Measuring Facial Expressions by Computer Image Analysis

Marian Stewart Bartlett[†]
Departments of Cognitive Science and Psychology
University of California, San Diego, and
The Salk Institute

Joseph C. Hager
Network Information Research Corp
Salt Lake City, Utah

Paul Ekman
Department of Psychiatry
University of California, San Francisco

Terrence J. Sejnowski
Department of Biology
University of California, San Diego, and
Howard Hughes Medical Institute
The Salk Institute

Accepted, *Psychophysiology*.

Running Head: Measuring Facial Expressions

This research was supported by NSF Grant No. BS-9120868, Lawrence Livermore National Laboratories Intra-University Agreement B291436, and Howard Hughes Medical Institute. We are indebted to FACS experts Harriet Oster, Linda Camras, Wil Irwin, and Erika Rosenberg for their time and assistance. We thank Gianluca Donato, Jan Larsen, and Paul Viola for contributions to algorithm development, Wil Irwin and Beatrice Golomb for contributions to project initiation, and Claudia Hilburn Methvin for image collection. Thanks to Gary Cottrell and two anonymous reviewers for valuable comments on earlier drafts of this paper. [†]Correspondence concerning this article should be addressed to Marian Stewart Bartlett, The Salk Institute - CNL, 10010 North Torrey Pines Road, La Jolla, CA 92037, email: marni@salk.edu, Fax: (619) 587-0417.

Abstract

Facial expressions provide an important behavioral measure for the study of emotion, cognitive processes, and social interaction. The Facial Action Coding System, (Ekman & Friesen, 1978), is an objective method for quantifying facial movement in terms of component actions. We have applied computer image analysis to the problem of automatically detecting facial actions in sequences of images. Three approaches were compared: Holistic spatial analysis, explicit measurement of features such as wrinkles, and estimation of motion flow fields. The three methods were combined in a hybrid system which classified six upper facial actions with 91% accuracy. The hybrid system outperformed human non-experts on this task, and performed as well as highly trained experts. An automated system would make facial expression measurement more widely accessible as a research tool in behavioral science and investigations of the neural substrates of emotion.

Keywords: Facial Expression / Facial Action Coding System / FACS / Emotion / Computer Image Analysis

Measuring Facial Expressions by Computer Image Analysis

Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski

Introduction

Facial expression measurement provides an indicator of emotion activity that is less intrusive than EEG, EMG, ANS or brain imaging measurements, and is presently used in a variety of areas of behavioral research, including the study of emotion, social interaction, communication, anthropology, personality, and child development (for reviews see Ekman, Huang, Sejnowski, & Hager, 1992; Ekman & Oster, 1979; Ekman & Rosenberg, 1997). Recent advances in computer vision and neural networks open up the possibility of automatic measurement of facial signals. An automated system would make facial expression measurement more widely accessible as a research tool in behavioral science and medicine, and would provide alternative measures of visual stimuli and behavioral responses in psychophysiological investigations into the neural substrates of emotion and facial expression recognition.

Measurement of facial signals

The Facial Action Coding System (FACS) (Ekman & Friesen, 1978) was developed to enable studies which directly measured facial behavior. Such studies included the differences in facial behavior when people are telling the truth versus lying, the patterns of central nervous system activity that accompany different facial movements, and whether facial behavior predicts clinical improvement. The difference between facial measurement as an approach to the study of facial expression versus measuring information that observers infer from facial expressions is reviewed in Ekman (1982a, 1982b).

FACS allows precise specification of the morphology and the dynamics of facial movement.[1] FACS was developed by determining from palpation, knowledge of anatomy, videotapes, and photographs how the contraction of each of the facial muscles changed the appearance of the face. Ekman and Friesen defined 46 Action Units, or AUs, to correspond to each independent motion of the face. FACS is coded from video, and a trained human FACS coder decomposes an observed expression into the specific AUs that occurred, and their duration, onset, and offset time. More than 300 people worldwide have achieved inter-coder agreement on the Facial Action Coding System. A number of studies have appeared showing the rich variety of information that can be obtained by using FACS (see Ekman & Rosenberg, 1997, for a review).

Although FACS is a promising approach, a major impediment to its widespread use is the time required to both train human experts and to manually score the video tape. It takes over 100 hours of training to achieve minimal competency on FACS, and each minute of video tape takes approximately one hour to score. Automating the FACS would make it more widely accessible as a research tool, and it would provide a good foundation for applications of automatic facial expression analysis in industry. An automated system would not only increase the speed of coding, it would also improve the reliability, precision, and temporal resolution of facial measurement.

Analysis of facial signals by computer

Some success has been achieved for automatic detection of facial actions by tracking the positions of dots attached to the face (Himer, Schneider, Kost, Heimann, 1991; Kaiser Wehrle, 1992). A system that detects facial actions from image sequences without requiring application of dots to the subjects face would have much broader utility. Efforts have recently turned to measuring facial actions by image processing of video sequences (Bartlett et al., 1996; Cohn, Zlochower, Lien, Wu, & Kanade, 1998; Lien, 1998). This paper explores and compares methods for classifying facial actions in image sequences of faces.[2]

[1]Electromyography also directly measures facial behavior, but it is obtrusive and not comprehensive.

[2]A brief report of this work appeared in Bartlett et al. (1996).

Recent advances have been made in computer vision for automatic recognition of facial expressions in images. The approaches that have been explored include analysis of facial motion (Mase, 1991; Yacoob & Davis, 1994; Rosenblum, Yacoob, & Davis, 1996, Essa & Pentland, 1997), measurements of the shapes and facial features and their spatial arrangements (Lanitis, Taylor, & Cootes, 1997), holistic spatial pattern analysis using techniques based on principal components analysis (Cottrell & Metcalfe, 1991; Padgett & Cottrell, 1997; Lanitis, Taylor, & Cootes, 1997) and methods for relating face images to physical models of the facial skin and musculature (Mase, 1991; Terzopoulos & Waters, 1993; Li, Roivainen, & Forcheimer, 1993; Essa & Pentland, 1997). These systems demonstrate approaches to face image analysis that are applicable to the present goals and are reviewed below, but the systems themselves are of limited use for behavioral and psychophysiological research.

Facial action codes versus emotion categories Most of the computer vision systems for recognizing facial expressions attempt to classify expressions into a few broad categories of emotion, such as happy, sad, or surprised. The evidence for seven universal facial expressions (see Ekman, 1989, for a review), does not imply that these emotion categories are sufficient to describe all facial expressions (Hager & Ekman, 1995). If automated facial measurement were to be constructed simply in terms of seven elementary emotional categories, much important information would be lost: blends of two emotions, variations within an emotional category (eg. vengeance vs. resentment), variations in intensity (annoyance vs. fury), conversational signals, and idiosyncratic facial movements.

Systems that only produced emotion category labels also could not be used in investigations of facial behavior itself. Several computer vision systems explicitly parameterize facial movement (Yacoob & Davis, 1994), and relate facial movements to the underlying facial musculature (Mase, 1991; Essa, & Pentland, 1997), but these descriptions are not readily interpretable in terms of facial action codes. It is unknown whether these descriptions are sufficient for describing the full range of facial behavior, and the relationship between these measures and internal state has not been established. A large body of empirical data already exists demonstrating the relationship of facial action codes to emotions, emotion intensity, variations, blends, and conversational signals.

