, could be detected from differential patterns of optical flow. Essa and Pentland (1994), for instance, found increased flow associated with action units in the brow and mouth region. The specificity of optical flow to action unit discrimination, however, was not tested. Discrimination of facial displays remained at the level of emotion expressions rather than the finer and more objective level of FACS action units. Bartlett et al. (1996) discriminated between action units in the brow and eye regions in a small number of subjects.

A question about optical-flow based methods is whether they have sufficient sensitivity to subtle differences in facial displays, as represented in FACS action units. Work to date has used aggregate measures of optical flow within relatively large facial regions (e.g., forehead or cheeks), including modal flow (Black & Yacoob, 1995; Rosenblum, Yacoob, & Davis, 1994; Yacoob & Davis, 1994) and mean flow within the region (Mase, 1991; Mase & Pentland, 1991). Black and Yacoob (1995) and Black, Yacoob, Jepson, and Fleet (1997) also disregard subtle changes in flow that are below an assigned threshold. Information about small deviations is lost when the flow pattern is aggregated or thresholds are imposed. As a result, the accuracy for discriminating FACS action units may be reduced.

The objective of the present study was to implement the first version of our automated method of face analysis and to assess its concurrent validity with manual FACS coding. Unlike

previous automated systems that use aggregate flow within large feature windows, our system tracks the movement of closely spaced feature points within very small feature windows (currently 13 by 13 pixels) and imposes no arbitrary thresholds. The feature points to be tracked are selected based on two criteria: they are in regions of high texture and represent underlying muscle activation of closely related action units. Discriminant function analyses are performed on the feature point measurements for action units in brow, eye, and mouth regions. The descriptive power of feature point marking is evaluated by comparing the results of a discriminant classifier based on feature point tracking with those of manual FACS coding.

Method

Image acquisition

Subjects were 100 university students enrolled in introductory psychology classes. They ranged in age from 18 to 30 years. Sixty-five percent were female, 15 percent were African-American, and three percent were Asian or Latino.

The observation room was equipped with a chair for the subject and two Panasonic WV3230 cameras, each connected to a Panasonic S-VHS AG-7500 video recorder with a Horita synchronized time-code generator. One of the cameras was located directly in front of the subject, and the other was positioned 30 degrees to the right of the subject. Only image data from the frontal camera are included in this report.

Subjects were instructed by an experimenter to perform a series of 23 facial displays that included single action units (e.g., AU 12, or lip corners pulled obliquely) and combinations of action units (e.g., AU 1+2, or inner and outer brows raised). Subjects began and ended each display from a neutral face. Before performing each display, an experimenter described and modeled the desired display. Six of the displays were based on descriptions of prototypic emotions (i.e., joy, surprise, anger, fear, disgust, and sadness). These six tasks and mouth opening in the absence of other action units were coded by one of the authors (AZ) who is certified in the use of FACS. Seventeen percent of the data were comparison coded by a second certified FACS coder. Inter-observer agreement was quantified with coefficient kappa, which is the proportion of agreement above what would be expected to occur by chance (Cohen, 1960; Fleiss, 1981). The mean kappa for inter-observer agreement was 0.86.

Action units which occurred a minimum of 25 times in the image data base were selected for analysis. This criterion ensured sufficient data for training and testing of Automated Face Analysis. When an action unit occurred in combination with other action units that may modify its appearance, the combination rather than the single action unit was the unit of analysis. Figure 1 shows the action units and action unit combinations thus selected. The action units we analyzed in three facial regions (brows, eyes, and mouth) are key components of emotion and other paralinguistic displays, and are common variables in emotions research. For instance, AU 4 is characteristic of negative emotion and mental effort, and AU 1+2 is a component of surprise. AU 6 differentiates felt, or Duchenne, smiles (AU 6+12) from non-Duchenne smiles (AU 12) (Ekman et al., 1990). In all three facial regions, the action units chosen are relatively difficult to discriminate because they involve subtle differences in appearance (e.g. brow narrowing due to AU 1+4 versus AU 4, eye narrowing due to AU 6 versus AU 7, three separate action unit combinations involving AU 17, and mouth widening due to AU 12 versus AU 20.). Unless otherwise noted, “action units” as used below refers to both single action units and action-unit combinations.

Figure 1 About Here

Image processing and analysis

Image sequences from neutral to target display (mean duration ~ 20 frames at 30 frames per second) were digitized automatically into 640 by 490 pixel arrays with 8-bit precision for gray scale values. Target displays represented a range of action unit intensities, including low, medium, and high intensity.

Figure 2 About Here

Image alignment. To remove the effects of spatial variation in face position, slight rotation, and facial proportions, images must be aligned and normalized prior to analysis. Three facial feature points were manually marked in the initial image: the medial canthus of both eyes and the uppermost point of the philtrum. Using an affine transformation, the images were then automatically mapped to a standard face model based on these feature points (Figure 2). By automatically controlling for face position, orientation, and magnification in this initial processing step, optical flows in each frame had exact geometric correspondence.

