Towards Automatic Recognition of Spontaneous Facial Actions

To appear in P. Ekman (Ed.), What the Face Reveals, 2nd Ed., Oxford University Press.

MARIAN STEWART BARTLETT, JAVIER R. MOVELLAN, GWEN LITTLEWORT, BJORN BRAATHEN, MARK G. FRANK & TERRENCE J. SEJNOWSKI

Charles Darwin (1872/1998) was the first to fully recognize that facial expression is one of the most powerful and immediate means for human beings to communicate their emotions, intentions, and opinions to each other. In addition to providing information about affective state, facial expressions also provide information about cognitive state, such as interest, boredom, confusion, and stress, and conversational signals with information about speech emphasis and syntax. Facial expressions also contain information about whether an expression of emotion is posed or felt (Ekman, 2001; Frank, Ekman, & Friesen, 1993). In order to objectively measure the richness and complexity of facial expressions, behavioral scientists have found it necessary to develop objective coding standards. The Facial Action Coding System (FACS) from Ekman and Friesen (1978) is arguably the most comprehensive and influential of such standards. FACS is based on the anatomy of the human face, and codes expressions in terms of component movements, called "action units" (AUs). Ekman and Friesen defined 46 AUs to describe each independent movement of the face. FACS measures all visible facial muscle movements, including head and eye movements, and not just those presumed to be related to emotion. When learning FACS, a coder is trained to identify the characteristic pattern of bulges, wrinkles, and movements for each facial AU. The AUs approximate individual facial muscle movements but there is not always a 1:1 correspondence.

FACS has been used to verify the physiological presence of emotion in a number of studies, with high (over 75%) agreement (e.g., Ekman, Friesen, & Ancoli, 1980; Ekman, Levenson, & Friesen, 1983; Ekman, Davidson, & Friesen, 1990; Levenson, Ekman, & Friesen, 1990; Ekman, Friesen, & O'Sullivan, 1988). Because it is comprehensive, FACS also allows for the discovery of new patterns related to emotional or situational states. For example, using FACS Ekman et al (1990) and Davidson et al (1990) found that smiles which featured both orbicularis oculi (AU6), as well as zygomatic major action (AU12), were correlated with self-reports of enjoyment, as well as different patterns of brain activity, whereas smiles that featured only zygomatic major (AU12) were not. Subsequent research demonstrated that the presence of smiles that involve the orbicularis oculi (hereafter "enjoyment smiles") on the part of a person who has

survived the death of their romantic partner predicts successful coping with that traumatic loss (Bonnano & Keltner, 1997). Other work has shown a similar pattern. For example, infants show enjoyment smiles to the presence of their mothers, but not to strangers (Fox & Davidson, 1988). Mothers do not show as many enjoyment smiles to their difficult children compared to their non-difficult children (Bugental, 1986). Research based upon FACS has also shown that facial expressions can predict the onset and remission of depression, schizophrenia, and other psychopathology (Ekman & Rosenberg, 1997), can discriminate suicidally from non-suicidally depressed patients (Heller & Haynal, 1994), and can predict transient myocardial ischemia in coronary patients (Rosenberg et al., 2001). FACS has also been able to identify patterns of facial activity involved in alcohol intoxication that observers not trained in FACS failed to note (Sayette, Smith, Breiner, & Wilson, 1992).

Although FACS is an ideal system for the behavioral analysis of facial action patterns, the process of applying FACS to videotaped behavior is currently done by hand and has been identified as one of the main obstacles to doing research on emotion (Frank, 2002, Ekman et al, 1993). FACS coding is currently performed by trained experts who make perceptual judgments of video sequences, often frame by frame. It requires approximately 100 hours to train a person to make these judgments reliably and pass a standardized test for reliability. It then typically takes over two hours to code comprehensively one minute of video. Furthermore, although humans can be trained to code reliably the morphology of facial expressions (which muscles are active) it is very difficult for them to code the dynamics of the expression (the activation and movement patterns of the muscles as a function of time). There is good evidence suggesting that such expression dynamics, not just morphology, may provide important information (Ekman & Friesen, 1982). For example, spontaneous expressions have a fast and smooth onset, with distinct facial actions peaking simultaneously, whereas posed expressions tend to have slow and jerky onsets, and the actions typically do not peak simultaneously (Frank, Ekman, & Friesen, 1993).

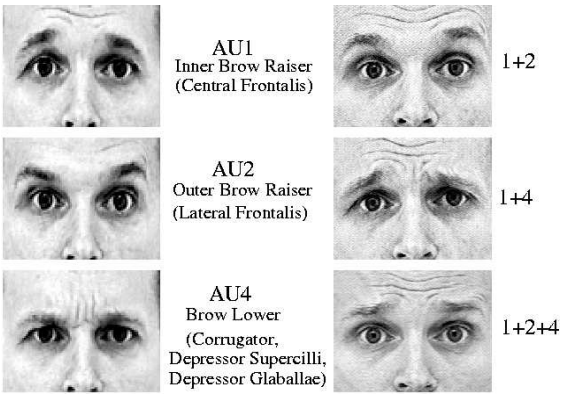


Figure 1: The Facial Action Coding System decomposes facial expressions into component actions. The three individual brow region actions and selected combinations are illustrated. When subjects pose fear they often perform 1+2 (top right), whereas spontaneous fear reliably elicits 1+2+4 (bottom right) (Ekman, 2001).

Within the past decade, significant advances in computer vision open up the possibility of automatic coding of facial expressions at the level of detail required for such behavioral studies. Automated systems would have a tremendous impact on basic research by making facial expression measurement more accessible as a behavioral measure, and by providing data on the dynamics of facial behavior at a resolution that was previously unavailable. Such systems would also lay the foundations for computers that can understand this critical aspect of human communication. Computer systems with this capability have a wide range of applications in basic and applied research areas, including man-machine communication, security, law enforcement, psychiatry, education, and telecommunications.

A number of ground breaking systems have appeared in the computer vision literature for facial expression recognition which use a wide variety of approaches, including optic flow (Mase, 1991;Yacoob & Davis, 1996; Rosenblum, Yacoob, & Davis, 1996; Essa & Pentland, 1997), tracking of high-level features (Tian, Kanade, & Cohn, 2001; Lien, Kanade, Cohn, & Li, 2000) methods that match images to physical models of the facial skin and musculature (Mase 1991; Terzopoulus & Waters, 1993; Li, Riovainen, & Forscheimer, 1993; Essa & Pentland, 1997), methods based on statistical learning of images (Cottrell & Metcalfe, 1991; Padgett & Cottrell, 1997; Lanitis, Taylor, & Cootes, 1997; Bartlett et al., 2000) and methods based on biologically inspired models of human vision (Zhang, Lyons, Schuster, & Akamatsu, 1998; Bartlett, 2001, Bartlett, Movellan, & Sejnowski, 2002). See Pantic (2000b) for a review.

Much of the early work on computer vision applied to facial expressions focused on recognizing a few prototypical expressions of emotion produced on command (e.g., "smile"). More recently there has been an emergence of groups that analyze facial expressions into elementary components. For example Essa and Pentland (1997) and Yacoob and Davis (1996) proposed methods to analyze expressions using an animation-style coding system inspired by FACS. Eric Petajan's group has also worked for many years on methods for automatic coding of facial expressions in the style of MPEG4 which codes movement of a set of facial feature points (Doenges, Lavagetto, Osterman, Pandzic and Petajan, 1997). While coding standards like MPEG4 are useful for animating facial avatars, behavioral research may require more comprehensive information. For example, MPEG4 does not encode some behaviorally relevant movements such as the contraction of the orbicularis oculi, which differentiates spontaneous from posed smiles (Ekman, 2001). It also does not measure changes in surface texture such as wrinkles, bulges, and shape changes that are critical for the definition of action units in the FACS system. For example, the vertical wrinkles and bulges between the brows are important for distinguishing AU 1 alone from AU 1+4 (see Figure 1b), both of which entail upward movement of the brows, but which can have different behavioral implications.

We present here an approach for developing a fully automatic FACS coding system. The approach uses state of the art machine learning techniques that can be applied to recognition of any facial action. The techniques were tested on a small sample of facial actions, but can be readily applied to recognition of other facial actions given a sample of images on which to train the system. We are presently collaborating with Mark Frank to collect more training data (see Afterword.) In this paper we show preliminary results for I. Recognition of posed facial actions in

controlled conditions, and II. Recognition of spontaneous facial actions in freely behaving subjects.

Two other groups have focused on automatic FACS recognition as a tool for behavioral research. One team, lead by Jeff Cohn and Takeo Kanade, present an approach based on traditional computer vision techniques such as using edge detection to extract contour-based image features and motion tracking of those features using optic flow. A comparative analysis of our approaches is available in (Bartlett et al, 2001; Cohn et al., 2001). Pantic & Rothcrantz (2000a) use robust facial feature detection followed by an expert system to infer facial actions from the geometry of the facial features. The approach presented here measures changes in facial texture that include not only changes in position of feature points, but also higher resolution changes in image texture such as those created by wrinkles, bulges, and changes in feature shapes. We explore methods that merge machine learning and biologically inspired models of human vision. Our approach differs from other groups in that instead of designing special purpose image features for each facial action, we explore general purpose learning mechanisms that can be applied to recognition of any facial action.