Analysis of facial motion. The majority of the computer vision work on facial expression recognition has focused on facial motion analysis through optic flow estimation. If the tissues and muscles are similar between different people, the motions that result from facial action should be similar, independent of surface level differences between faces. In an early exploration of facial expression recognition, Mase (1991) used optic flow to estimate the activity in 12 of the 44 facial muscles. For each muscle he defined a window in the face image and an axis along which each muscle expands and contracts. The mean similarity of the flow vectors inside the window to this axis provided a coarse estimate of the activity of the muscle. Yacoob & Davis (1994) constructed a mid-level representation of facial motion from the optic flow output, which consisted of such descriptions as "right mouth corner raises.". The mid-level representation was then classified into one of six facial expressions using a set of heuristic rules. Rosenblum, Yacoob & Davis (1996) expanded this work to analyze facial expressions using the full temporal profile of the expression, from initiation, to apex, and relaxation. They trained radial basis function neural networks to estimate the stage of an expression from a facial motion description, and constructed separate networks for each expression. Radial basis functions approximate nonlinear mappings by Gaussian interpolation of examples, and are well suited to modeling systems with smooth transitions between states. Beymer, Shashua, and Poggio (1993) trained radial basis function neural networks to learn the transformation from optic flow fields to pose and expression coordinates, and from pose and expression coordinates back to optic flow fields. The estimated optic flow fields could be used to synthesize new poses or expressions from an example image by image warping techniques. The work most closely related to the approach in this paper is by Cohn et al. (1997), who are building a system to classify facial actions by facial feature point tracking. Over 40 points were manually located in the initial face image, and the displacements of these feature points were estimated by optic flow. Discriminant functions classified the displacements into 3 action classes in the brow region, 3 in the eye region, and 9 in the mouth region.

Model-based techniques. Several facial expression recognition systems have employed explicit physical models of the face (Mase, 1991; Terzopoulos & Waters, 1993; Li, Roivainen, & Forcheimer, 1993; Essa & Pentland, 1997). Essa & Pentland (1997) extended a detailed anatomical and physical model of the face developed by Terzopoulos and Waters (1993) and applied it to both recognizing and synthesizing facial expressions. The model consisted of a geometric mesh with 44 facial muscles, their points of attachment to the skin, and the elastic properties of the skin. Images of faces were mapped onto the physical model by image warping based on the locations of six points on the face. Motion estimates from optic flow were refined by the physical model in a recursive estimation-and-control framework, and the estimated forces were used to classify the facial expressions. In a model-based system, classification accuracy is limited by the validity of the model. There are numerous factors that influence the motion of the skin following muscle contraction, and it would be difficult to accurately account for all of them in a deterministic model. In this paper we take a neural network approach to image analysis in which facial action classes are learned directly from example image sequences of the actions, bypassing the physical model.

Feature-based approaches. One of the earliest approaches to recognizing facial identity in images was based on a set of feature measurements such as nose length, chin shape, and distance between the eyes (Kanade, 1977; Brunelli & Poggio, 1993). Lanitis, Taylor, & Cootes, (1997), recognized identity, gender, and facial expressions by measuring shapes and spatial relationships of a set of facial features using a flexible face model. An advantage of the feature-based approach is that it drastically reduces the number of input dimensions. A disadvantage is that the specific image features relevant to the classification may not be known in advance, and vital information may be lost when compressing the image into a limited set of features. Moreover, holistic graylevel information appears to play an important role on human face processing (Bruce, 1988; in press).

Holistic analysis. The alternative to feature-based image analysis, holistic analysis, emphasizes preserving the original images as much as possible and allowing the classifier to discover the relevant features in the images (Movellan, 1994). An example of this approach is template matching. Templates capture information about configuration and shape that can be difficult to parameterize. In related neural network approaches to image analysis, the physical properties relevant to the classification need not be specified in advance, and can be learned from the statistics of the image set. This is particularly useful when the specific features relevant to the classification are unknown (Valentin, Abdi, O'Toole, and Cottrell, 1994).

One holistic spatial representation is based on the principal components of the image pixels (Cottrell & Fleming, 1990; Turk & Pentland, 1991). Principal component analysis (PCA) finds an orthogonal set of dimensions that account for the principal directions of variability in the dataset. The component axes are template images that can resemble ghost-like faces which have been labeled "Holons" (Cottrell & Flemming, 1990) and "Eigenfaces" (Turk & Pentland, 1991). A low-dimensional representation of the face images with minimum reconstruction error is obtained by projecting the images onto the first few principal component axes. Principal components analysis has been applied successfully to recognizing both facial identity (Cottrell & Fleming, 1990; Turk & Pentland, 1991), and facial expressions (Cottrell & Metcalfe, 1991; Bartlett, Viola, Sejnowski, Golomb, Larsen, Hager & Ekman, 1996; Padgett & Cottrell, 1997). Another holistic spatial representation is obtained by a class-specific linear projection of the image pixels (Belhumeur, Hespanha, & Kriegman, 1997). Accurate alignment of the faces is critical to the success of such image-based approaches. Feature-based and template-based methods need not be mutually exclusive. Lanitis, Taylor, & Cootes, (1997), recognized identity, gender, and facial expressions by measuring shapes and spatial relationships of a set of facial features using a flexible face model. Performance improved by augmenting a set of feature measurements with parameters containing information about modes of variation in graylevel images based on principal component analysis.

Automating the Facial Action Coding System (FACS)

We explored three different methods for classifying facial actions that were suited to detecting different kinds of image cues: holistic spatial analysis based on principal components, a feature based approach that measures facial wrinkles and eye opening, and facial motion analysis based on template matching of optic flow fields. The performances of the three systems were compared and then combined into a single system that pools their strengths. One benchmark for the performances of the automated systems was provided by the ability of naive human subjects to classify the same images. A second benchmark was provided by the agreement rates of expert coders on these images.

Methods

Image Database We collected a database of image sequences of subjects performing specified facial actions. The full database contained over 1100 sequences containing over 150 distinct actions, or action combinations. The image database was obtained from 24 Caucasian subjects, 12 males and 12 females. Their ages ranged from 19 to 61 with a median of 30. 13 were experienced FACS coders, 8 had some FACS training, and 3 were naive. Each image sequence consisted of six frames, beginning with a neutral expression and ending with a high magnitude muscle contraction (Figure 1). The database therefore contained examples of the facial actions at low and medium magnitude as well as at high magnitude.[3] Trained FACS experts provided demonstrations and instructions to subjects on how to perform each action. The selection of images was based on stop motion video coded by three experienced FACS coders certified with high intercoder reliability. The criterion for acceptance of images was that the requested action and only the requested action was present.

For this investigation, we used data from 20 subjects and attempted to classify the six individual upper face actions illustrated in Figure 2. This set of actions was chosen for this study because the facial actions in the upper face comprise a relatively independent subset of facial actions; facial actions in the upper face have little influence on facial motion in the lower face, and vice versa (Ekman & Friesen, 1978). Most subjects were able to perform only a subset of the actions without interference from other facial muscles. Each subject performed a mean of 4 actions. The dataset therefore contained, aside from the neutral frame, a total of 400 images of facial actions (20 subjects X 4 actions X 5 frames per action). 9 subjects performed AU1, 10 performed AU2, 18 performed AU4, all 20 performed AU 5, 5 performed AU6, and 18 performed AU7.

Faces were aligned, cropped, and scaled based on the locations of two points in the first frame of each sequence. The two points were indicated by a single mouse click at the center of each eye. All other procedures were fully automated. Accurate image registration is critical for principal components based approaches. The variance in assigned eye location using this procedure was 0.4 pixels in the 640 x 480 pixel images.

The eye positions from frame 1 were used to crop all subsequent frames, and scale the faces to 45 pixels between the eyes. The images were rotated in the plane so that the eyes were horizontal, and the luminance brightness values were linearly rescaled to [0, 255]. The images were cropped to contain only the upper half of the face, as shown in Figure 2. The final images contained 66 x 96 pixels. Difference images, which were used in the holistic analysis, were obtained by subtracting the neutral expression frame (the first frame in each sequence), from the five subsequent frames. Advantages of difference images include robustness to changes in illumination, removal of surface variations in facial appearance, and emphasis of the dynamic aspects of the image sequence (Movellan, 1995).