Figure 3 About Here

Feature point tracking. In the first frame, 37 features were manually marked using a computer mouse (leftmost image in Figure 3): 6 feature points around the contours of the brows, 8 around the eyes, 13 the nose, and 10 around the mouth. The inter-observer reliability of feature point marking was assessed by independently marking 33 of the initial frames. Mean inter-observer error was 2.29 and 2.01 pixels in the horizontal and vertical dimensions, respectively. Mean inter-observer reliability, quantified with Pearson correlation coefficients, was 0.97 and 0.93 in the horizontal and vertical dimensions, respectively.

The movement of feature points was automatically tracked in the image sequence using an *optical flow* algorithm (Lucas & Kanade, 1981). Given an n by n feature region R and a gray-scale image I, the algorithm solves for the displacement vector $\boldsymbol{d} = (d_x, d_y)$ of the original n by n feature region by minimizing the residual $E(\boldsymbol{d})$, which is defined as

$$E(\boldsymbol{d}) = \sum_{\boldsymbol{x} \in R} [\, I_{t+1}(\boldsymbol{x}+\boldsymbol{d}) - I_t(\boldsymbol{x}) \,]^2$$

where $\boldsymbol{x} = (x, y)$ is a vector of image coordinates. The Lucas-Kanade algorithm performs the minimization efficiently by using spatio-temporal gradients, and the displacements d_x and d_y are solved with sub-pixel accuracy. The region size used in the algorithm was 13-by-13. The algorithm was implemented by using an iterative hierarchical 5-level image-pyramid (Poelman, 1995), with which rapid and large displacements of up to 100 pixels (e.g., as found in sudden mouth opening) can be robustly tracked while maintaining sensitivity to subtle (sub-pixel) facial motion. On a dual-processor 300 MHz Pentium II computer with 128 megabytes of random access memory, processing time is approximately 1 second per frame.

The two images on the right in Figure 3 show an example of feature-point-tracking results. The subject's face changes from neutral (AU 0) to brow raise (AU 1+2), eye widening (AU 5), and jaw drop (AU 26), which is characteristic of surprise. The feature points are precisely tracked across the image sequence. Lines trailing from the feature points represent changes in their location during the image sequence. As the action units become more extreme, the feature point trajectory becomes longer.

Data analysis and action unit recognition

To evaluate the descriptive power of feature point tracking measurements, discriminant function analysis was used. Separate discriminant function analyses (DFA) were conducted on the measurement data for action units within each facial region. In the analyses of the brow region, the measurements consisted of the horizontal and vertical displacements of the 6 feature points around the brows. In the analyses of the eye region and of Duchenne versus non-Duchenne smiles, the measurements consisted of the horizontal and vertical displacements of the 8 feature points around the eyes. In analyses of the mouth region, the measurements consisted of the horizontal and vertical displacements of the 10 feature points around the mouth and four on the either side of the nostrils because of the latter's relevance to the action of AU 9. Therefore, each measurement was represented by a *2p* dimensional vector by concatenating *p* feature point displacements; that is $\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots\ldots, \boldsymbol{d}_p) = (d_{1x}, d_{1y}, d_{2x}, d_{2y}, \ldots d_{px}, d_{py})$.

The discrimination between action units was done by computing and comparing the *a posteriori* probabilities of action units AUs; that is

$$\boldsymbol{D} \rightarrow AU_k \quad \text{if } p(AU_k \mid \boldsymbol{D}) > p(AU_j \mid \boldsymbol{D}) \; j \neq k$$

where

$$p(AU_i \mid \boldsymbol{D}) = \frac{p(\boldsymbol{D} \mid AU_i)p(AU_i)}{p(\boldsymbol{D})} = \frac{p(\boldsymbol{D} \mid AU_i)p(AU_i)}{\sum_{j=1}^{k} p(\boldsymbol{D} \mid AU_j)p(AU_j)}$$

The discriminant function between AU_i and AU_j is therefore the log-likelihood ratio

$$f_{ij}(\boldsymbol{D}) = \log \frac{p(AU_i \mid \boldsymbol{D})}{p(AU_j \mid \boldsymbol{D})} = \log \frac{p(\boldsymbol{D} \mid AU_i)p(AU_i)}{p(\boldsymbol{D} \mid AU_j)p(AU_j)}$$

The $p(\boldsymbol{D} \mid AU_i)$ was assumed to be a multivariate Gaussian probability distribution $N(\boldsymbol{u}_i, \boldsymbol{\Sigma}_i)$, where the mean $\boldsymbol{u}_i$, and the covariance matrix $\boldsymbol{\Sigma}_i$ were estimated by the sample means and sample covariance matrices of the training data. This discriminant function is a quadratic discriminant function in general; but if covariance matrices $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_j$ are the same, it reduces to a linear discriminant function. Because we were interested in the descriptive power of the feature point displacement vector itself, rather than relying on other information (e.g., relative frequencies of action units in our specific samples), *a priori* probabilities $p(AU_i)$s were assumed to be equal.