Study I: Automatic FACS coding of posed facial actions, controlled conditions

A database of directed facial actions was collected by Paul Ekman and Joe Hager at the University of California, San Francisco. The full database consists of 1100 image sequences containing over 150 distinct actions and action combinations, and 24 subjects. These images were collected in a constrained environment. Subjects deliberately faced the camera and held their heads as still as possible. Each sequence contained 7 frames, beginning with a neutral expression and ending with the action unit peak. For this investigation, we used 111 sequences from 20 subjects and attempted to classify 12 actions: 6 upper face actions (Aus 1, 2, 4, 5, 6, and 7) and 6 lower face actions (Aus 9, 10, 16, 17, 18, 20). Upper and lower-face actions were analyzed separately. A sample of facial actions from this database is shown in Figure 1b.

We developed and compared techniques for automatically recognizing these facial actions by computer (Bartlett et al., 1996; Bartlett, Hager, Ekman, & Sejnowski, 1999; Donato, Bartlett, Hager, Ekman, & Sejnowski, 1999; Bartlett, Donato, Hager, Ekman, & Sejnowski, 2000). Our work focused on comparing the effectiveness of different image representations, or feature extraction methods, for facial action recognition. We compared image filters derived from supervised and unsupervised machine learning techniques. These data-driven filters were compared to Gabor filter banks, which closely model the response transfer function of simple cells in primary visual cortex. In addition, we also examined motion representations based on optic flow, and an explicit feature-extraction technique that measured facial wrinkles in specified locations (Bartlett et. al. 1999; Donato et al. 1999). These techniques are briefly reviewed here. More information is available in the journal papers cited above, and in Bartlett (2001).

Adaptive methods

In contrast to more traditional approaches to image analysis in which the relevant structure is decided by the human user and measured using hand-crafted techniques, adaptive methods learn about the image structure directly from the image ensemble. We draw upon principles of machine learning and information theory to adapt processing to the immediate task environment. Adaptive methods have proven highly successful for tasks such as recognizing facial identity (e.g. Brunelli & Poggio, 1993; Turk & Pentland, 1991; Penev & Atick, 1996; Belhumeur et al., 1997; Bartlett, Movellan, & Sejnowski, 2002; see Bartlett, 2001 for a review), and can be applied to recognizing any expression dimension given a set of training images.

We compared four techniques for developing image filters adapted to the statistical structure of face images. (See Figure 2.) The techniques were Principal Component Analysis (PCA), often termed Eigenfaces (Turk & Pentland 1991), Local Feature Analysis (LFA) (Penev & Atick, 1996), Fisher's linear discriminants (FLD), and Independent Component Analysis (ICA). Except for FLD, all of these techniques are unsupervised; image representations are developed without knowledge of the underlying action unit categories. Principal component analysis, Local Feature Analysis and Fisher discriminant analysis are a function of the pixel by pixel covariance matrix and thus insensitive to higher-order statistical structure. Independent component analysis is a generalization of PCA that learns the high-order relations between image pixels, not just pair-wise linear dependencies. We employed a learning algorithm for ICA developed in Terry Sejnowski's laboratory based on the principle of optimal information transfer between neurons (Bell & Sejnowski, 1995; Bartlett, Movellan, & Sejnowski, 2002).

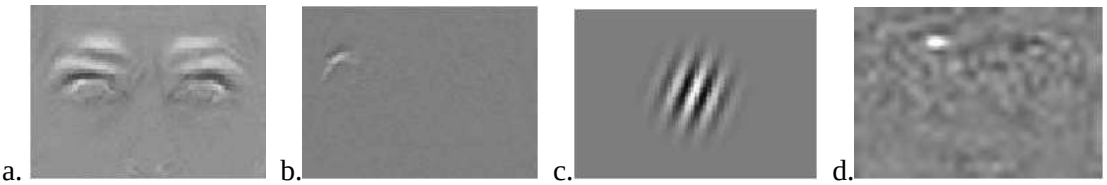


Figure 2. Sample image filters for the upper face. a. Eigenface (PCA). b. Independent component analysis (ICA) c. Gabor. d. Local Feature Analysis (LFA).

Predefined image features

Gabor wavelets

An alternative to the adaptive methods described above are wavelet decompositions based on predefined families of image kernels. We employed Gabor kernels, which are 2-D sine waves modulated by a Gaussian. Gabor kernels model the response functions of cells in the primate visual cortex (Daugman, 1988), and have proven successful as a basis for recognizing facial identity in images (Lades et al., 1993).

Explicit Feature Measures

A more traditional approach to computer vision is to apply hand-crafted image features explicitly designed to measure components of the image that the engineer has decided are relevant. We applied a method developed by Jan Larson (Bartlett et. al., 1996) for measuring changes in facial wrinkling and eye opening. Facial wrinkling was measured by the sum of the squared derivatives of the image pixels along line segments in 4 facial regions predicted to contain wrinkles due to the facial actions in question. Eye opening was measured as the area of visible sclera. Changes in wrinkling or eye opening were measured by subtracting baseline measured for the neutral image. See Bartlett et al. (1999) for more information on this technique.

Optic Flow

The majority of the work on automatic facial expression recognition has focused on facial motion analysis through optic flow estimation. Here, optic flow fields were calculated by employing a correlation-based technique developed by Singh (1992). Optic flow fields were classified by template matching. (See Donato et al., 1999, for more information).

Classification Procedure

The face was located in the first frame in each sequence using the centers of the eyes and mouth. These coordinates were obtained manually by a mouse click. The coordinates from Frame 1 were used to register the subsequent frames in the sequence. The aspect ratios of the faces were warped so that the eye and mouth centers coincided across all images. The three coordinates were then used to rotate the eyes to horizontal, scale, and finally crop a window of 60 x 90 pixels containing the upper or lower face. To control for variations in lighting, logistic thresholding and luminance scaling was performed (Movellan, 1995). Difference images were obtained by subtracting the neutral expression in the first image of each sequence from the subsequent images in the sequence. Individual frames of each action unit sequence were otherwise analyzed separately, with the exception of optic flow which analyzed three consecutive frames.

Each image analysis algorithm produced a feature vector f. We employed a simple nearest neighbor classifier in which the similarity of a training feature vector f_t and a novel feature vector f_n was measured as the cosine of the angle between them. The test vector was assigned the class label of the training vector for which the cosine was highest. We also explored template matching, where the templates were the mean feature vectors for each class. Generalization to novel faces was evaluated using leave-one-out cross-validation.

Human Subject Comparisons

The performance of human subjects provided benchmarks for the performances of the automated systems. Naïve subjects benchmarked the difficulty of the visual classification task. The agreement rates of FACS experts benchmarked how close we were to the goal of replacing expert human coders with an automated system. Naïve subjects were 10 adult volunteers with no prior knowledge of facial expression measurement. Upper and lower facial actions were tested separately. Subjects were provided with a guide sheet which gave an example of each of the 6 lower or upper facial actions along with written descriptions from Ekman & Friesen (1978). Each subject was given a training session in which the facial actions were described and demonstrated, and visual cues were pointed out in the example images. The subject kept the guide sheet as a reference during the task. Face images were preprocessed identically to how they had been for the automated systems, and then printed using a high resolution laser printer. Face images were presented in pairs, with the neutral image and the test image presented side by side. Subjects made a 6-alternative forced choice on 93 pairs of upper face and 93 pairs of lower face actions. Expert subjects were 4 certified FACS coders. Expert subjects were not given additional training or a guide sheet.

Overall Findings

Image decomposition with gray-level image filters outperformed explicit extraction of facial wrinkles or motion flow fields. Best performance was obtained with the Gabor wavelet decomposition and independent component analysis, each of which gave 96% accuracy for classifying the 12 facial actions (see Table 1). This performance equaled the agreement rates of expert human subjects on this set of images. The Gabor and ICA representations were both sensitive to high-order dependencies among the pixels (Field, 1994; Simoncelli, 1997), and have relationships to visual cortical neurons (Daugman, 1988; Bell & Sejnowski, 1997). See (Bartlett, 2001) for a more detailed discussion. We also obtained evidence that high spatial frequencies are important for classifying facial actions. Classification with the three highest frequencies of the Gabor representation (15,18,21 cycles/face) was 93% compared to 84% with the three lowest frequencies (9,12,15 cycles/face).

Computational Analysis	Eigenfaces	79.3 $\pm$4
	Local Feature Analysis	81.1 $\pm$4
	Independent Component Analysis	95.5 $\pm$2
	Fisher's Linear Discriminant	75.7 $\pm$4
	Gabor Wavelet Decomposition	95.5 $\pm$2
	Optic Flow	85.6 $\pm$3
	Explicit Features (wrinkles)	57.1 $\pm$6
Human Subjects	Naïve	77.9 $\pm$3
	Expert	94.1 $\pm$2

Table 1: Summary of results for recognition of directed facial actions. Performance is for novel subjects on frame 5. Values are percent agreement with FACS labels in the database.

We also investigated combining multiple sources of information in a single classifier. Combining the wrinkle measurements with PCA in a three layer perceptron resulted in a 0.3 percentage point improvement in performance over PCA alone (Bartlett et al., 1999).