Because faces tend to be asymmetric, and the contractions of facial muscles are also frequently asymmetric, we generated additional training data by reflecting each image about the vertical axis. Mirror reversed images of *test* subjects were never included in the training set, so the classifiers had no access to information about reflected test images either during parameter estimation or classification. The reflected images were not assumed to be independent of their originals, and were not counted in the N for statistical comparisons. All 400 difference images in the dataset were asymmetric. The

[3] The term "magnitude" replaces the term "intensity" used in FACS to avoid confusion with image intensity.

reflected images differed from their originals in 6125 of the 6336 pixels on average, and the mean magnitude of the difference was 5.36. Images differed *between* individuals in an average of 6179 pixels, and the mean magnitude of the difference *between* individuals was 7.17. The symmetry of the training set also ensured that the classifiers had no asymmetric bias.

Holistic spatial analysis We first evaluated the ability of a back-propagation network to classify facial actions given principal components of graylevel images as input. This approach is based on (Cottrell & Metcalfe, 1991) and (Turk & Pentland, 1991), with the primary distinction in that we performed principal components analysis on the dataset of difference images. The remaining variation in the dataset of difference images was that due to the facial dynamics. Each of the 800 difference images was converted to a vector by concatenating the rows of pixel intensities. The principal component axes of the difference image data were then calculated by finding the eigenvectors of the pixelwise covariance matrix. The axes were ordered by the magnitude of the corresponding eigenvalue. Figure 3 shows the first 12 principal components of the difference images.

The principal component representation consisted of a set of coefficients obtained by projecting each difference image onto the component axes. These coefficients comprised the input to a 2 layer neural network with 10 hidden units, and six output units, one per action. The network was feedforward, with each unit connected to all of the units in the layer above (see Haykin, 1994). The activities of the hidden and output units were calculated sequentially as the weighted sum of their inputs, passed through a sigmoidal hyperbolic tangent transfer function. The network was trained by back-propagation of error to output a 1 for the appropriate action, and zeros everywhere else, using conjugate gradient descent on the summed squared error. Stopping criterion was the inflection point in the mean test error. The output unit with the highest activity determined the classification.

Feature measurement Four of the upper face actions produce wrinkles in distinct locations on the face, and the remaining two alter the amount of visible sclera. We applied a method developed by Jan Larsen (Bartlett et. al. 1996) for measuring changes in facial wrinkling and eye opening. The feature measurements were carried out on 360 x 240 pixel images. Facial wrinkles were measured at the four facial positions shown in Figure 4a, which were located in the image automatically from the eye position information. These image locations were selected for detecting wrinkles produced by AUs 1,2,4, and 6. At each location, mean pixel intensities of a five pixel wide segment were extracted and then smoothed lengthwise by a median filter. Figure 4b shows the smoothed pixel intensities along the image segment labeled A. The pixel intensities drop sharply at the two major wrinkles.

We chose as a measure of facial wrinkling the sum squared derivative of the pixel intensities along the segment. This value is estimated by P (Figure 4c.) This measure is sensitive to both the deepening of existing wrinkles and the addition of new wrinkles. To control for permanent wrinkles, P values for the neutral image were subtracted. Figure 4d shows P values along line segment A, for a subject performing each of the six actions. The P values remain at zero except for AU 1, for which it increases as action magnitude increases. Only AU 1 produces wrinkles in the center of the forehead.

For detecting and discriminating AUs 5 and 7, we defined an eye opening measure as the area of visible sclera lateral to the iris. This area was found by starting at the pupil and searching laterally for connected rows of pixels above threshold. Again, differences from baseline were measured. A three-layer neural network was trained to classify each image from the five feature measures, consisting of the wrinkle feature measured at 4 locations and the eye opening measure. The network had 15 hidden units and six output units.

Optic flow Local estimates of motion in the direction of the image gradient were obtained by an algorithm based on the brightness constraint equation (Horn & Schunk, 1981):

$$\frac{dI(x,y,t)}{dt} = \frac{\partial x}{\partial t}\frac{\partial I(x,y,t)}{\partial x} + \frac{\partial y}{\partial t}\frac{\partial I(x,y,t)}{\partial y} + \frac{\partial I(x,y,t)}{\partial t} = 0 \qquad (2)$$

This equation assumes that there is no overall gain or loss of brightness in the image I over time, and any changes in brightness can be accounted for by shifts in spatial position. The local image velocities, $v_x = \frac{\partial x}{\partial t}$ and $v_y = \frac{\partial y}{\partial t}$, are defined in terms of the spatial and temporal gradients of the image, $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$, and $\frac{\partial I}{\partial t}$.

Optic flow was estimated between image pairs, a given frame in an action sequence, t_i, and the neutral frame, t_0. Images were smoothed by a 5 x 5 Gaussian kernel. Estimates of the spatial gradients, ΔI_x and ΔI_y, were obtained with horizontal and vertical Sobel edge filters. The temporal gradient was estimated by $\Delta I_t = I(x, y, t_i) - I(x, y, t_0)$. Local estimates of image velocity in the direction of the gradient were obtained by $v_x = \frac{\Delta I_t}{\Delta I_x}$ and $v_y = \frac{\Delta I_t}{\Delta I_y}$.

Gradient-based techniques for estimating optic flow give reliable estimates only at points where the gradient is high (ie. at moving edges). Velocity estimates were set to zero at locations at which the total edge measure $r = {\Delta I_x}^2 + {\Delta I_y}^2$ was beneath a threshold of 0.2. An example flow field is shown in Figure 5. One of the advantages of this simple local estimate of flow was speed. It took 0.13 seconds on a 120 MHz Pentium to compute one flow field.

The flows fields were classified by a template matching procedure. A weighted template for each of the actions was calculated from the training images as the mean flow field at medium action magnitude (frame 4 of the sequence). Novel flow patterns, f^n, were compared to the template f^t by the correlational similarity measure S:

$$S(f^n, f^t) = \frac{\sum_i f_i^n \cdot f_i^t}{\sqrt{\sum_i f_i^n \cdot f_i^n}\sqrt{\sum_i f_i^t \cdot f_i^t}} \tag{3}$$

where i indexes image location. $S(f^n, f^t)$ is the cosine of the angle between the two flow vectors.

Naive human subjects Subjects were nine adult volunteers with no prior knowledge of facial expression measurement. Subjects were provided with a guide sheet similar to Figure 2 which contained an example image of each of the six actions along with a written description of each action and a list of image cues for detecting and discriminating the actions from Ekman and Friesen (1978). Each subject was given a training session in which the facial actions were described and demonstrated, and the image cues listed on the guide sheet were reviewed and indicated on the example images. The subjects kept the guide sheet as a reference during the task.

Face images were cropped and scaled identically as they had been for the automated systems, with 45 pixels between the eyes, and printed using a high resolution HP Laserjet 4si printer with 600 dpi. Because the automated systems had information about the test image and the neutral image only when making a classification, face images were presented to the human subjects in pairs, with the neutral image and the test image presented side by side. Subjects were instructed to compare the test image with the neutral image and decide which of the actions the subject had performed in the test image. Subjects were given a practice session with feedback consisting of one example of each action at high magnitude. Neither the practice face nor the reference face was used for testing. The task contained ninety-six image pairs, consisting of low, medium, and high magnitude examples of the six actions from six different faces, three male and three female. Subjects were allowed to take as much time as they needed to perform the task, which ranged from 30 minutes to one hour.

Expert coders Subjects were four certified FACS coders. The task was identical to the naive subject task with the following exceptions: Expert subjects were not given a guide sheet or additional training, and the complete face was visible, as it would normally be during FACS scoring. One hundred and fourteen image pairs were presented, consisting of low, medium, and high action magnitude examples of the six actions from seven faces. Time to complete the task ranged from 20 minutes to one hour and 15 minutes.

Results

Generalization to novel faces was tested using leave-one-out cross-validation (Tukey, 1958). This procedure makes maximal use of the available data for estimating parameters. System parameters were estimated 20 times, each time using images from 19 subjects for training and and reserving all of the images from one subject, including the reflected images, for testing. The system parameters were deleted and re-estimated for each test. Mean classification performance across all test images in the 20 cross-validation runs was then calculated.