The analyses used 872 samples of 15 action units or action unit combinations that occurred 25 or more times in 504 image sequences of 100 subjects. The samples were randomly divided into a training and a cross-validation set. However, if an action unit occurred in more than one image sequence from the same subject, all of the samples of that action unit by that subject were assigned to the training set. Thus, for each action unit, samples from the same subject belonged exclusively either to the training or the cross-validation set but not both. This strict criterion ensured that the training and the cross-validation set were uncorrelated with respect to subjects for each action unit, and thus that what was recognized was the action unit rather than the subject.

The agreement of action unit discrimination between manual FACS coding and Automated Face Analysis by feature point tracking was quantified. We used coefficient kappa (κ) to measure the proportion of agreement above what would be expected to occur by chance (Cohen, 1960; Fleiss, 1981). In preliminary analyses, subjects' race and gender were unrelated to classification accuracy and therefore were not included as factors in the discriminant function analyses and classification results reported below.

Results

Action units in the brow region

Three action units or action unit combinations (AU 1+2, AU 1+4, and AU 4) were analyzed in the brow region. Wilk's lambda and two discriminant functions were highly significant (λ = .07, p < .001, canonical correlations = .93 and .68, p < .001). In the training set, 93% of the action units were correctly classified (κ = .88). In the cross-validation set (Table 1), 91% were correctly classified (κ= 0.87); accuracy ranged from 74% for AU 1+4 to 95% and 97% for AU 1+2 and AU 4, respectively.

Insert Table 1 About Here

Action units in the eye region

Three action units (AU 5, AU 6, and AU 7) in the eye region were analyzed. Wilk's Lambda and two discriminant functions were highly significant (λ = 0.09, p < .001; canonical correlations = .91 and .67, p < .001). In the training set, 92% of action units were correctly classified (κ = .88). In the cross-validation set (Table 2), 88% were correctly classified (κ = 0.82). Disagreements that occurred were between AU 6 and AU 7.

We also evaluated recognition accuracy for Duchenne versus Non-Duchenne smiles; that is, a comparison of AU 6+12 with AU 12. Feature point data were restricted to the eye region. Wilk's lambda and one discriminant function were significant (λ = 0.45, p < .025; canonical correlation = .74, p < .05). In the training set, classification accuracy was 83% (κ = .67). In the cross-validation set, accuracy was 82% (κ = .63). (See Table 3). Errors resulted from over classification of AU 12.

Insert Tables 2 and 3 About Here

Action units in the mouth region

Nine action units were analyzed in the mouth region. Wilk's Lambda and five discriminant functions were highly significant (λ = 0.0006 , canonical correlations = .94, .93, .87, .76, and .63, p < .001). In the training set, 94% (κ = 0.93) were correctly classified. In the cross-validation set (Table 4), 81% were correctly classified (κ = 0.79). Accuracy was low for discriminating AU 26 from AUs 25 and 27 while accuracy for all other action units ranged from 73% to 100%.

Insert Table 4 About Here

Discussion

Facial displays are a rich source of information about human behavior, but that information has been difficult to obtain efficiently. Manual methods are labor intensive, semi-quantitative, difficult to standardize, and often subjective. Several recent studies have used computer-vision based approaches to discriminate facial displays (Bartlett et al., 1996; Cottrell & Metcalfe, 1991; Padgett et al., 1996; Essa & Pentland, 1994; Yacoob & Davis, 1994). Except for a study by Bartlett et al. (1996), this work has focused on discriminating a small number of emotion expressions (e.g., joy, surprise, and fear) that differ from each other in many facial regions, and the sample sizes used have been small, ranging from 7 to 20 subjects. We have developed an automated face analysis method that discriminates FACS action units, which are the smallest visibly discriminable facial displays with well-established objective criteria. We tested Automated Face Analysis with a large, varied data set.

To discriminate FACS action units, feature points in regions of moderate to high texture were automatically tracked in image sequences, and the effects of spatial variation were removed using an affine transformation of the feature point displacements. Using a discriminant classifier, average accuracy in the training set was above 90% for action units in the brow, eye, and mouth regions, and 83% for discriminating Duchenne from non-Duchenne smiles. In the cross-validation set, average accuracy was 91%, 88%, and 81% in the brow, eye, and mouth regions, respectively, and accuracy for Duchenne versus non-Duchenne smiles was 82%.

Automated Face Analysis demonstrated high concurrent validity with manual coding for action units in each of the facial regions studied. The level of inter-method agreement for action units was comparable to the accepted standard in tests of inter-observer agreement in FACS. The inter-method disagreements that did occur were generally the same ones that are common in FACS, such as the distinction between AU 25 and AU 26, and between AU 1+4 and AU 4.