In addition, we trained a dedicated system to distinguish felt from unfelt smiles (Littlewort-Ford, Bartlett, & Movellan, 2001) based on the findings of Ekman, Friesen, and O'Sullivan (1988) that felt smiles include the contraction of the orbicularis oculi. This system was trained on two FACS-coded databases of images, the DFAT-504 and the Ekman-Hager databases. There were 157 examples of smiles scored as containing both AU 12 (zygomatic major) and AU 6 (orbicularis oculi) and 72 examples of smiles scored as containing 12 but not AU 6. This system obtained 87% correct discrimination of felt from unfelt smiles. This is encouraging given that non-expert humans detected AU 6 about 50% of the time and false alarmed about 25% of the time on a 6-alternative forced choice (Bartlett et al., 1999).

Study II: Automatic FACS coding of spontaneous facial expressions[1]

Prior to 2000, work in automatic facial expression recognition was based on datasets of posed expressions collected under controlled conditions with subjects deliberately facing the camera at all times. In 2000-2001 our group at UCSD, along with the Cohn/Kanade group at CMU, undertook the first attempt that we know of to automate FACS coding of spontaneous facial expressions in freely behaving individuals (Bartlett et. al, 2001; Cohn et al. 2001). Extending these systems to spontaneous facial behavior was a critical step forward towards development of tools with practical applications in behavioral research.

Spontaneous facial expressions differ substantially from posed expressions, similar to how continuous, spontaneous speech differs from isolated words produced on command. Spontaneous facial expressions are mediated by a distinct neural pathway from posed expressions. The pyramidal motor system, originating in the cortical motor strip, drives voluntary facial actions, whereas involuntary, emotional facial expressions originate subcortically and involve the basal ganglia, limbic system, and the cingulate motor area (e.g. Rinn, 1984). Psychophysical work has shown that spontaneous facial expressions differ from posed expressions in a number of ways (Ekman, 2001). Subjects often contract different facial muscles when asked to pose an emotion such as fear versus when they are actually experiencing fear. (See Figure 1.) In addition, the dynamics are different. Spontaneous expressions have a fast and smooth onset, with apex coordination, in which muscle contractions in different parts of the face peak at the same time. In posed expressions, the onset tends to be slow and jerky, and the muscle contractions typically do not peak simultaneously.

The goal of this study was to classify facial actions in twenty subjects who participated in a high stakes mock crime experiment previously conducted by Mark Frank and Paul Ekman (Frank and Ekman, under review). The results were evaluated by a team of computer vision experts (Yaser Yacoob, Pietro Perona) and behavioral experts (Paul Ekman, Mark Frank). These

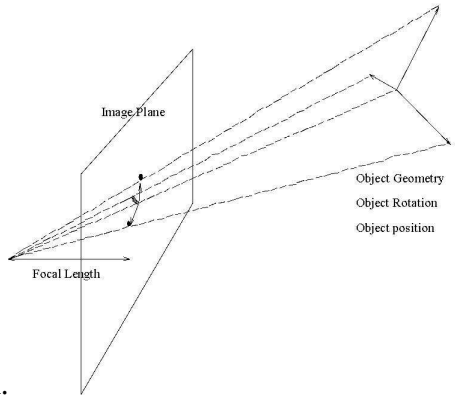
experts produced a report identifying the feasibility of this technology and the steps necessary for future progress.

Factorizing rigid head motion from nonrigid facial deformations

The most difficult technical challenge that came with spontaneous behavior was the presence of out-of-plane rotations due to the fact that people often nod or turn their head as they communicate with others. Our approach to expression recognition is based on statistical methods applied directly to filter bank image representations. While in principle such methods may be able to learn the invariances underlying out-of-plane rotations, the amount of data needed to learn such invariances was not available to us. Instead, we addressed this issue by means of deformable 3D face models. We fit 3D face models to the image plane, texture those models using the original image frame, then rotate the model to frontal views, warp it to a canonical face geometry, and then render the model back into the image plane. (See Figures 3-5.) This allowed us to factor out image variation due to rigid head rotations from variations due to nonrigid face deformations. The rigid transformations were encoded by the rotation and translation parameters of the 3D model. These parameters are retained for analysis of the relation of rigid head dynamics to emotional and cognitive state.

Since our goal was to explore the use of 3D models to handle out-of-plane rotations for expression recognition, we first tested the system using hand-labeling to give the position of 8 facial landmarks. The average deviation between human coders was 1/5 of an iris. We are currently obtaining similar precision using automatic feature detectors (See Afterword).

a.



b.

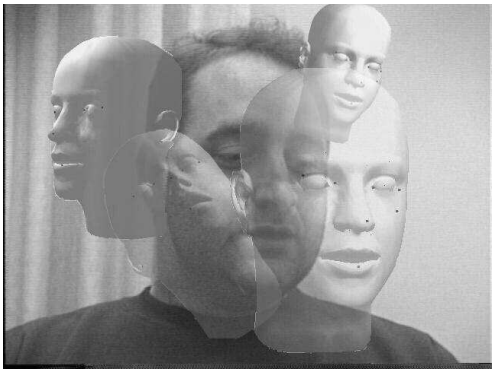


Figure 3: Head pose estimation. a. First camera parameters and face geometry are jointly estimated using an iterative least squares technique b. Next head pose is estimated in each frame using stochastic particle filtering. Each particle is a head model at a particular orientation and scale.

When landmark positions in the image plane are known, the problem of 3D pose estimation is relatively easy to solve. We begin with a canonical wire-mesh face model and adapt it to the face of a particular individual by using 30 image frames in which 8 facial features have been labeled by hand. Using an iterative least squares triangulation technique, we jointly estimate camera parameters and the 3D coordinates of these 8 features. A scattered data interpolation technique is then used to modify the canonical 3D face model so that it fits the 8 feature positions (Pighin et al., 1998). Once camera parameters and 3D face geometry are known, we used a stochastic particle filtering approach (Kitagawa, 1996) to estimate the most likely rotation and translation parameters of the 3D face model in each video frame. (See Braathen, Bartlett, Littlewort, & Movellan, 2001).

Action unit recognition

Database of spontaneous facial expressions

We employed a dataset of spontaneous facial expressions from freely behaving individuals. The dataset consisted of 300 Gigabytes of 640 x 480 color images, 8 bits per pixels, 60 fields per second, 2:1 interlaced. The video sequences contained out of plane head rotation up to 75 degrees. There were 17 subjects: 3 Asian, 3 African American, and 11 Caucasians. Three subjects wore glasses. The facial behaviors in one minute of video per subject were scored frame by frame by 2 teams experts on the FACS system, one lead by Mark Frank at Rutgers, and another lead by Jeffrey Cohn at U. Pittsburgh.

While the database we used was rather large for current digital video storage standards, in practice the number of spontaneous examples of each action unit in the database was relatively small. Hence, we prototyped the system on the three actions which had the most examples: Blinks (AU 45 in the FACS system) for which we used 168 examples provided by 10 subjects, Brow raises (AU 1+2) for which we had 48 total examples provided by 12 subjects, and Brow lower (AU 4) for which we had 14 total examples provided by 12 subjects. Negative examples for each category consisted of randomly selected sequences matched by subject and sequence length. These three facial actions have relevance to applications such as monitoring of alertness, anxiety, and confusion (Holland 1972, Karson, 1988; Orden, Jung & McKeig, 2000; Ekman, 2001).

The system presented here employs general purpose learning mechanisms that can be applied to recognition of any facial action once sufficient training data is available. There is no need to develop special purpose feature measures to recognize additional facial actions.

Recognition system

An overview of the recognition system is illustrated in Figures 4 and 5. Head pose was estimated in the video sequences using a particle filter with 100 particles. Face images were then warped onto a face model with canonical face geometry, rotated to frontal, and then projected back into the image plane. This alignment was used to define and crop a subregion of the face image containing the eyes and brows. The vertical position of the eyes was 0.67 of the window height.

There were 105 pixels between the eyes and 120 pixels from eyes to mouth. Pixel brightnesses were linearly rescaled to [0,255]. Soft histogram equalization was then performed on the image gray-levels by applying a logistic filter with parameters chosen to match the mean and variance of the gray-levels in the neutral frame (Movellan, 1995).

The resulting images were then convolved with a bank of Gabor kernels at 5 spatial frequencies and 8 orientations. Output magnitudes were normalized to unit length and then downsampled by a factor of 4. The Gabor representations were then channeled to a bank of support vector machines (SVM's). Nonlinear SVM's were trained to recognize facial actions in individual video frames. The training samples for the SVM's were the action peaks as identified by the FACS experts, and negative examples were randomly selected frames matched by subject. Generalization to novel subjects was tested using leave-one-out cross-validation. The SVM output was the margin (distance along the normal to the class partition). Trajectories of SVM outputs for the full video sequence of test subjects were then channeled to hidden Markov models (HMM's). HMMs are probabilistic dynamical models that learn probability distributions of sequences. They are the dominant approach in current speech recognition systems, where the task is to recognize sequences of sounds. HMMs were trained to learn the sequences of SVM outputs typically produced for each AU. One HMM was trained on a single AU unit and thus that HMM can be considered as an expert for that AU. A similar approach is used in speech recognition where each HMM becomes an expert on a given phoneme. At test time a new sequence was presented and fed to each HMM to get an estimate of the likelihood of each sequence given each possible AU under consideration. The AU corresponding to the HMM that provided maximum likelihood was chosen. Note the approach classifies facial actions without using information about which frame contained the action peak. Generalization to novel subjects was again tested using leave-one-out cross-validation.