Under this procedure there were 800 test images, containing low, medium and high magnitude examples of the facial actions. The systems classified the test images one frame at a time, without reference to previous outputs. Figure 6 plots the overall mean performances of the classifiers on novel faces. Performances by facial action are the diagonal entries in the confusion matrices in Tables 1 and 2.

Holistic spatial analysis Classification performance was evaluated for two scales of difference images, 66 x 96 and 22 x 32, and for five quantities of principal components in the network input: 10, 25, 50, 100, and 200. There was a trade-off between increasing the amount of information in the input and increasing the number of free parameters to be estimated. The higher principal components may also include more information on between subject variations. We obtained the best performance of 88.6% using the first 50 principal components of the 22 x 32 difference images.

The holistic system with 50 principal components had 580 parameters, while our training set in a given training run contained on average 760 images. Over-parameterization is a risk with such high dimensional networks. Performance for generalization to novel faces provided a measure of how well the system performed the general class discrimination, as opposed to finding a trivial solution that minimized the error for the training samples without learning the class discrimination.

The performance of 88.6% is substantially higher than the 70% performance reported by Padgett & Cottrell (1997) for facial expression classification using full-face Eigenfaces. The success of the present system could be attributable to reduced variability due to the use of difference images, or to the smaller original image size, so that 50 principal components accounted for a greater percentage of the variability. In addition, we employed a region of interest analysis, consisting of half of the face image, which is similar to the "Eigenfeature" approach that gave Padgett & Cottrell better performance.

Feature measurement The performance of the feature-based classifier on novel faces was lower than the other methods, at 57% correct. Normalization of the feature measures with Z-scores did not improve performance. The classifier was most accurate for the two actions that involved changes in eye opening, AU5 and AU7, at 74% and 62% correct respectively. The poor performance for novel faces may be attributable to the differences in facial wrinkling patterns between subjects depending on skin elasticity, facial structure, and fat stores. The feature-based classifier performed well for new images of a face used for training, with classification accuracy of 85.3%.

Optic flow Template matching of motion flow fields classified the facial actions with 84.5% accuracy for novel subjects. The performance of the motion-based classifier was similar to that of the holistic classifier, giving highest accuracy for AUs 2. 4, 5, and 7, and lowest for AUs 1 and 6.

Hybrid System We obtained the best performance when we combined all three sources of information into a single neural network. The classifier was a feed forward network with 10 hidden units taking 50 component projections, 5 feature measures, and 6 template matches as input. The hybrid system improved the generalization performance to 90.9%, over the best individual method at 88.6%. While the increase is small, it constitutes about 20% of the difference between the best individual classifier and perfect performance.

We examined how the hybrid system benefited from the multiple sources of input information by looking at correlations in the performances of the three individual classifiers. The contribution of additional inputs to the signal-to-noise ratio depends on their correlations. Each data point in the correlation was mean percent correct for one of the twenty faces, across all actions and action magnitudes. The performances of the holistic and the flow field classifiers were correlated ($r^2 = 0.36, t(18) = 2.96, p < 0.01$). The feature-based system was not correlated with either the holistic or flow field classifiers ($r^2 = 0.05, t(18) = 0.85, p > 0.4$) and ($r^2 = 0.02, t(18) = 0.65, p > 0.5$), respectively. Although the stand-alone performance of the feature-based system was low, it contributed to the hybrid system by providing estimates that were uncorrelated with the two template-based systems. Without the feature measures, 17% of the improvement was lost.

Human subjects A benchmark for the performance of the automated systems was provided by the performance of naive human subjects on the same set of images with identical cropping and scaling. Human non-experts classified the images with 73.7% accuracy. This is a difficult classification problem that requires considerable training for people to be able to perform well. Performance of the naive human subjects was significantly lower than that of the hybrid system on the subset of images used in the human study ($Z = 2.04, p < 0.05$).

A second benchmark was provided by the agreement rates of expert coders on these images. The expert human subjects classified the actions with 91.8% agreement with the class labels assigned during database collection, which is well above the FACS inter-coder agreement standard for proficiency. The majority of the disagreement was on the low magnitude examples of the actions, and the absence of video motion could account for much of the disagreement. Because the images were originally labeled by two expert coders with access to stop-motion video, this data provides a measure of inter-coder *agreement* between coding stop-motion video and static images. The performance of the holistic and hybrid computer systems did not differ significantly from that of the human experts (Z = 1.63; Z = 1.86), but the expert coders did outperform the optic flow and feature-based classifiers ($Z = 3.17, p < 0.01$) and ($Z = 7.2, p < 0.001$).

Error analysis The action confusions made by both naive and expert human subjects are presented in Table 1. Naive subjects made the most confusions between AUs 6 and 7, which both alter the appearance underneath the eye, followed by AUs 2 and 5, which both give an eye widening appearance by raising the outer brows and the upper lid respectively, followed by AUs 1 and 2, which raise the inner and outer portions of the eyebrows, respectively. The majority of the disagreements for the experts were between AUs 6 and 7.

Table 2 shows the action confusions made by the three image analysis systems and the hybrid system. Correlations among the action confusions are given in Table 3. Consistent with the performance rate comparisons, the confusions made by the holistic system were highly correlated with those of the motion-based system, whereas the confusions made by the feature-based system were less correlated with those of the holistic system, and uncorrelated with those of the motion-based system.

Of the four automated systems, the holistic system had the most similar pattern of confusions to both the naive human subjects and to the expert coders. This finding is consistent with previous reports that principal component representations of face images account well for human perception of distinctiveness and recognizability of faces (O'Toole, Deffenbacher, Valentin, & Abdi, 1994; Hancock, Burton, & Bruce, 1996). The confusions of the feature-based system were least correlated with those of the human subjects, with a low but significant correlation with the expert coders, and no significant correlation with the naive subjects.

Discussion

Facial action codes provide a rich description of facial behavior that enables investigation of the relationship of facial behavior to internal state. We developed methods for automatically classifying facial actions from image sequences. The approach presented here differed from other computer facial expression analysis systems in that we focused on classifying the basic elements that comprise complex facial movements rather than classifying emotion categories. Classification was learned directly from images of facial actions without mediation of a physical model.

We compared the performance of three diverse approaches to processing face images for classifying facial actions: holistic spatial analysis, feature measurement, and analysis of motion flow fields. Best performance of 92% correct for classifying 6 actions was achieved by combining the three methods of image analysis in a single system. The hybrid system classified an image in less than a second on a 120 MHz Pentium. Our initial results are promising since some of the upper facial actions included in this study require extensive training for humans to discriminate reliably. The holistic and hybrid automated systems outperformed human non-experts on this task, and the hybrid system performed as well as highly trained experts.

The image analysis methods did not depend on the precise number of video frames, nor that the actions be of any particular magnitude beyond the neutral frame. For applications in which neutral images are unavailable, principal component analysis could be performed on the original graylevel images. Methods based on principal component analysis have successfully classified static graylevel images of facial expressions (Padgett & Cottrell, 1997). The image analysis also required localization of the face in the image. For this study, the localization was carried out by making two mouse clicks, one at the center of each eye, in the first frame of the sequence. All other aspects of the systems were fully automated. Highly accurate eye location algorithms are available (eg. Beymer, 1994), and automating this step is a realistic option. The image alignment procedure ignored out-of-plane rotations, which could be handled by methods for estimating the frontal view of a face from a non-frontal view (eg. Beymer Shashua & Poggio, 1993; Vetter & Poggio, 1997).

There are 46 action units, of which we have presented classification results for 6. The holistic and motion-based systems are not specific to particular actions, and can be applied to any other facial motion. The image analysis in these systems was limited to the upper half of the face because upper facial actions have little effect on motion in the lower face, and vice versa (Ekman & Friesen 1978). We are presently applying these techniques to images of the lower half of the face to classify the lower facial actions as well.