Note that this test of the concurrent validity of Automated Face Analysis was performed with a larger, more heterogeneous data set than previous work. The data set consisted of more than 500 image sequence samples with 15 action units and action-unit combinations of 100 subjects. The image sequences contained positional and rotational motions of the face, and the set of action units spanned those in three facial regions (both upper and lower face). Action unit could occur either alone or embedded in others. Also, subjects included men and women of African-American and Asian ethnicity, providing a more adequate test of how well action unit discrimination would generalize to image sequences in new subjects. Automated Face Analysis was comparable to the accepted standard for manual coding, FACS.

In the present study, we used a restricted number of distinct features for action unit discrimination: feature points around the brows, eyes, nose, and mouth. We have not used other features in other regions, such as the forehead, glabella, infra-orbital furrow, cheeks, and the chin boss. Manual FACS coding looks for many types of movement in all of these facial regions when coding the action units analyzed here. AU 6, for instance, produces skin movement across the cheeks which is useful in discriminating AU 6 from AU 7. Feature point tracking in the cheek region would detect skin movement due to AU 6 and likely increase the accuracy of AU discrimination.

Many action units involve changes in transient features, such as lines or furrows, that may occur or vary across an image sequence. “Crows-feet” wrinkles, for instance, form at the eye corners from contraction of the orbicularis oculi in AU 6, and increases in the sclera above the eyeball occur with AU 5. These features can be represented by intensity gradients in the

image sequence and are quantified by the computer vision method of edge detection. For some action units, the use of edge detectors should prove essential. To discriminate between AU 25 and AU 26, FACS specifies a requisite distance between upper and lower teeth, which is readily detected by edge detectors but not by optical flow. By increasing the number of feature regions and supplementing feature point tracking and optical flow estimation with edge detection, further improvement in facial feature analysis can be achieved (Lien, Kanade, Zlochower, Cohn, & Li, 1998; Lien, Zlochower, Cohn, & Kanade, 1998).

In comparison with manual FACS coding, Automated Face Analysis represents a substantial improvement in efficiency. Manual FACS coding requires lengthy training and is time intensive. The current Automated Face Analysis requires feature point marking in a single frame of each image sequence, but it is fast and reliable with little training. It took us about 4 hours to mark manually the first frame in each of the 504 image sequences analyzed in the present study. After the initial reference points were marked, the facial features were tracked automatically in all subsequent images. On a 333 MHz Pentium II computer, the processing rate of automatic feature tracking was approximately 1 frame per second; processing of the 504 image sequences analyzed here required under 3 hours to complete. By contrast, manual FACS coding would require as much as 10 hours for each minute of image data (Ekman, 1982).

A major source of error in analyzing facial displays is global motion of the head across an image sequence. Movement toward, away from, or parallel to the image plane of the camera, as well as rotation in the image plane, is readily accommodated by automatically scaling, translating, and rotating the digitized images so that they are normalized with respect to the initial frame. When out-of-plane rotation varies within about ± 5 degrees, which was the case in the image sequences analyzed here, we found that these normalizations are sufficient. In many applications, however, larger out-of-plane rotations may occur. Intermediate rotations can be normalized by using an eight-parameter planar model to warp images to match with the initial frame (Black & Yacoob, 1985; Wu, Kanade, Cohn, & Li, 1998). For larger rotations, however, higher-degree motion models or multiple-camera setups may be needed (Basu, Essa, & Pentland, 1996; DeCarlo & Metaxis, 1995; Narayanan, Rander, & Kanade, 1998; Vetter, 1995). Multiple camera setups already are common in observational research, so the necessary recording capability is present in many laboratories.

The present analyses focused on the concordance between Automated Face Analysis and manual FACS coding in classifying action units and action unit combinations. Automated Face Analysis also provides a powerful tool with which to quantify the temporal dynamics of emotion displays. Ekman and Friesen (1982) theorized that false emotion expressions have a different temporal pattern than genuine ones (e.g., latency to apex is faster in false emotion expressions and they are punctuated by the occurrence of rapid micro-displays). Until now, hypotheses such as these have been difficult to test (See, for example, Frank, Ekman, & Friesen, 1993). Human observers have difficulty locating precise changes in behavior as well as in estimating changes in intensity of expression. Inter-observer agreement in locating the timing of action unit changes within a sequence is generally low (e.g., Ekman & Friesen, 1978b). Automated Face Analysis, by contrast, can precisely track quantitative changes on a frame-by-frame basis (Cohn, Zlochower, Lien, Wu, & Kanade, 1996). Small pixel-wise changes from frame to frame may be measured, and the temporal dynamics of facial displays can be determined.

In summary, Automated Face Analysis by feature point tracking demonstrated high concurrent validity with manual FACS coding. In the cross-validation set which included

subjects of mixed ethnicity average recognition accuracy for 15 action units in the brow, eye, and mouth regions was 81% to 91%. This is comparable to the level of inter-observer agreement achieved in manual FACS coding. We are extending Automated Face Analysis to incorporate convergent methods of quantifying facial displays, increase the number of action units and action unit combinations that can be recognized, and increase the generalizability of the system to a wide range of image orientations. We also have begun to use Automated Face Analysis to study emotion expression in infants (Zlochower, Cohn, Lien, & Kanade, 1998). With continued development, Automated Face Analysis will greatly reduce or eliminate the need for manual coding, make feasible the use of larger, more representative data sets, and open new areas of investigation.