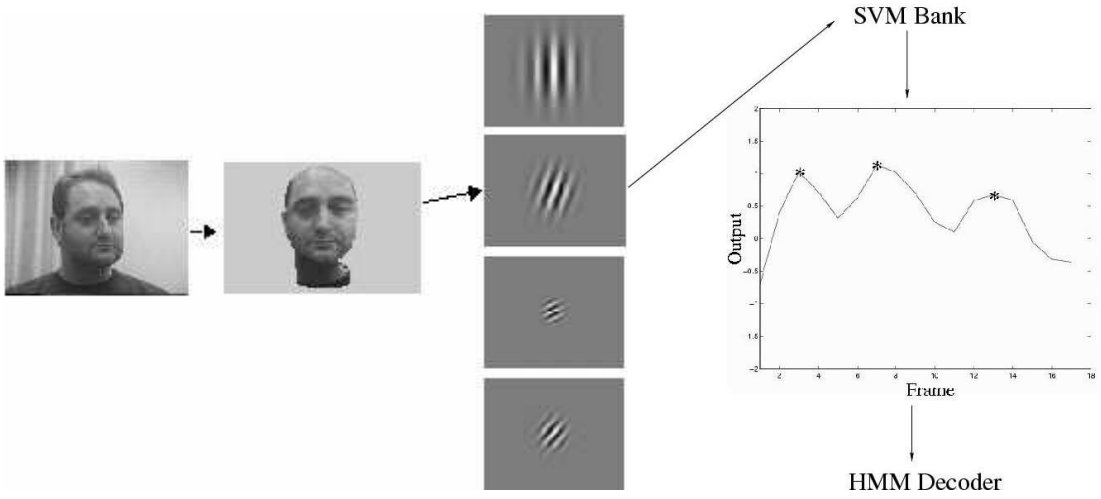


Figure 4: Flow diagram of recognition system. First, head pose is estimated, and images are warped to frontal views and canonical face geometry. The warped images are then passed through

a bank of Gabor filters. SVM's are then trained to classify facial actions from the Gabor representation in individual video frames. The output trajectories of the SVM's for full video sequences are then channeled to hidden Markov models.

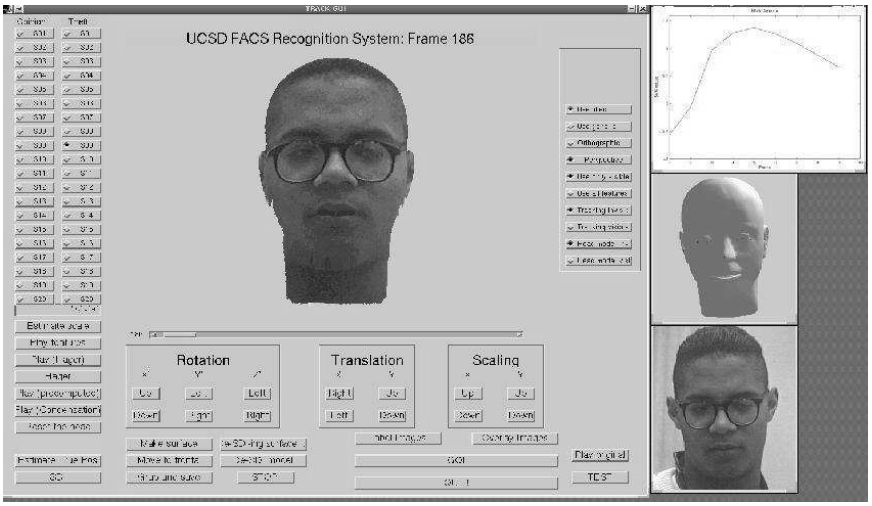


Figure 5: User interface for the FACS recognition system. Bottom right: Frame from the dataset. Middle right: Estimate of head pose. Center: Warped to frontal view and conical geometry. Top right: The curve shows the output of the blink detector for the video sequence. This frame is in the relaxation phase of a blink.

Results

Classifying individual frames with SVM's

SVM's were first trained to discriminate images containing the peak of blink sequences from randomly selected images containing no blinks. A nonlinear SVM applied to the Gabor representations obtained 95.9% correct for discriminating blinks from non-blinks for the peak frames. The nonlinear kernel was of the form $1/(k+d)^2$ where d is Euclidean distance, and k is a constant. Here k=4.

Recovering FACS dynamics

Figure 6a shows the time course of SVM outputs for complete sequences of blinks. Although the SVM was only trained to discriminate open from closed eyes, its output produced a continuous trajectory that correlated well with the amount of eye opening at each video frame. The SVM outputs provide information about FACS dynamics that was previously unavailable by human coding due to time constraints. Current coding methods provide only the beginning and end of the action, along with the location and magnitude of the action unit peak. This information about dynamics may be useful for future behavioral studies.

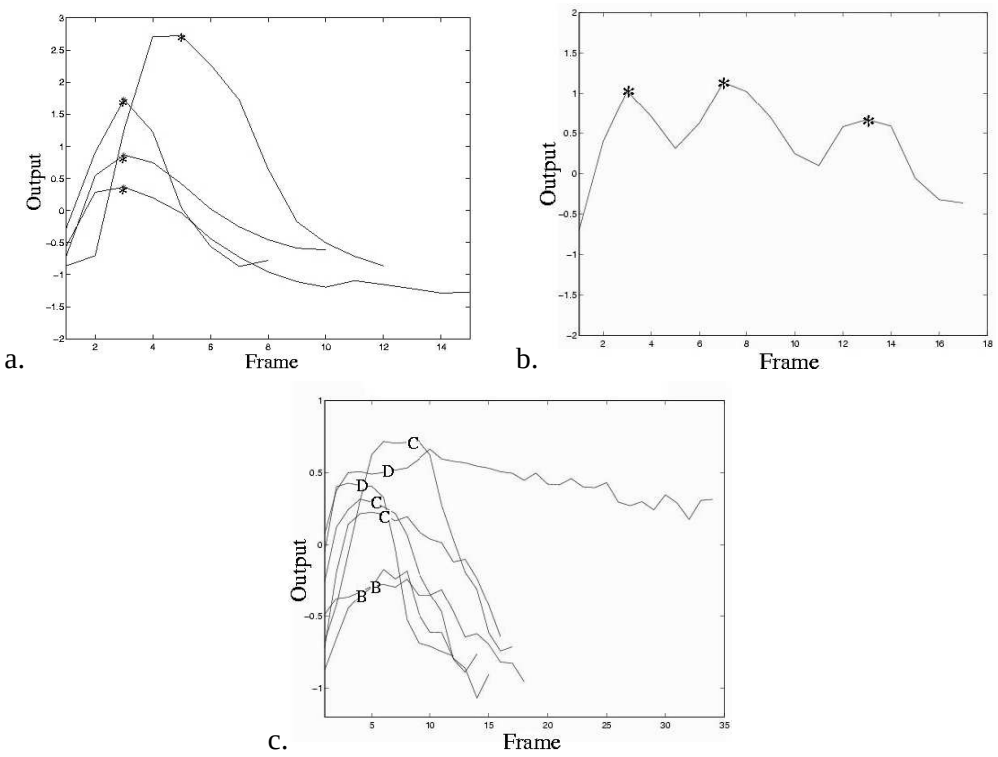


Figure 6: a. Blink trajectories of SVM outputs for four different subjects. Star indicates the location of the AU peak as coded by the human FACS expert. b. SVM output trajectory for a blink with multiple peaks (flutter). c. Brow raise trajectories of SVM outputs for one subject. Letters A-D indicate the intensity of the AU as coded by the human FACS expert, and are placed at the peak frame.

One approach to detecting action units in continuous video would be to simply choose a threshold and decide that an action unit is present if the output of an SVM reaches that threshold. However, even when the output does not reach threshold, there may be information in the output trajectory to indicate an action unit. Figure 6c illustrates a case in point. Choosing a threshold of 0 would

miss the actions labeled intensity B. However, the action can be detected by examining the pattern of rise and fall of the sub-threshold output. To capture these dynamics we used the HMM approach previously described. Two hidden Markov models, one for Blinks and one for random sequences matched by subject and length, were trained and tested using leave-one-out cross-validation. The number of states was varied from 1-10 and the number of Gaussian mixtures per state vas varied from 1-7. Best performance of 98.2% correct was obtained using 6 hidden states and 7 Gaussians per state. .

Brow movement discrimination

The goal was to discriminate three action units localized around the eyebrows. Since this is a 3-category task and SVMs are originally designed for binary classification tasks, we trained a different SVM on each possible binary decision task: Brow Raise (AU 1+2) versus matched random sequences, Brow Lower (AU 4) versus another set of matched random sequences, and Brow Raise versus Brow Lower. The output of these three SVM's was then fed to an HMM for classification. The input to the HMM consisted of three values which were the outputs of each of the three 2-category SVM's. As for the blinks, the HMM's were trained on the ``test'' outputs of the SVM's. The HMM's achieved 78.2% accuracy using 10 states, 7 Gaussians per state and including the first derivatives of the observation sequence in the input. Separate HMM's were also trained to perform each of the 2-category brow movement discriminations in image sequences. These results are summarized in Table 2.

Figure 6c shows example output trajectories for the SVM trained to discriminate Brow Raise from Random matched sequences. As with the blinks, we see that despite not being trained to indicate AU intensity, an emergent property of the SVM output was the magnitude of the brow raise. Maximum SVM output for each sequence was positively correlated with action unit intensity, as scored by the human FACS expert ($r = .43$, $t(42) = 3.1$, $p = 0.0017$).