It remains an empirical question to determine whether this approach will have the same success when dealing with spontaneous rather than deliberately made facial actions. While the morphology of the facial actions should not differ in spontaneous as compared to deliberate facial actions, the timing of the activity and the complexity of facial actions may well be different. Evaluating spontaneous facial movement is an important next step.

Cohn et al. (1998) are developing a related system for automatic facial action coding that takes advantage of the precision obtainable through human-computer interaction. In their system, more than 40 feature points are manually identified in the initial image. The system presented here is more automatable since human interaction in our system was limited to the two mouse clicks in the initial image, described above. Another difference between the two systems is that the Cohn et al. system estimates displacements in a select set of feature points whereas our system captures full field information on skin motion.

Most automatic facial expression analysis systems have focused on either motion or surface graylevels, but not both. It should be noted that while human subjects *can* recognize facial expressions from motion signals alone (Bassili, 1979), recognition rates are only just above chance. Likewise, although humans can recognize facial expressions quite well from static graylevel images, expression recognition improves with high temporal resolution video (Wallbott, 1992). This system integrates both analysis of surface graylevels and motion information.

We found that the two template-based methods, holistic spatial analysis and motion analysis, outperformed the feature-based method for facial action recognition. This supports previous findings that template approaches outperformed feature-based systems for recognizing faces (Brunelli & Poggio, 1993; Lanitis, Taylor, & Cootes, 1997). This result is also supported by Lien (1998), who found that facial furrow measurement based on analysis of high image gradients did not perform as well as full field motion analysis for facial action classification.

Our results also suggest that hand-crafted features plus templates may be superior to either one alone, since their performances may be uncorrelated. Classification of local feature measurements is heavily dependent on exactly which features were measured. Padgett & Cottrell (1997) found that local principal component analysis was superior to full-face Eigenfaces for expression recognition. These local features were based on data-driven kernels obtained from the graylevels of the face images, as opposed to the hand-crafted feature measures that performed poorly in this study and others (e.g. Brunnelli & Poggio, 1993). We are presently exploring local representations of faces based on the outputs of local filters such as Gabor wavelets and local principal component analysis for facial action classification.

A completely automated method for scoring facial actions in images would make facial expression measurement more widely accessible as a research tool in behavioral science, medicine, and psychophysiology. Facial action codes have already proven a useful behavioral measure in studies of emotion (e.g. Ekman, 1984), human interaction and communication (e.g. Ekman & Oster, 1979), cognition (e.g. Zajonc, 1984), and child development (e.g. Camras, 1977). Measurement of observable facial behavior has been combined with simultaneous scalp EEG in the study of physiological patterns associated with emotional states (e.g. Davidson, Ekman, Saron, Senulis, & Friesen, 1990), and with measures of autonomic nervous system activity to study the relationship of emotion to facial muscles and the autonomic nervous system (Ekman, Levenson, & Friesen, 1983).

Neuropsychological investigations in humans and physiological recordings in primates have indicated a separate neural substrate for recognizing facial expression independent of identity (Tranel, Damasio, & Damasio, 1988; Adolphs, Tranel, Damasio, & Damasio, 1995; Hasselmo, Rolls & Baylis, 1989), and there is evidence that the recognition of specific facial expressions depends on distinct systems (e.g. Adolphs, Damasio, Tranel, & Damasio, 1996). Neural substrates for the perception of two negative emotions, fear and disgust, have recently been differentiated using fMRI (Phillips et al., 1997). Whereas perception of expressions of fear and anger produced activation in the amygdala (Brieter et al. 1996; Morris et al., 1996), perception of disgust in others activated interior insular cortex, an area involved in responses to offensive tastes (Yaxley, Rolls, & Seinkiewitz, 1988; Kinomura et al. 1994).

Automated facial action coding would provide an objective measure of visual stimuli in such investigations of the neural substrates for the perception of facial expressions, as well as providing a behavioral measure of emotional state. An automated system would improve the reliability, precision, and temporal resolution of facial measurement, and would facilitate the use of facial measurement in psychophysiological investigations into the neural systems mediating emotion.

Acknowledgements

This research was supported by NSF Grant No. BS-9120868, Lawrence Livermore National Laboratories Intra-University Agreement B291436, and Howard Hughes Medical Institute. We are indebted to FACS experts Harriet Oster, Linda Camras, Wil Irwin, and Erika Rosenberg for their time and assistance. We thank Gianluca Donato, Jan Larsen, and Paul Viola for contributions to algorithm development, Wil Irwin and Beatrice Golomb for contributions to project initiation, and Claudia Hilburn Methvin for image collection. Thanks to Gary Cottrell and two anonymous reviewers for valuable comments on earlier drafts of this paper.

References

Adolphs, R. Damasio, H., Tranel, D., & Damasio, A. (1996). Cortical systems for the recognition of emotion in facial expressions. *Journal of Neuroscience*, 16(23), 7678-7687.

Adolphs, R. Tranel, D. Damasio, H., & Damasio, A. (1995). Fear and the human amygdala. *Journal of Neuroscience,* 15(9), 5879-5891.

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. Touretzki, M. Mozer, & M. Hasselmo (Eds.) *Advances in Neural Information Processing Systems 8* (pp. 823-829). Cambridge, MA: MIT Press.

Bassili, J., (1979). Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology, 37*, 2049-2059.

Belhumeur, P.N., Hespanha, J.P., & Kriegman, D.J. (1997). Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE transactions on Pattern Analysis and Machine Intelligence* 15(6), 711-720.

Beymer, D. 1994. Face recognition under varying pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Los Alamitos, CA: IEEE Comput. Soc. Press: 756-61.

Beymer, Shashua, & Poggio (1993). Example based image analysis and synthesis. Massachusetts Institute of Technology AI Memo No. 1431.

Breiter H.C., Etcoff N.L., Whalen P.J., Kennedy W.A., Rauch S.L., Buckner R.L., Strauss M.M., Hyman S.E., Rosen B.R. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron 17*(5), 875-87.

Bruce, V. (1988). *Recognising Faces.* London: Lawrence Erlbaum Assoc.

Bruce, V. (in press). Human face perception and identification. In H. Wechsler, V. Bruce, J. Phillips, S. Fogelman-Soulie, & T. Huang (Eds.) *Face Recognition, from Theory to Applications.* NATO ASI Series F. New York, London: Springer-Verlag.

Brunelli, R., & Poggio, T. (1993). Face recognition: Features versus templates. *IEEE transactions on pattern analysis and machine intelligence 15*(10), 1042-1052.

Camras, L.A., (1977). Facial expressions used by children in conflict situations. *Child Development, 48*, 1431-35.

Cohn, J.F., Zlochower, A.J., Lien, J.J., Wu, Y-T, & Kanade, T. (1998). Automated face coding: A computer-vision based method of facial expression analysis. *Psychophysiology*, in press.

Cottrell, G.W. & Fleming M.K. (1990). Face recognition using unsupervised feature extraction. *Proceedings of the International Neural Network Conference* (pp. 322-325). Paris, France, Dordrecht: Kluwer.

Cottrell, G.,& Metcalfe, J. (1991). EMPATH: Face, gender and emotion recognition using holons. In R.P. Lippman, J. Moody, & D. Touretzky, (Eds.) *Advances in Neural Information Processing Systems 3,* (pp. 564 - 571). San Mateo, CA: Morgan & Kaufman.

Davidson, R.J., Ekman, P., Saron, C. Senulis, J., & Friesen, W.V. (1990). Emotional expression and brain physiology I. Approach/withdrawal and cerebral asymmetry. *Journal of Personality and Social Psychology* 58, 330-341.

Ekman, P. (Ed.) (1982a). *Emotion in the Human Face, Second Edition.* New York: Cambridge University Press.

Ekman, P. (1982b) Methods for measuring facial action. In Scherer, K.R. and Ekman, P. (Eds.) *Handbook of Methods in Nonverbal Behavior Research.* New York, Cambridge University Press, 45-135.