References

Bakeman, R. & Gottman, J.M. (1986). Observing behavior: An introduction to sequential analysis. Cambridge: Cambridge University.

Bartlett, M. Stewart, Viola, P. A., Sejnowski, T. J., Golomb, B.A., Larsen, J., Hager, J. C., and Ekman, P. (1996). Classifying facial action, in D. Touretski, M. Mozer & M. Hasselmo (Eds.), Advances in Neural Information Processing Systems 8, pp. 823-829. Cambridge, MA: MIT.

Basu, S., Essa, I., & Pentland, A. (1996). Motion regularization for model-based head tracking. Proceedings of the International Conference on Pattern Recognition '96, 611-616.

Black, M.J. and Yacoob, Y. (June, 1995). Recognizing facial expressions under rigid and non-rigid facial motions. Proceedings of the IEEE International Workshop on Automatic Face and Gesture Recognition '95, pp. 12-17.

Black, M.J., Yacoob, Y., Jepson, A.D., and Fleet, D.J. (June, 1997). Learning Parameterized Models of Image Motion. Proceedings of the International Conference on Computer Vision and Pattern Recognition '97, 561-567.

Camras, L. (1992). Expressive development and basic emotions. Cognition and Emotion, 6, 269-283.

Campos, J.J., Bennett, I.B., & Kermoian, R. (1992). Early experience and emotional development: The emergence of wariness of heights. Psychological Science, 3, 61-64.

Carroll, J. M., & Russell, J. A. (1997). Facial expressions in Hollywood's portrayal of emotion. Journal of Personality and Social Psychology, 72, 164-176.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20, 37-46.

Cohn, J.F. & Elmore, M. (1988). Effect of contingent changes in mothers' affective expression on the organization of behavior in 3-month-old infants. Infant Behavior and Development, 11, 493-505.

Cohn, J.F. and Tronick, E.Z. (1988). Mother-infant interaction: Influence is bidirectional and unrelated to periodic cycles in either partner's behavior. Developmental Psychology, 24, 386-392.

Cohn, J.F., Zlochower, A., Lien, J., Wu, Y.T., & Kanade, T. (August, 1996). In N.H. Frijda (Ed.), Proceedings of the 9th Conference of the International Society for Research on Emotions, (pp. 329-333). Toronto, Canada.

Cottrell, G.W. & Metcalfe (1991). EMPATH: Face, emotion, and gender recognition using hologons. In R.P. Lippmann, J.E. Moody, & D.S. Touretzky (Eds.), Neural information processing systems, 3, (pp. 564-571).

DeCarlo, D. & Metaxas, D. (1995). Deformable model-based face shape and motion estimation. Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition '95, 146-150.

DePaulo, B.M. (1992). Nonverbal behavior and self-presentation. Psychological Bulletin, 111, 203-243.

Craig, K. D., Hyde, S. A., & Patrick, C. J. (1991). Genuine, suppressed and faked facial behavior during exacerbation of chronic low back pain. Pain, 46, 161-171.

Duda, R.O. & Hart, P.E. (1973). Pattern classification and analysis. NY: Wiley.

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.

Ekman, P. (1993). Facial expression and emotion. American Psychologist, 48, 384-392.

Ekman, P., Davidson, R.J., & Friesen, W.V. (1990). The Duchenne smile: Emotional expression and brain physiology II. Journal of Personality and Social Psychology, 58, 342-353.

Ekman, P. & Friesen, W.V. (1978a). Facial action coding system. Palo Alto: Consulting Psychologist Press.

Ekman, P. & Friesen, W.V. (1978b). Facial action coding system: Investigator's guide Part I. Palo Alto: Consulting Psychologist Press.

Ekman, P. & Friesen, W.V. (1982). Felt, false, and miserable smiles. Journal of Nonverbal Behavior, 6, 238-252.

Ekman, P. & Rosenberg, E. (1997). What the face reveals. NY: Oxford University.

Emde, R.N., Gaensbauer, T.J. & Harmon, R.J. (1976). Emotional expression in infancy: A biobehavioral study. Psychological Issues, 10 (No. 37), NY: International Universities.

Essa, I.A. & Pentland, A. (1994). A vision system for observing and extracting facial action parameters. Proceedings of the International Conference on Computer Vision and Pattern Recognition '94, 76-83..

Fleiss, J.L. (1981). Statistical methods for rates and proportions. NY: Wiley.

Fox, N. & Davidson, R.J. (1988). Patterns of brain electrical activity during facial signs of emotion in ten-month-old infants. Developmental Psychology, 24, 230-236.