Action	**Percent Correct (HMM)**	**N**
Blink vs. Matched Random Seq.	98.2	168
Brow Raise vs. Matched Random Seq.	90.6	48
Brow Lower vs. Matched Random Seq.	75.0	14
Brow Raise vs. Brow Lower	93.5	31
Brow Raise vs. Lower vs. Random	78.2	62

Table 2: Summary of results. All performances are for generalization to novel subjects. Random: Random sequences matched by subject and length. N: Total number of positive (and also negative) examples.

The contribution of Gabor filtering of the image was examined by comparing linear and nonlinear SVM's applied directly to the difference images versus to Gabor outputs. Consistent with our previous findings (Littlewort, Bartlett & Movellan, 2001), Gabor filters made the space

more linearly separable than the raw difference images. For blink detection, a linear SVM on the Gabors performed significantly better (93.5%) than a linear SVM applied directly to difference images (78.3%). Using a nonlinear SVM with difference images improved performance substantially to 95.9%, whereas the nonlinear SVM on Gabors gave only a small increment in performance, also to 95.9%. A similar pattern was obtained for the brow movements, except that nonlinear SVMs applied directly to difference images did not perform as well as nonlinear SVM's applied to Gabors. The details of this analysis, and also an analysis of the contribution of SVM's to system performance, are available in Bartlett et al., (2001).

Conclusions

The results of Study I provided guidance as to which image representations, or feature extraction methods, are most effective for facial action recognition. We found that Gabor wavelets and Independent Component Analysis gave best performance. These methods rely on precise alignment of the face image. Out-of-plane head rotations present a major challenge.

Study II explored an approach for handling out-of-plane head rotations in automatic recognition of spontaneous facial expressions from freely behaving individuals. The approach fits a 3D model of the face and rotates it back to a canonical pose (e.g., frontal view). We found that machine learning techniques applied directly to the warped images is a promising approach for automatic coding of spontaneous facial expressions.

This approach employed general purpose machine learning techniques that can be applied to the recognition of any facial action. The approach is parsimonious and does not require defining a different set of feature parameters or image operations for each facial action. While the database we used was rather large for current digital video storage standards, in practice the number of spontaneous examples of each action unit in the database was relatively small. We therefore prototyped the system on the three actions which had the most examples. Inspection of the performance of our system shows that 14 examples was sufficient to successfully learn an action, an order of 50 examples was sufficient to achieve performance over 90%, and an order of 150 examples was sufficient to achieve over 98% accuracy and learn smooth trajectories. Based on these results, we estimate that a database of 250 minutes of coded, spontaneous behavior would be sufficient to train the system on the vast majority of facial actions.

One exciting finding is the observation that important measurements emerged out of filters derived from the statistics of the images. For example, the output of the SVM filter matched to the blink detector could be potentially used to measure the dynamics of eyelid closure, even though the system was not designed to explicitly detect the contours of the eyelid and measure the closure. (See Figure 6.)

The results presented here employed hand-labeled feature points for the head pose tracking step. We are presently developing a fully automated head pose tracker (see Afterword).

All of the pieces are in place for the development of automated systems that recognize spontaneous facial actions at the level of detail required by FACS. Collection of a much larger, realistic database to be shared by the research community is a critical next step.

Acknowledgments

Support for this project was provided by ONR N00014-02-1-0616, NSF-ITR IIS-0220141 and IIS-0086107, DCI contract No.2000-I-058500-000, and California Digital Media Innovation Program DiMI 01-10130.

Notes

1. This section originally appeared in the following: Bartlett, M.S., Littlewort, G., Braathen, B., Sejnowski, T.J., & Movellan, J.R. (2003). A prototype for automatic recognition of spontaneous facial actions. In S. Becker & K. Obermayer, (Eds.) Advances in Neural Information Processing Systems, Vol 15. MIT Press. Reprinted with permission.

References

Bartlett, M.S. (2001). Face image analysis by unsupervised learning, Vol. 612 of The Kluwer International Series on Engineering and Computer Science. Boston: Kluwer Academic Publishers.

Bartlett, M.S., Braathen, B., Littlewort-Ford, G., Hershey, J., Fasel, I., Marks, T., Smith, E., and Movellan, J.R. (2001) Automatic Analysis of Spontaneous Facial Behavior: A Final Project Report. Institute for Neural Computation MPLab TR2001.08, University of California, San Diego.

Bartlett, M., Donato, G., Movellan, J., Hager, J., Ekman, P., & Sejnowski, T. (2000). Image representations for facial expression coding. In S. Solla, T. Leen, & K.-R. Muller (Eds.), Advances in Neural Information Processing Systems, Vol. 12. MIT Press.

Bartlett, M., Hager, J., Ekman, P., & Sejnowski, T. (1999). Measuring facial expressions by computer image analysis. Psychophysiology, 36, 253-263.

Bartlett, M., Movellan, J., & Sejnowski, T. (2002). Image representations for facial expression recognition. IEEE Transactions on Neural Networks 13(6) p. 1450-64.

Bartlett, Viola, Sejnowski, Golomb, Larsen, Hager, & Ekman, (1996). Classifying facial action. In Advances in Neural Informaiton Processing Systems 8. Cambridge, MA: MIT Press. p. 823-829.

Belhumeur, P., Hispanha, J., & Kriegman, D. (1997). Eigenfaces versus Fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7) p. 711-720.

Bell, A., & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 7(6), 1129--1159.

Bell, A., & Sejnowski, T. (1997). The independent components of natural scenes are edge filters. Vision Research, 37(23), 3327--3338.

Bonanno, G .A., & Keltner, D. (1997). Facial expressions of emotion and the course of conjugal bereavement. Journal of Abnormal Psychology, 106, 126-137.

Braathen, B., Bartlett, M.S., Littlewort-Ford, G., and Movellan, J.R. (2001). First Steps Towards Automatic Recognition of Spontaneous Facial Action Units. Proceedings of the ACM Conference on Perceptual User Interfaces.

Brand, M. (2001). Flexible flow for 3d nonrigid tracking and shape recovery. CVPR.

R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates, IEEE Trans Pattern Analysis and Machine Intelligence, vol. 15, no. 10, pp. 1,042-1,052, Oct. 1993.

Bugental, D. B. (1986). Unmasking the "polite smile": Situational and personal determinants of managed affect in adult child interaction. Personality and Social Psychology Bulletin, 12, 7-16.

Cohn, J., Kanade, T., Moriyama, T., Ambadar, Z., Xiao, J., Gao, J., and Imamura, H. (2001). A comparative study of alternative FACS coding algorithms. Robotics Institute Technical Report, Carnegie-Mellon Univerisity.

Darwin, C. (1872/1998). The expression of the emotions in man and animals. New York: Oxford. (3rd Edition, w/ commentaries by Paul Ekman).

J.G. Daugman, "Complete Discrete 2D Gabor Transform by Neural Networks for Image Analysis and Compression, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 36, pp. 1,169-1,179, 1988.

Doenges, P., Lavagetto, F., Ostermann, J., Pandzic, I.S., Petajan, E. (1997). MPEG-4: Audio/Video and Synthetic Graphics/Audio for Real-Time, Interactive Media Delivery. Image Communications Journal, Vol. 5, No. 4.

Donato, G., Bartlett, M., Hager, J., Ekman, P., & Sejnowski, T. (1999). Classifying facial actions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(10), 974--989.

Efron, D. (1941). Gesture and Environment. New York: King's Crown.

Ekman, P. (2001). Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage, 3rd Edition. New York: W.W. Norton.

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. (1978). The Facial Action Coding System. Palo Alto: Consulting Psychologists Press.

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

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

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. Journal of Personality and Social Psychology, 54, 414-420.

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

Ekman & E. L. Rosenberg (Eds.) (1997). What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System. . New York: Oxford.

Essa, I., & Pentland, A. (1997). Coding, analysis, interpretation, and recognition of facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7), 757--63.

Fasel, I.R., Smith, E.C., Bartlett, M.S. & Movellan, J.R. (2002). A comparison of Gabor filter methods for automatic detection of facial landmarks. Fifth International Conference on automatic face and gesture recognition. (accepted).

Fox, N.A., & Davidson, R.J. (1988). Patterns of brain electrical activity during facial signs of emotion in 10-month old infants. Developmental Psychology, *24*, 230-236.

Frank, M.G. (2002). Facial expressions. In N. Eisenberg (Ed.) International Encyclopedia of the Social and Behavioral Sciences. (in press). Oxford: Elsevier.

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.

Heller, M. and Haynal, V. (1994). The faces of suicidal depression (Translation). Les visages de la depression de suicide. Kahiers Psychiatriques Genevois (Medecine et Hygiene Editors) V. 16, p. 107-117.

Holland, M.K. and Tarlow G. (1972) Blinking and mental load Psychological Reports, 2, 31, 119-127.

Karson, C.N. (1988) Physiology of normal and abnormal blinking, Advances in Neurology, 49, 119-127.

Kitagawa, G. (1996). Monte carlo filter and smoother for non-Gaussian nonlinear state space models. Journal of Computational and Graphical Statistics, 5(1), 1--25.