Ekman, P. (1984). Expression and the nature of emotion. In K. Scherer and P. Ekman, Eds.

Approaches to Emotion (pp 319-343). Hillsdale, NJ: Lawrence Erlbaum.

Ekman, P. (1989). The argument and evidence about universals in facial expressions of emotion. In H. Wagner & A. Manstead, Eds. *Handbook of Social Psychophysiology, 58* 342-353.

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

Ekman, P., Huang, T., Sejnowski, T., & Hager, J. (1992). *Final Report to NSF of the Planning Workshop on Facial Expression Understanding.* Available from UCSF, HIL-0984, San Francisco, CA 94143. Online version: http://alt.xmission.com/ nirc1/Research/research.html

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

Ekman, P. & Oster, H. (1979) Facial expressions of emotion. *Annual Review of Psychology, 30*, 527-554.

Ekman, P. & Rosenberg, E.L. (1997) *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS).* New York: Oxford University Press.

Essa, I., & Pentland A. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7), 757 - 763.

Hager, J., & Ekman, P., (1995). The essential behavioral science of the face and gesture that computer scientists need to know. In M. Bichsel (Ed.) *Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition* (pp. 7-11). Available from University of Zurich, Department of Computer Science, Winterhurrerstrasse 190, CH-8057. Online version: http://alt.xmission.com/ nirc1/Research/research.html

Hancock, P.J.B., Burton, A.M., & Bruce, V. (1996). Face processing: human perception and principal components analysis. *Memory and Cognition 24,* 26-40.

Hasselmo M.E., Rolls E.T., & Baylis G.C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research*, 32(3), 203-18.

Haykin, S. (1994). Neural Networks: A Comprehensive Foundation. New Jersey: MacMillan.

Himer, W., Schneider, F., Kost, G., Heimann, H. (1991). Computer-based analysis of facial action: A new approach. *Journal of Psychophysiology, 5*(2), 189-195.

Horn, B. & Schunk, B. (1981). Determining optical flow. *Artificial Intelligence 17*, 185 - 203.

Jain, R. Kasturi, R, & Schunk, B. G.(1995). *Machine Vision.* New York: McGraw-Hill.

Kaiser, S., & Wherle, T. (1992). Automated coding of facial behavior in human-computer interactions with FACS. *Journal of Nonverbal Behavior, 16*(2), 65-140.

Kanade, T. (1977). *Computer recognition of human faces.* Basel & Stuttgart: Birkhauser Verlag.

Kinomura S., Kawashima R., Yamada K., Ono S., Itoh M., Yoshioka S., Yamaguchi T.,Matsui H., Miyazawa H., Itoh H., et al. (1994) Functional anatomy of taste perception in the human brain studied with positron emission tomography. *Brain Research 659*(1-2), 263-6.

Lanitis, A., Taylor, C.J., & Cootes, T.F. (1997). Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7), 743-756.

Li, H., Roivainen, P., & Forcheimer, R. (1997). 3-D motion estimation in model-based facial image coding. *IEEE transactions on Pattern Analysis and Machine Intelligence* 15(6), 545-55.

Lien, J-J. (1998). Automatic recognition of facial expressions using hidden markov models and estimation of expression intensity. Doctoral Dissertation, University of Pittsburgh, School of Engineering.

Mase, K. (1991): Recognition of facial expression from optical flow. *IEICE Transactions E, 74*(10), 3474-3483.

Morris J.S., Frith C.D., Perrett D.I., Rowland D., Young A.W., Calder A.J., Dolan R.J., (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature 383*(6603), 812-5.

Movellan, J.R. (1995) Visual Speech Recognition with Stochastic Networks. In G. Tesauro, D. Touretzky, & T. Leen (Eds.) *Advances in Neural Information Processing Systems 7*, Cambridge, MA: MIT Press, 851-858.

O'Toole, A.J., Deffenbacher, K.A., Valentin, D., & Abdi, H. (1994). Structural aspects of face recognition and the other race effect. *Memory and Cognition,* 22(2) 208-224.

Padgett, C., Cottrell, G., (1997). Representing face images for emotion classification. In M. Mozer M. Jordan, & T. Petsche (Eds.) *Advances in Neural Information Processing Systems 9*, Cambridge, MA: MIT Press, 894-900.

Phillips, M.L., Young, A.W., Senior, C., Brammer, M. Andrews, C., Calder, A.J., Bullmore, E.T., Perrett, D.I., Rowland, D., Williams, S.C.R., Gray, A.J., & David, A.S. (m1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature 389*, 495-498.

Rosenblum, M., Yacoob, Y., & Davis, L. (1996). Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks, 7*(5), 1121-1138.

Terzopoulos, D., & Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 15(6), 569-79.

Tranel, D. Damasio, A., Damasio, H. (1988). Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology*, 38(5), 690-696.

Tukey, J.W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics 29*, 614.

Turk, M., & Pentland, A. (1991): Eigenfaces for Recognition. *Journal of Cognitive Neuroscience, 3*(1), 71-86.

Valentin, D., Abdi, H., O'Toole, A.J., Cottrell, G.W. (1994). Connectionist models of face processing: a survey. *Pattern Recognition, 27*(9), 1209-30.

Vetter, T., & Poggio, T. (1997). Linear object classes and image synthesis from a single example image. *IEEE transactions on Pattern Analysis and Machine Intelligence* 15(6), 733-741.

Wallbott, H. (1992). Effects of distortion of spatial and temporal resolution of video stimuli on emotion attributions. *Journal of Nonverbal Behavior, 16*(1), 5-20.

Yacoob, Y., & Davis, L. (1994): Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence 18*(6), 636-642.

Yaxley S., Rolls E.T., Sienkiewicz Z.J. (1988). The responsiveness of neurons in the insular gustatory cortex of the macaque monkey is independent of hunger. *Physiology and Behavior 42*(3), 223-9.

Zajonc, R.B. (1984). The interaction of affect and cognition. In K. Scherer & P. Ekman, (Eds.) *Approaches to Emotion* (pp. 239-246). Hillsdale, NJ: Lawrence Erlbaum.

Action	Responses											
	AU1		AU2		AU4		AU5		AU6		AU7	
	Nv	Ex	Nv	Ex	Nv	Ex	Nv	Ex	Nv	Ex	Nv	Ex
AU1	**.84**	**.99**	.08	.00	.03	.00	.02	.00	.02	.00	.02	.01
AU2	.12	.04	**.83**	**.93**	.00	.00	.03	.00	.01	.00	.00	.02
AU4	.03	.00	.03	.01	**.88**	**.96**	.01	.00	.02	.00	.03	.02
AU5	.09	.00	.20	.01	.00	.01	**.64**	**.98**	.03	.00	.03	.01
AU6	.04	.00	.03	.01	.04	.00	.00	.00	**.55**	**.41**	.34	.58
AU7	.00	.00	.04	.00	.05	.02	.00	.00	.26	.09	**.65**	**.89**

Table 1: **Confusion Matrix for Naive and Expert Human Subjects.** Rows give the percent occurrence of each response for a given action. Nv: Naive subject data, Ex: Expert subject data.