Frank, M.G., Ekman, P., & Friesen, W.V. (1993). Behavioral markers and recognizability of the smile of enjoyment. Journal of Personality and Social Psychology, 64, 83-93.

Fridlund, A.J. (1994). Human facial expression: An evolutionary view. NY: Academic.

Friesen, W. V., & Ekman, P. (1984). EMFACS-7: Emotional Facial Action Coding System. Unpublished manuscript, University of California at San Francisco.

Friesen, W.V. & Ekman, P. (1992). Changes in FACS scoring. Unpublished manuscript, University of California, San Francisco.

Gosselin, P., Kirouac, G., & Dore, F. Y. (1995). Components and recognition of facial expression in the communication of emotion by actors. Journal of Personality and Social Psychology, 68, 83-96.

Izard, C.E. (1983). The Maximally Discriminative Facial Movement Coding System. Unpublished Manuscript, University of Delaware.

Izard, C.E., Dougherty, L.M., & Hembree, E.A. (1983). A system for identifying affect expressions by holistic judgments. Unpublished Manuscript, University of Delaware.

Kanade, T. (1973). Picture processing system by computer complex and recognition of human faces. Doctoral dissertation, Kyoto University.

Kanade, T. (1977). Computer recognition of human faces. Stuttgart and Busel: Birkhauser Verlag.

Katsikitis, M. & Pilowsky, I. (1988). A study of facial expression in Parkinson's disease using a novel microcomputer-based method. Journal of Neurology, Neurosurgery, and Psychiatry, 51, 362-366.

Lien, J., Kanade, T., Zlochower, A., Cohn, J.F., and Li, C.C. (June, 1998). A multi-method approach for discriminating between similar facial expressions, including expression intensity

estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, '98, xxx-xxx.

Lien, J.J., Zlochower, A., Cohn, J.F., & Kanade, T. (1998). Automated Facial Expression Recognition. Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, '98, 390-395.

Lucas, B.D. & Kanade, T. (1981). An iterative image registration technique with an application in stereo vision. Proceedings of the International Joint Conference on Artificial Intelligence, 7, 674-679.

Malatesta, C.Z., Culver, C., Tesman, J.R., & Shephard, B. (1989). The development of emotion expression during the first two years of life. Monographs of the Society for Research in Child Development, 54 (Serial No. 219).

Martin, P. & Bateson, P. (1986). Measuring behavior: An introductory guide. Cambridge: Cambridge University.

Mase, K. (1991). Recognition of facial expression from optical flow. IEICE of Japan Transactions, 10, 3474-3473.

Mase, K. & Pentland, A. (1990) Lip reading by optical flow, IEICE of Japan, Transactions, 6, 796-803.

Matias, R., Cohn, J.F., & Ross, S. (1989). A comparison of two systems to code infants' affective expression. Developmental Psychology, 25, 483-489.

Moore, G., Cohn, J.F., & Campbell, S.B. (1997). Mothers' affective behavior with infant siblings: Stability and change. Developmental Psychology, 33, 856-860.

Narayanan, P.J., Rander, P., & Kanade, T. (1998). Constructing virtual worlds using dense flow. Proceedings of the International Conference on Computer Vision '98, 3-10.

Oster, H. & Rosenstein, D. (1993). Baby FACS: Analyzing facial movement in infants. Unpublished manuscript, New York University, NY, NY.

Oster, H., Hegley, D., & Nagel, L. (1992). Adult judgments and fine-grained analysis of infant facial expressions: Testing the validity of a priori coding formulas. Developmental Psychology, 28, 1115-1131.

Padgett, C., Cottrell, G.W., & Adolphs, B. (1996, in press). Categorical perception in facial emotion classification. Proceedings of the Cognitive Science Conference, 18, 249-253.

Parke, F.I. & Waters, K.(1996). Computer facial animation. Wellesley, MA: A.K. Peters.

Philips, J. (1996, October). Face recognition: Where are we now and where are we going. Panel Session conducted at the International Conference on Automatic Face and Gesture Recognition, Killington, VT.

Poelman, C.J. (1995). The paraperspective and projective factorization methods for recovering shape and motion. (Technical report CMU-CS-95-173). Pittsburgh, PA: Carnegie Mellon University, Robotics Institute.

Rinn, W.E. (1984). The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. Psychological Bulletin, 95, 52-77.

Rinn, W.E. (1991). Neuropsychology of facial expression. In R.S. Feldman & B. Rime (Eds.), Fundamentals of nonverbal behavior. NY: Cambridge University.

Rosenblum, M., Yacoob, Y., and Davis, L.S. (November, 1994). Human emotion recognition from motion using a radial basis function network architecture. Proceedings of the Workshop on Motion of Non-rigid and Articulated Objects '94, 43-49.

Russell, J.A. (1994). Is there universal recognition of emotion from facial expression? Psychological Bulletin, 115, 102-141.

Tronick, E., Als, H., & Brazelton, T.B. (1980). Monadic phases: A structural descriptive analysis of infant-mother face-to-face interaction. Merrill-Palmer Quarterly of Behavior and Development, 26, 3-24.