M. Lades and J. Vorbruggen, J. Buhmann, J. Lange, W. Konen, C. von der Malsburg, and R. Wurtz, Distortion Invariant Object Recognition in the Dynamic Link Architecture, IEEE Trans. Computers, vol. 42, no. 3, pp. 300-311, Mar. 1993.

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. Psychophysiology, 27, 363-384.

Littlewort, G. ., Bartlett, M.S. and Movellan, J.R. (2001). Are your eyes smiling? Detecting genuine smiles with support vector machines and Gabor wavelets. Proceedings of the 8th Annual Joint Symposium on Neural Computation.

Movellan, J. (1995). Visual speech recognition with stochastic networks. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), Advances in Neural Information Processing Systems, Vol. 7 (pp. 851--858). Cambridge, MA: MIT Press.

Pantic, M., and Rothcrantz, L.J.M. (2000a). Expert System for automatic analysis of facial expressions. Image and Vision Computing 18, p. 881-905.

Pantic, M., and Rothcrantz, L.J.M. (2000b). Automatic Analysis of Facial Expressions: The State of the Art. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), p. 1424-1445.

Penev, P., & Atick, J. (1996). Local Feature Analysis: A general statistical theory for object representation. Network: Computation in Neural Systems 7(3):477-500.

Pighin, F. D. H, Szekiski, R. and Salesin, D. (19898) Synthesizing realistic facial expressions from photographs, Proc SIGGRAPH.

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.

Rosenberg, E. L; Ekman, P, & Blumenthal, J.A. (1998). Facial expression and the affective component of cynical hostility in male coronary heart disease patients. Health Psychology, 17, 376-380.

Rosenberg, E.L.. Ekman, P., Jiang,W., Babyak, M., and others (2001). Linkages between facial expressions of anger and transient myocardial ischemia in men with coronary artery disease. American Psychological Assn, US. Emotion 1(2) p. 107-115.

Sayette, M. A., Smith, D. W., Breiner, M.J., & Wilson, G. T. (1992). The effect of alcohol on emotional response to a social stressor. Journal of Studies on Alcohol, 53, 541-545.

Singh, A. Optic Flow Computation. Los Alamitos, Calif.: IEEE CS Press, 1991.

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

Van Orden K. , Jung, T.P. and Makeig, S. (2000) Eye Activity Correlates of Fatigue, Biological Psychology, 52, 3, 221-240.

Yacoob, Y., & Davis, L. (1996). Recognizing human facial expressions from long image sequences using optical flow. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(6), 636--642.

AFTERWORD

The Next Generation of Automatic Facial Expression Measurement

---

JAVIER R. MOVELLAN & MARIAN STEWART BARTLETT

The research presented in the previous chapter demonstrated proof that automatic recognition of facial actions in indoor environments is achievable with current technology. The system we explored used general purpose machine learning techniques that can be applied to recognition of any facial action provided enough data is available. One of the most useful aspects of the experiment was to help us identify the challenges that need to be met for automatic facial expression measurement systems to become a useful a tool for behavioral scientists:

- Availability of training data. The primary limitation in the number of action units recognized in the previous studies is the availability of training data. The computer system requires labeled training images of many subjects performing the facial actions we wish to recognize.

- Automatic face detection and alignment.

- Handling out-of-plane head rotations. Current approaches to facial expression recognition deteriorate with out-of-plane head rotations beyond about 10 degrees.

- Robustness to lighting conditions. Even in controlled experimental settings, lighting changes as the subject moves his or her head.

- Recognizing action unit combinations. Over 7000 distinct action unit combinations have been reported in the literature. It is impractical to train on all possible combinations. An approach to this issue is described in Smith et al. (2001).

- Temporal segmentation of facial actions. How does a system determine when the action begins, ends, and peaks?

- Handling missing image data. During out-of-plane head rotations, a portion of the face image is missing. One approach to this problem is multiple camera recording.

Three of these issues are discussed in more detail below: (1) Collection of a database of spontaneous facial expressions, (2) fully automatic face detection and tracking, and (3) fully automatic 3D head pose estimation.

Image databases

An important lesson learned from the speech recognition community is the need for large, shared image databases for training, testing, and evaluation. If an effort is made to develop and share such databases, automatic FACS recognition systems that work in real time in unconstrained environments will emerge, as occurred in speech recognition in the last decade. Development of these databases is a priority that will require joint effort from the computer vision, machine learning, and psychology communities. Because of differences between posed and spontaneous facial expressions, it is important that some of the databases contain spontaneous expressions. This makes collaboration with the behavioral science community all the more essential, as the computer vision community has little experience eliciting spontaneous expressions.

Mark Frank at Rutgers University, in collaboration with our laboratory and Paul Ekman, is collecting a new state of the art database of spontaneous facial expressions to serve as training data for computer vision systems. The database will be FACS coded by two certified FACS coders. The database will consist of a terabyte of uncompressed digital video from three cameras. The database will contain 100 subjects videotaped for 2.5 minutes each while they participate in a false opinion paradigm. This paradigm was selected for the variety of expressions it tends to elicit, including basic emotions and language related symbolic expressions. This database will provide the basis for training the computer to recognize a much larger set of spontaneous facial actions, and for examination of facial action dynamics. More databases such as this one will enable advancement of automated FACS recognition.

Fully automatic face detection and expression recognition

We have recently paired automatic facial expression measurement with fully automatic face detection and tracking. The combined system works in real-time, at about 15 frames per second

with no manual intervention. One version of the system was trained and tested on two publicly available datasets of FACS-coded expressions of basic emotions: Pictures of Facial Affect, (Ekman and Friesen, 1976), and DFAT-504 (Kanade Cohn & Tian, 2000). Below, we present results for coding expressions in terms of 7 dimensions: Joy, sadness, surprise, anger, disgust, fear, and neutral. The mechanism is the same as for recognizing facial actions. The only difference is the training data and how they are labeled. When large datasets of spontaneous facial actions become available such as the one described above, this system can be trained to code expressions in terms of facial actions. The system is reviewed here. More information is available in Bartlett et al. (2003) and Littlewort et al. (in press).

Figure 1. The face detector results for two complex background and illumination conditions. The system works in real time at 30 frames per second on a fast PC.

Face detection

The face detector works in real time, and is based on a state-of-the-art face detection system developed by Viola & Jones (2001). We have developed methods in our lab to make the system significantly faster and more robust to difficult illumination and background conditions (see Figure 1). The full system, including enhancements to the Viola & Jones approach, is described in Bartlett et al., (2003) and Littlewort et al. (in press). We made source code for the face detector freely available at http://kolmogorov.sourceforge.net. Performance on standard test sets are equal to the state-of-the-art in the computer vision literature (e.g. 90% detection and 1 in a million false alarms on the CMU face detection test set). The CMU test set has unconstrained lighting and background. When lighting and background can be controlled, such as in behavioral experiments, accuracy is much higher. We are also using the same technology to detect facial features within the face (Fortenberry et al., submitted). The precision of the current systems is in the order of 1/4 of an iris, similar to the precision obtained by human labelers in our previous study.

Facial expression recognition

The output of the face detector is scaled to 90x90 and fed directly to the facial expression analysis system (see Figure 2). The system is essentially the one described in the previous chapter. First the face image is passed through a bank of Gabor filters at 8 orientations and 9 scales (2-32 pixels/cycle at 0.5 octave steps). The filterbank representations are then channeled to a classifier to code the image in terms of a set of expression dimensions. We have found support vector machines to be very effective for classifying facial expressions (Littlewort et al., in press, Bartlett et al., 2003). Recent research at our lab has demonstrated that both speed and accuracy are enhanced by performing feature selection on the Gabor filters prior to classification (e.g. Bartlett et al., 2003). This approach employs Adaboost (Freund & Shapire, 1996) a state of the art technique for feature selection that sequentially selects the feature that gives the most information about classification *given* the features that have been already selected.

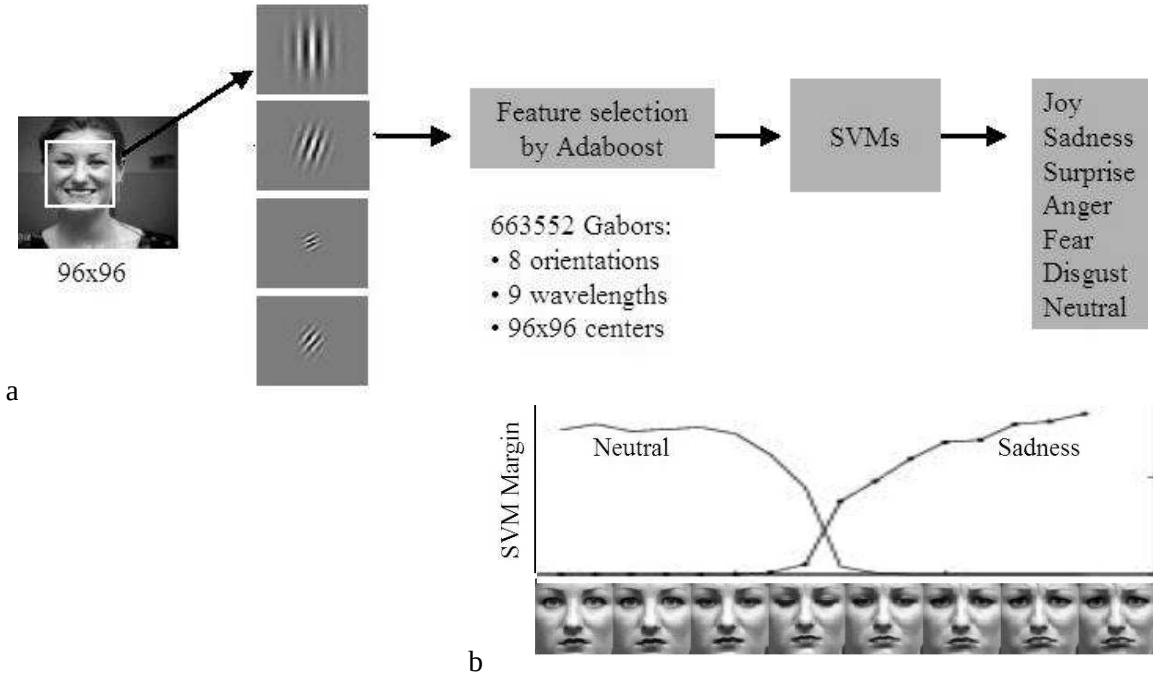


Figure 2. a. Facial expression recognition system. b. Outputs of the SVMs trained for neutral and sadness for a full image sequence of a test subject performing sadness.