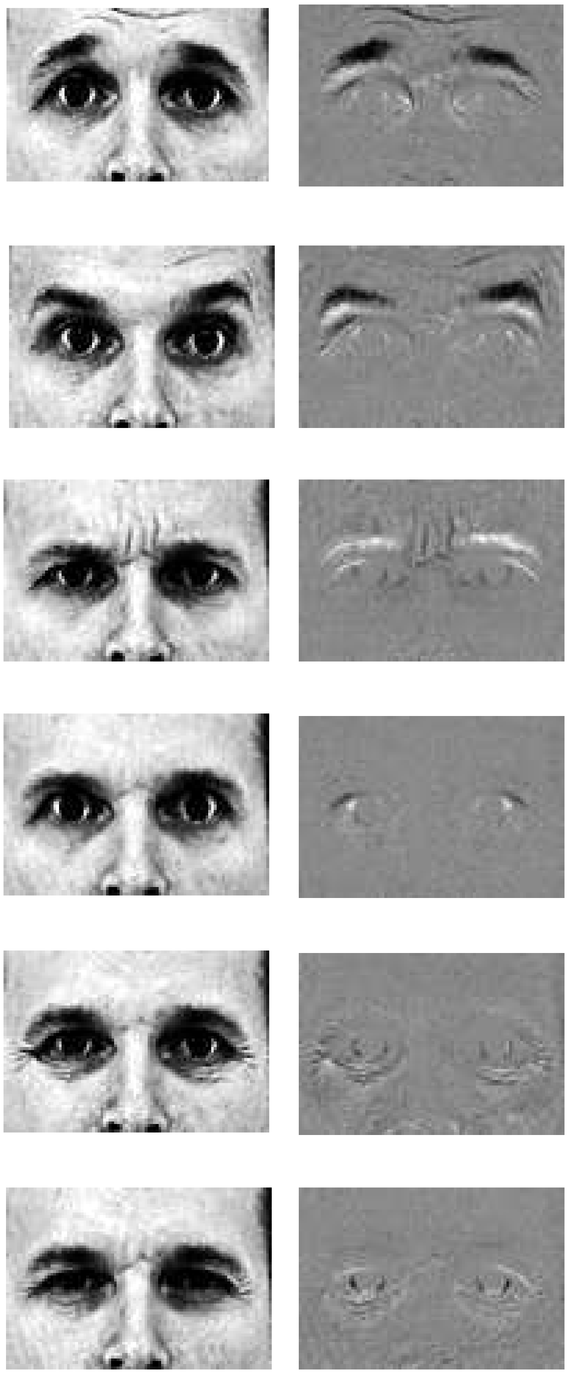
Action	**Responses**					
	AU1	AU2	AU4	AU5	AU6	AU7
	Hol Mt Ft Hyb	Hol Mt Ft Hyb	Hol Mt Ft Hyb	Hol Mt Ft Hyb	Hol Mt Ft Hyb	Hol Mt Ft Hyb
AU1	**.58 .20 .50 .57**	.19 .31 .04 .17	.00 .00 .29 .01	.10 .33 .14 .08	.03 .00 .00 .00	.10 .15 .02 .18
AU2	.12 .02 .10 .10	**.83 .94 .36 .85**	.01 .00 .04 .00	.01 .02 .41 .00	.00 .00 .00 .00	.03 .02 .09 .05
AU4	.00 .00 .08 .00	.00 .01 .01 .00	**.96 .97 .54 .99**	.00 .00 .26 .00	.06 .00 .00 .00	.04 .02 .10 .01
AU5	.01 .00 .07 .00	.15 .00 .35 .00	.00 .00 .10 .00	**.98 1.0 .74 1.0**	.00 .00 .00 .00	.00 .00 .06 .00
AU6	.00 .00 .00 .00	.00 .00 .02 .00	.00 .16 .00 .02	.06 .04 .20 .02	**.56 .40 .38 .74**	.38 .40 .40 .22
AU7	.00 .00 .06 .00	.00 .00 .03 .00	.01 .00 .06 .01	.00 .02 .21 .01	.00 .03 .03 .00	**.99 .94 .62 .98**

Table 2: **Confusion Matrix for the automated classifiers.** Rows give the percent occurrence of each response for a given action. Hol: Holistic, Mt: Motion, Ft: Feature, Hyb: Hybrid.

	Expert	Holistic	Motion	Feature	Hybrid
Naive	.58***	.36**	.18*	.05	.19*
Expert		.66***	.36**	.23**	.36**
Holistic			.70***	.17*	.82***
Motion				.09	.69***
Feature					.07

Table 3: **Action confusion correlations.** Entries are squared correlation coefficients. Stars indicate statically significant correlation based on a t-test with 28 degrees of freedom at the .05* level, .01**, and .001***.

Figure 1: Example action sequence from the database. The example shows a subject performing AU1 starting from a neutral expression and ending with a high magnitude action.



AU 1. Inner Brow Raiser
Central frontalis muscle contracts.
Inner corners of brows raise up,
In some subjects, brows make inverted V.
Wrinkles in center of forehead, slight U-shape.

AU 2. Outer Brow Raiser
Lateral frontalis muscle contracts.
Pulls lateral portion of the eyebrows upward, producing
Wrinkles on sides of forehead.
Lateral portion of eye cover fold stretches upwards.

AU 4. Brow Lower
Brows lower and / or draw together.
Produces deep vertical and / or 45 degree wrinkles between brows.
Pushes eye cover fold downwards and may narrow eye opening.
May also produce horizontal wrinkles at root of nose and oblique bulge over medial corner of brow.

AU 5. Upper Lid Raiser
Upper lid retracts, widens eye opening.
Decreases visibility of upper lid.
Increases exposure of sclera above the iris, and on the sides.
Eyeballs appear to protrude.

AU 6. Cheek Raiser
Orbital muscles around the eye contract
May cause crows feet to appear or deepen.
Makes infraorbital furrow evident - straight or crescent shape.
Causes bagging or wrinkling of the skin below the eye.
Changes under eye are lower on face than AU 7

AU 7. Lid Tightener
Tightens eyelids, narrowing the eye opening.
More apparent in lower lid. Lower lid covers more sclera and iris.
Lower lid straightens or medial portion slightly inverts.
Bulge appears in lower lid and lower eyelid furrow may deepen
Can produce single "crow's toe," not crows feet as in AU 6.

Figure 2: Examples of the six actions used in this study. From left to right: cropped image of the action at highest magnitude; difference image obtained by subtracting the neutral image (frame 1 of the sequence); Action unit number; Action unit description adapted from Ekman & Friesen (1978).

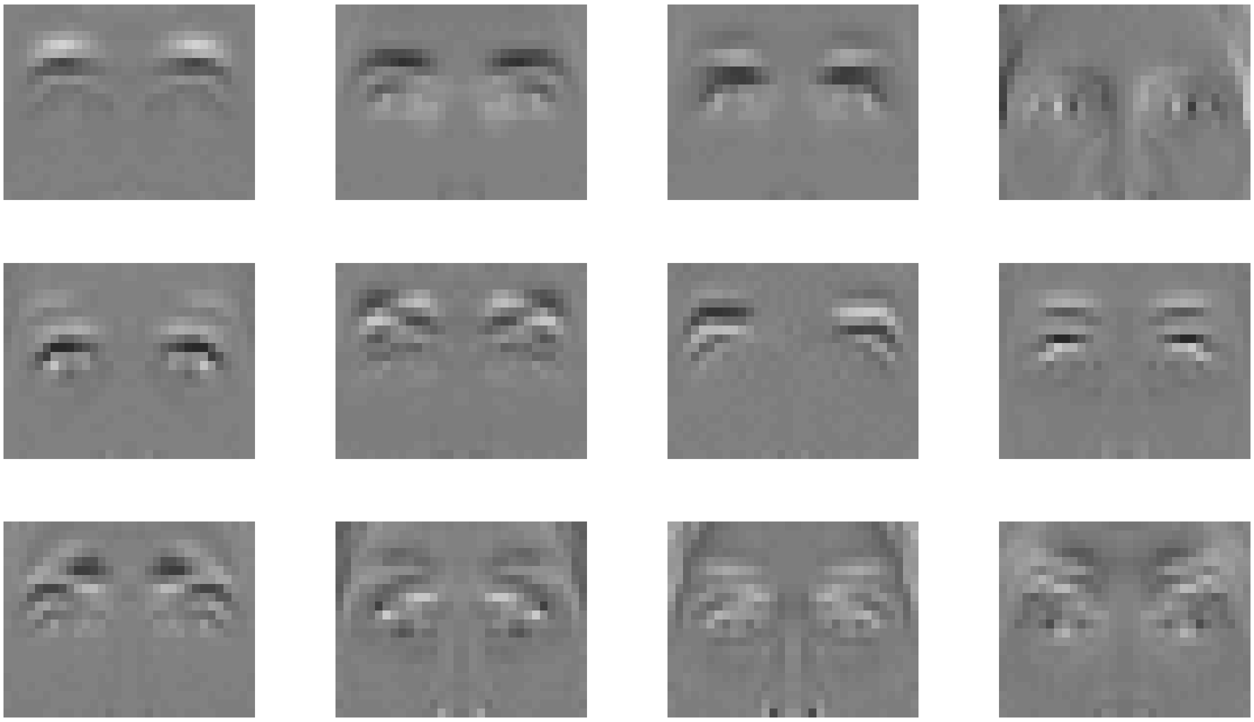
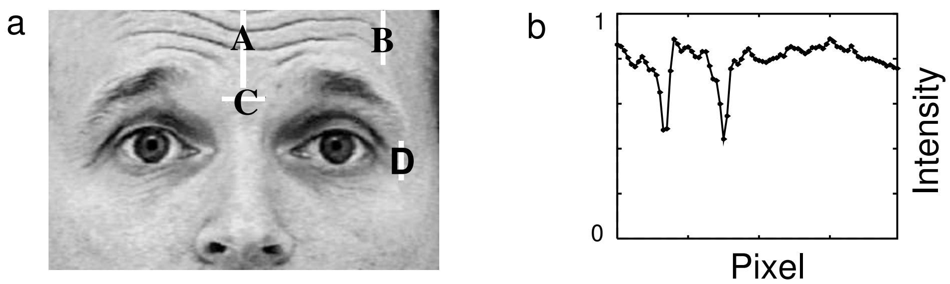