Turk, M.A. & Pentland, A.P. (1991). Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3, 71-86.

Vetter, T. (1995). Learning novels views to a single face image. Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition '95, 22-29.

Wu, Y.T., Kanade, T., Cohn, J.F. and Li, C.C. (1998). Optical Flow Estimation Using Wavelet Motion Model. Proceedings of the International Conference on Computer Vision '98, 992-992.

Yacoob, Y. & Davis, L. (1994). Computing spatio-temporal representations of human faces. Proceedings in Computer Vision and Pattern Recognition '94, 70-75.

Zlochower, A.J., Cohn, J.F., Lien, J.J., & Kanade, T. (April, 1998). Automated face coding: A computer vision based method of facial expression analysis in parent-infant interaction. International Conference on Infant Studies, Atlanta, Georgia.

Author Note

This research was supported by NIMH grant R01 MH51435 to Jeffrey F. Cohn. Portions of the data were presented at the Seventh European Conference on Facial Expression, Measurement, and Meaning in Salzburg, Austria, August 1997 and the IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan 1998. Thanks are due to Ginger Moore for help in FACS coding, and Simon Cohn, Peter Rander, and Yu-Te Wu for technical assistance. Address correspondence to Jeffrey F. Cohn, Clinical Psychology Program, 4015 O'Hara Street, Pittsburgh, PA 15260.
Electronic mail: jeffcohn@vms.cis.pitt.edu

Table 1
Proportion of Agreement Between Automated Face Analysis and Manual Coding in Identifying Action Units in the Eyebrow Region

		Automated Face Analysis		
Manual Coding		AU 1+2	AU 1+4	AU 4
Training Set				
AU 1+2	(42)	.91	.10	.00
AU 1+4	(25)	.04	.88	.08
AU 4	(84)	.00	.05	.95
Cross-Validation Set				
AU 1+2	(43)	.95	.05	.00
AU 1+4	(19)	.00	.74	.26
AU 4	(32)	.00	.03	.97

Note. Number of samples of each AU appears in parentheses. κ = .88 and .87 in the training and the cross-validation sets, respectively.

Table 2

Proportion of Agreement between Automated Face Analysis and Manual Coding in Identifying Action Units in the Eye Region

	Automated Face Analysis		
Manual	AU 5	AU 6	AU 7
Coding			
Training Set			
AU 5 (41)	1.00	.00	.00
AU 6 (35)	.00	.77	.23
AU 7 (34)	.00	.03	.97
Cross-Validation Set			
AU 5 (28)	.93	.00	.07
AU 6 (33)	.00	.82	.18
AU 7 (14)	.00	.07	.93

Note. Number of samples of each AU appears in parentheses. κ = .88 and .82 in the training and the cross-validation set, respectively.

Table 3

Proportion of Agreement Between Automated Face Analysis and Manual Coding in Discriminating Between Non-Duchenne and Duchenne Smiles

		Automated Face Analysis	
Manual Coding		Non-Duchenne	Duchenne
Training Set			
Non-Duchenne	(25)	.88	.12
Duchenne	(35)	.20	.80
Cross-Validation Set			
Non-Duchenne	(12)	1.00	.00
Duchenne	(33)	.24	.76

Note. Number of samples of each AU appears in parentheses. κ = .67 and .63 in the training and the cross-validation set, respectively.

Table 4

Proportion of Agreement Between Automated Face Analysis and Manual Coding in Identifying Action Units in the Mouth Region

		Automated Face Analysis								
Manual Coding		AU 27	AU 26	AU 25	AU 12	AU 12+25	AU 20+25±16	AU 15+17	AU17+23+24	AU 9+17±25
Training Set										
AU 27	(28)	1.00	.00	.00	.00	.00	.00	.00	.00	.00
AU 26	(28)	.05	.86	.09	.00	.00	.00	.00	.00	.00
AU 25	(22)	.00	.00	1.00	.00	.00	.00	.00	.00	.00
AU 12	(15)	.00	.00	.00	.94	.06	.00	.00	.00	.00
AU 12+25	(37)	.00	.00	.00	.05	.95	.03	.00	.00	.00
AU	(31)	.03	.00	.00	.03	.06	.88	.00	.00	.00
20+25±16										
AU 15+17	(34)	.00	.00	.03	.00	.00	.00	.94	.03	.00
AU	(14)	.00	.00	.00	.00	.00	.00	.07	.93	.00
17+23+24										
AU 9+17±25	(18)	.00	.00	.00	.00	.00	.00	.00	.00	1.00