Performance

The facial expression recognition system can be trained to recognize any target expression dimension. Here we present results for recognizing 7 basic emotions (joy, sadness, surprise, fear, disgust, anger, neutral). The system was trained and tested on Cohn and Kanade's DFAT-504 dataset (Kanade, Cohn, & Tian, 2000). This dataset consists of video sequences of university students posing facial expressions. An experimenter described and modeled each desired facial display for each subject. For our study, we selected the 313 sequences from the dataset that were labeled as one of the 6 basic emotions. The sequences came from 90 subjects, with 1 to 6 emotions per subject. Subjects ranged in age from 18 to 30 years. 65% were female, 15% were African-American, and 3% were Asian or Latino.

All faces in this dataset were successfully detected. The expression recognition system was trained on the last frame of each sequence, which contained the highest magnitude of the target expression (peak frames). Neutral expression samples consisted of the first frame of each sequence. Seven support vector machines, one for each expression, were trained using one-versus-all partitioning (e.g. joy vs. everything else). A nonlinear radial basis function kernel was employed. The emotion category decision was then implemented by choosing the classifier with the maximum output for the test example.

Performance on novel subjects was tested using leave-one-out cross-validation (Tukey, 1951). The system obtained 93% agreement with the emotion category labels assigned in the database. We were encouraged by these results, since the best published results on this database by other systems is 81%-83% accuracy.

Performance of the system was also evaluated on a second publicly available dataset, "Pictures of Facial Affect", collected by Paul Ekman and Wallace Friesen (1976). This dataset contains images of 20 Caucasian adults, male and female, posing 6 expressions of basic emotion, plus neutral. Subjects were directed to move specific facial muscles posited by Ekman and Friesen to comprise the universal expressions of emotion, and variations thereof. The automatic facial expression recognition system obtained 97% accuracy for generalization to novel subjects, trained by leave-one-subject-out cross-validation. This is about 10 percentage points higher than the best previously reported results on this dataset. A demo of the system is available on our webpage: http://mplab.ucsd.edu. Users can upload an image and run the face detector and expression classifier on their own image.

Figure 2b illustrates system outputs for a full sequence of a facial expression. An emergent property was that the outputs of the classifier change smoothly as a function of time, providing a potentially valuable representation to code facial expression dynamics in a fully automatic and unobtrusive manner. This would provide information about expression dynamics at a temporal resolution previously intractable by human coding. The time courses of the system outputs will be analyzed with dynamical models in the next phase of development.

While the system gives the best performance we know of for recognition of basic emotions on standard datasets, we have found that in unconstrained environments the system is still sensitive to changes in illumination. (Performance is about 80% correct on unconstrained images from the web.) The next phase of development will work on making this system more robust, particularly to variations in lighting conditions.

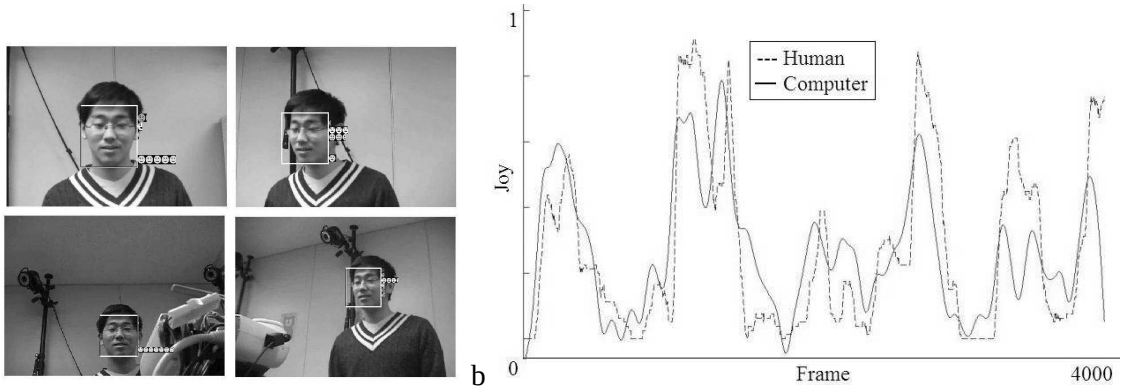


Figure 3. a. Facial expression is measured during interaction with the Robovie robot from the continuous output of four video cameras. b. Mean human ratings compared to automated system outputs for 'joy' (one subject).

Pilot study: measuring spontaneous expressions in unconstrained environments

We conducted a pilot study at the Intelligent Robotics and Communication laboratories at ATR, Japan, to evaluate the expression recognition system in unconstrained environments. Subjects interacted in an unconstrained manner with RoboVie, a communication robot developed at ATR and the University of Osaka (Ishiguro, 2001).

To improve performance of the system we simultaneously recorded video from 4 video cameras. 14 paid participants recruited from the university of Osaka were invited to interact with RoboVie for a 5 minute period. Faces were automatically detected and facial expressions classified independently on the four cameras. This resulted in a 28 dimensional vector per video frame (7 emotion scores x 4 cameras), The output of the 4 cameras was then combined using a standard probabilistic fusion model. To assess the validity of the system, four naive human observers were presented with the videos of each subject at 1/3 speed. The observers indicated the amount of happiness shown by the subject in each video frame by turning a dial, a technique commonly used in marketing research. Figure 3 compares human judgments with the automated system. The frame-by frame correlation of the human judges averaged across subjects and judge pairs was 0.54, The average correlation between the 4 judges and the automated system was 0.56, which does not differ significantly from the human/human agreement ($t(13) = 0.15$, $p<0.875$). Figure 3b shows frame by frame the average scores given by the 4 human judges for a particular subject, and the scores predicted by the automatic system. We are presently evaluating the 4 camera version of the system as a potential new tool for research in behavioral and clinical studies.

Tracking Out-of-Plane Head Motion

The previous chapter showed that 3D alignment and rotation to frontal views is a viable approach to recognizing facial actions in the presence of out-of-plane head rotations. 3D alignment may give more precise facial expression measurements than the 4-camera approach described above. At the time of the study in the previous chapter, head pose estimates were obtained from a set of eight hand-marked feature points. Since then we developed a system for fully automatic head pose estimation without hand-marking (Marks, Hershey, Roddey, & Movellan, 2003). See Figure 4. The system, which we call *Gflow*, dynamically adjusts the relative contributions of optic flow and template based tracking information in a generative model, making it quite robust to noise. Recent developments in the computer vision field (e.g. Brand, 2001) enable this kind of nonlinear filtering operation to be performed in real time. Optic flow estimation during nonrigid deformations due to speech and changes in facial expression is enabled using a set of morph bases Brand (2001). We are presently collecting training data so that the system can be applied to arbitrary subjects. The system will be demonstrated at NIPS 2003, and is scheduled to be available in the spring of 2004. This is another example of the tools that will be ready to apply to recognition of spontaneous facial actions when the datasets become available.

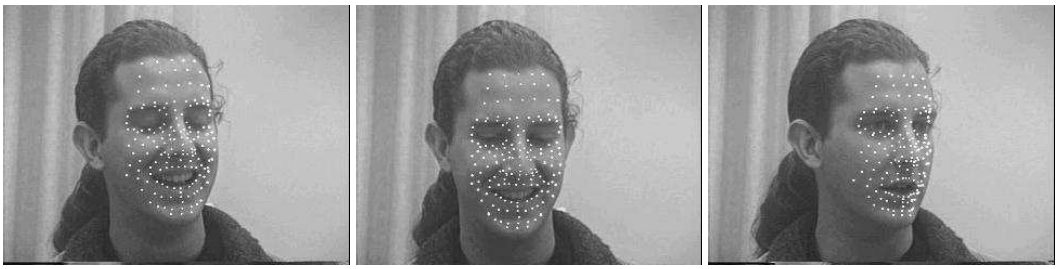


Figure 4: A demonstration of 3D head pose tracking on a subject from Frank & Ekman (1997). The dots are not painted on the face. The dots indicate the computer's estimate of the location of points on the face in each frame. Given these points, head pose can be estimated and the face image mapped to a frontal view.