Figure 3: First 12 principal components of the dataset of difference images, ordered left to right, top to bottom. The first component appears to code for vertical brow position. The sixth component axis appears to differentiate between AU1, raising the inner corners of the brow, and AU2, raising the lateral portions of the brows. Component 7 appears to be an axis of left-right asymmetry in the lateral brow movement, and component 5 appears to be an eye opening axis.



c

$$P = \sum (I_i - I_{i-1})^2 \quad (1)$$

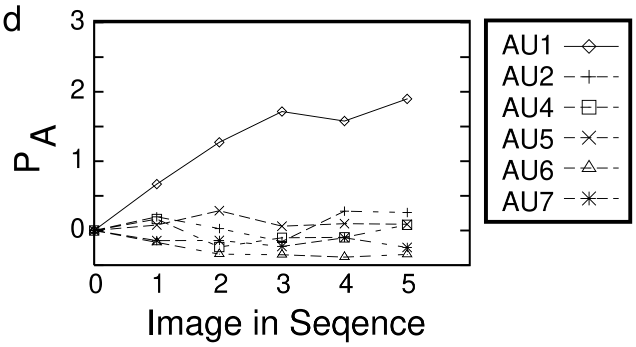


Figure 4: a) Wrinkling was measured at four image locations, A-D. b) Smoothed pixel intensities along the line labeled A. c) The wrinkle measure, P. I_i is the intensity of the ith pixel of the segment. Pixel differences approximate the derivative (Jain, Kasturi, & Schunk, 1995). d) P measured at image location A for one subject performing each of the six actions.

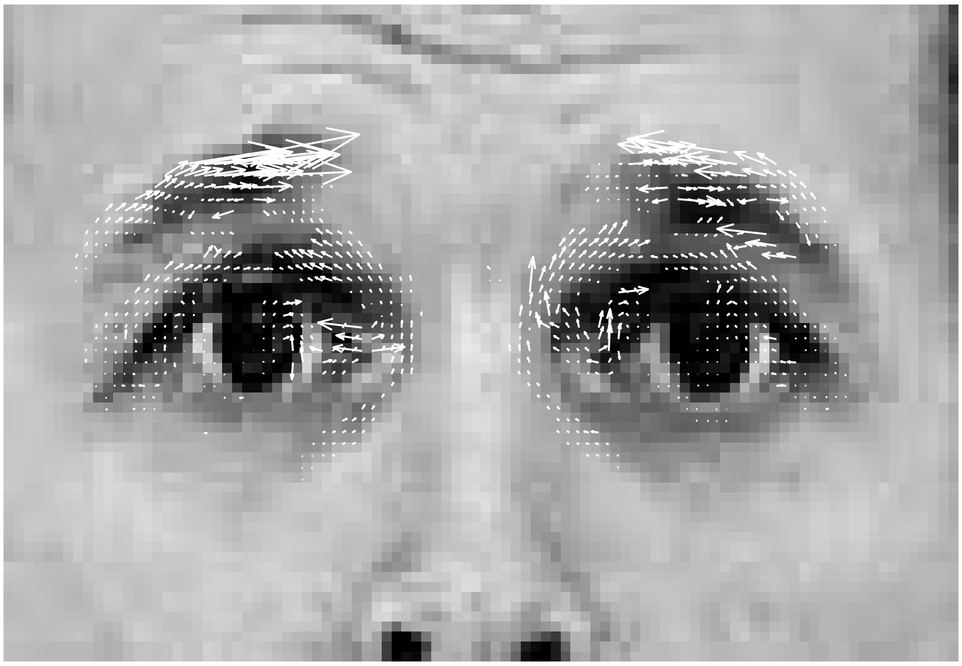


Figure 5: Example flow field of a subject performing AU1, inner brow raiser. The flow vector at each image location is plotted as an arrow with length proportional to the local estimate of velocity.

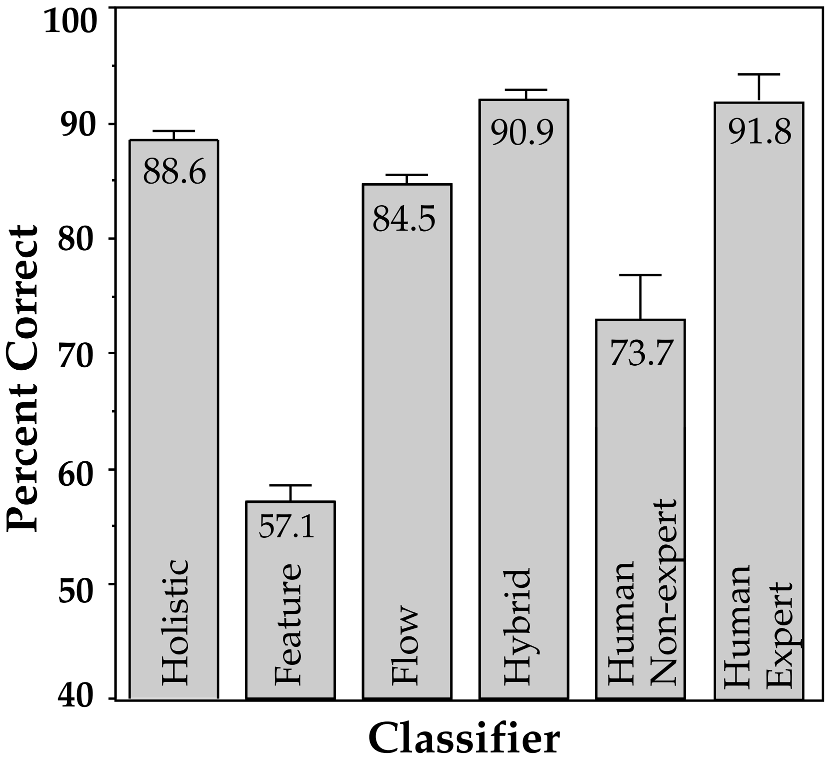


Figure 6: Performance comparisons for generalization to novel subjects. Values are percent correct across all test images. Error bars are one standard deviation of the estimate of the success rate in a Bernoulli distribution. Human results were prorated by action and action magnitude to match the proportions in the complete image set.