Table 4 continued

		AU 27	AU 26	AU 25	AU 12	AU 12+25	AU 20+25±16	AU 15+17	AU17+23+24	AU 9+17±25
Cross-Validation Set										
AU 27	(29)	.79	.10	.03	.00	.00	.07	.00	.00	.00
AU 26	(20)	.30	.52	.18	.00	.00	.00	.00	.00	.00
AU 25	(22)	.00	.14	.73	.00	.00	.00	.14	.00	.00
AU 12	(18)	.00	.00	.00	.83	.17	.00	.00	.00	.00
AU 12+25	(35)	.00	.00	.03	.00	.81	.17	.00	.00	.00
AU 20+25±16	(26)	.00	.03	.00	.00	.08	.89	.00	.00	.00
AU 15+17	(36)	.00	.00	.00	.00	.00	.03	.92	.06	.00
AU 17+23+24	(12)	.00	.00	.00	.08	.00	.00	.00	.92	.00
AU 9+17±25	(17)	.00	.00	.00	.00	.00	.00	.00	.00	1.00

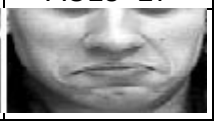
Note. Number of samples of each AU appears in parentheses. κ = .93 and .79 in the training and the cross-validation set, respectively.

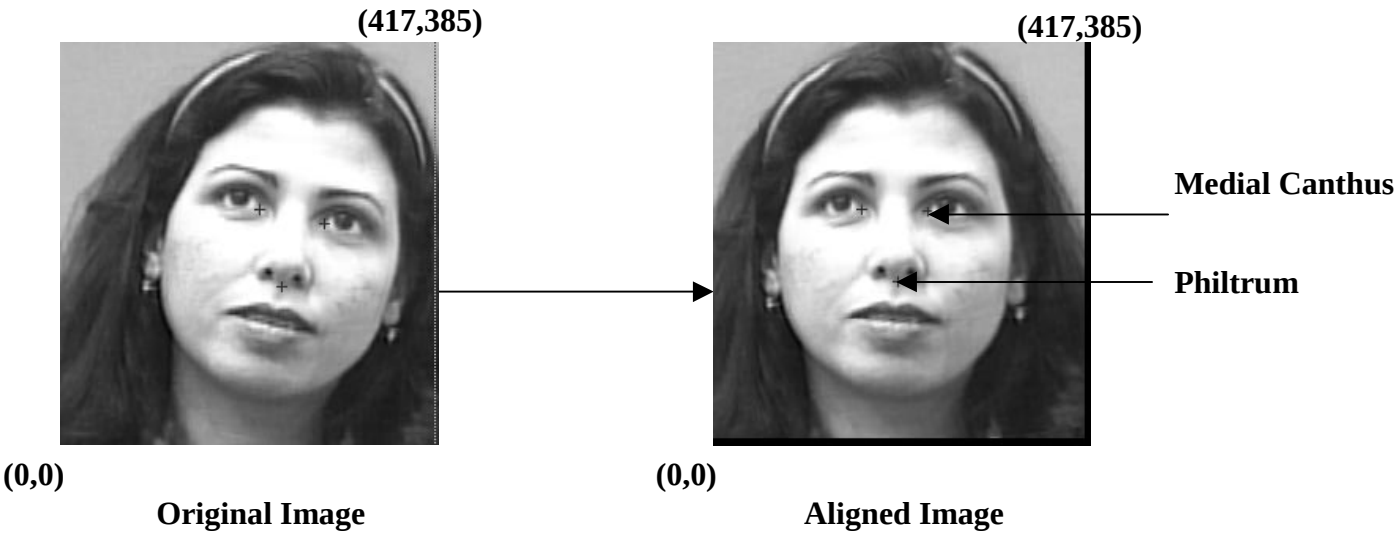
Figure Captions

Figure 1. Facial displays studied for Automated Face Analysis. Adapted from Ekman and Friesen (1978a).

Figure 2. Image alignment by affine transformation, which includes translation, scaling, and rotation factors. The medial canthi and philtrum are used as reference points.

Figure 3. Example of manually located feature points (leftmost image) and results of automated feature point tracking (two images on the right). The subject's expression changes from neutral (AU 0) to surprise (AU 1+2+5+26).

Upper Face		
AU4	AU1+4	AU1+2
		
Brows lowered and drawn together	Medial portion of the brows is raised and pulled together	Inner and outer portions of the brows are raised
AU5	AU6	AU7
		
Upper eyelids are raised	Cheeks are raised and eye opening is narrowed	Lower eyelids are raised
Lower Face		
AU25	AU26	AU27
		
Lips are relaxed and parted	Lips are relaxed and parted; mandible is lowered	Mouth is stretched open and the mandible pulled down
AU12	AU12+25	AU20+25
		
Lip corners are pulled obliquely	AU12 with mouth opening	Lips are parted and pulled back laterally
AU9+17	AU17+23+24	AU15+17
		
The infraorbital triangle and center of the upper lip are pulled upwards and the chin boss is raised (AU17)	AU17 and lips are tightened, narrowed, and pressed together	Lip corners are pulled down and chin is raised


(417,385)
(0,0)
Original Image
(417,385)
Medial Canthus
Philtrum
(0,0)
Aligned Image