Applications

Clinical Applications

We are beginning to explore applications of automatic facial expression measurement for clinical assessment and basic research on mental disorders. Forty-five percent of all diagnoses listed in the official diagnostic manual of the American Psychiatric Association involve abnormal, distorted, or squelched emotional responses (Thoits, 1985). Facial expression measurement (coded by hand) has already proven useful for diagnosis, predicting improvement, and measuring

response to treatment (Ekman, Matsumoto, & Friesen, 1997; Berenbaum & Oltmanns, 1992; Katsikitis & Pilowsky, 1991;Steiner, 1986). In addition to aiding in diagnosis and treatment, facial expression measurement has helped provide insight into the clinical nature of psychopathologies including suicidal depression (Heller & Haynal, 1994), neurological disorders including blunted affect in schizophrenia (Barenbaum & Oltmanns, 1992, Kring & Neale, 1996), and parkinsonism (Ellgring, 1997), social disorders including aggressive adolescence (Keltner, Moffit, & Strouthamer-Loeber, 1995), and cardiac pathology (Rosenberg, Ekman, & Blumenthal, 1998; Rosenberg et al, 2001). For example, facial expression measurement supported a qualitatively difference between suicidal and major depression, involving contempt (Heller & Haynal, 1994). Facial expression measurement during marriage counseling both predicted the outcome and provided clinical insights into marital failure. The expression of disgust or contempt, but not anger, predicted divorce (Gottman, 1994). Thus Facial expression measurement can aid in the diagnosis and treatment of psychopathology and social disorders, and also contribute to the understanding of the underlying condition.

The work described above required 100 hours of training, and two hours to manually score each minute of video tape. The time required for hand-coding of videotaped behavior has been identified as one of the main obstacles to doing research on emotion (Frank, 2002; Ekman et al., 1993). Computer vision systems will unlock this demonstrably important area in the diagnosis and treatment of psychopathology and social disorders. The computer vision tools developed here also have the potential to measure facial expression dynamics at a temporal resolution previously intractable by hand-coding. Hence we may be able to expand the analysis of these clinical groups to include dynamic qualities of their expression, rather than just the morphology of their expression. For example, with hand coding of facial expressions, it has been found that schizophrenic patients have a more disorganized facial muscle movement pattern (Krause, et al, 1989). It may be thus possible for computer vision systems to make differentiations such as between schizophrenic patients and psychotic depressed patients from their facial dynamics.

Connecting Perception and Action

Researchers have begun to explore and develop digital creatures that have the ability to express emotions, recognize emotions, and whose behavior is modulated by synthetic emotional dynamics. This area of research is known in the computer science literature as ``affective computing" (Picard, 1997). See Pantic & Rothcrantz (2003) for a recent survey of affect sensitive human-computer interaction. Intelligent digital devices that are personal, emotional and engaging may revolutionize the way we interact with and think of computers. For example, we are applying affective computing technology to develop a new generation of automatic tutors, in collaboration with Ron Cole at U.C. Colorado. These tutors will interact with the students via computer animated agents which will adapt to the cognitive and emotional state of the students, the way good teachers do.

Affective computing technology will also have a significant impact on the digital entertainment industry. We are taking the first steps to realize the idea of personal robots that are

aware of their human companions, understand their emotional expressions, and that develop personable, engaging interactions with humans. (See Figure 5).

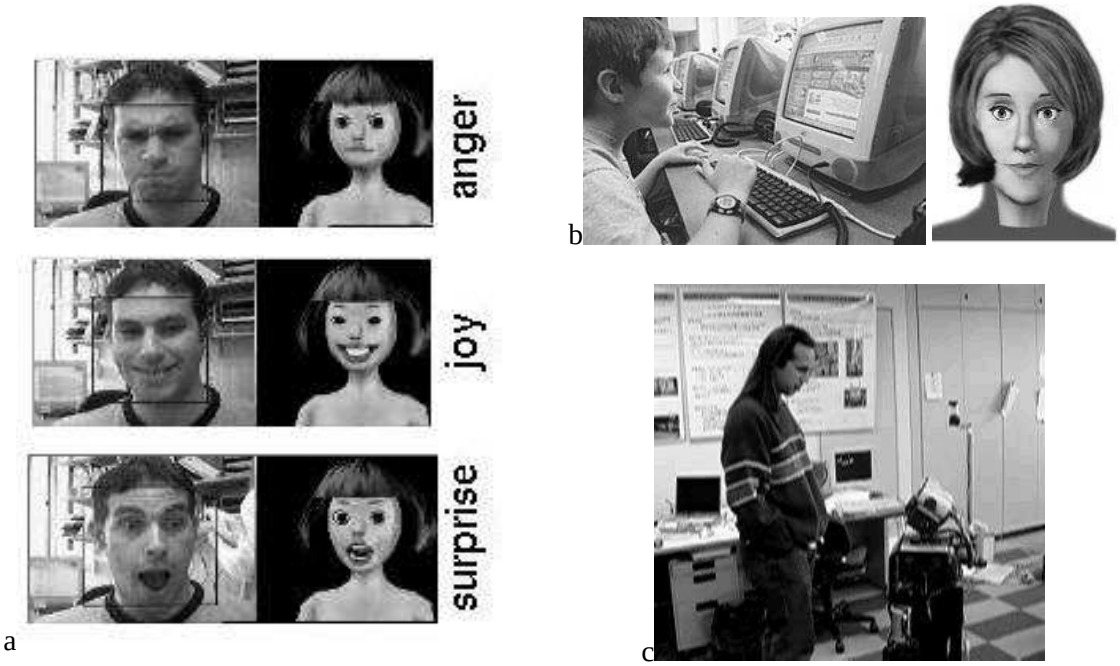


Figure 5. Connecting perception and action a. The animated character mirrors the facial expression of the user. b. Facial expression measurement is being deployed in automatic tutoring systems. c. Face detection and tracking software was ported to the Robovie robot, developed by M. Ishiguro at ATR.

Emotion Mirror

The emotion mirror is a prototype system that recognizes the emotion of the user and responds in an engaging way. It involves real time interaction between machine perception and animation software. The larger objective is to incorporate this system into robotic and computer animation applications in which it is important to engage the user at an emotional level and/or have the computer recognize and adapt to the emotions of the user. In the emotion mirror, the face-finder captures a face image which is sent to the emotion classifier. The outputs of the 7-emotion classifier is a seven-dimensional vector that encodes the expression of the user at the current video frame. This code is sent to CU Animate, a computer animation tool developed at CU Boulder, to render a computer animated character in real time. The character then mimics the facial expression of the user. Figure 5a shows the prototype system.

Deployment in real-time robotic environments

We implemented this system on an active camera with 5 degrees of freedom (roll, pitch, and yaw, plus one degree of freedom each on two directional cameras). The system contains a third omnidirectional camera as well. The robot head tracks faces in mobile subjects as they move about the room. See Figure 5c. Facial expressions are automatically classified from the video stream. This system used to collect data for evaluating system performance in a real-time environment.

Summary and Conclusions

The automatic analysis of the face and facial expressions is rapidly evolving into a mature scientific discipline. The next ten years are likely to see the development of a new generation of systems capable of recognizing the human face and facial expressions at levels of accuracy similar to that of human experts. This technology will provide wonderful new tools to behavioral scientists that will help make dramatic progress in our understanding of the human mind. These tools are likely to produce paradigmatic changes in the cognitive and behavioral sciences. We will also see the progressive development of machines that interact with us in ways we cannot currently conceive, including robots and computer animated characters that are aware of our presence and make inferences about our mental states the way other humans do. These socially aware systems may provide highly effective treatments (e.g. speech therapy) to populations that cannot currently afford them or who do not have daily access to therapists. They may liberate teachers from the more automatic parts of the educational experience to let them concentrate on the more creative aspects of it. They may also help us measure and track the effect of new drugs to address mental and affective disorders. The scientific community, as well as society at large, needs to begin to address the ethical and political challenges that this technology will bring about. Difficult decisions will need to be made about the pros and cons of these technologies and decide when it should and should not be used. The challenges are great and the potential impact both on science and on our daily life are enormous.

References

Bartlett, M.S., Littlewort, G., Braathen, B., Sejnowski, T.J., & Movellan, J.R. (2003). A prototype for automatic recognition of spontaneous facial actions. In S. Becker & K. Obermayer, (Eds.) *Advances in Neural Information Processing Systems, Vol 15*. MIT Press.

Berenbaum, H., & Oltmanns, T. F. (1992). Emotional experience and expression in schizophrenia and depression. *Journal of Abnormal Psychology, 101,* 37-44.

Brand, M. (2001). Flexible flow for 3d nonrigid tracking and shape recovery. *CVPR.*

Ekman, P., Huang, T., Sejnowski, T. and Hager, J. (1993). *Final Report to NSF of the Planning Workshop on Facial Expression Understanding,* 1992. Available from UCSF, HIL-0984, San Francisco, CA 94143.

Ekman, P., Matsumoto, D., & Friesen, W.V. (1997). Facial expression in affective disorders. In P. Ekman & E.L. Rosenberg (Eds.), What the face reveals*: Basic and applied studies of spontaneous expression using the Facial Action Coding System* (pp. 331-341). New York: Oxford University Press.

Ekman, P., & Friesen, W. V. (1976). *Pictures of Facial Affect.* Palo Alto: Consulting Psychologists Press.

Ellgring, H. (1997). Nonverbal expression of psychological states in psychiatric patients: Afterword. In Ekman & E. L. Rosenberg (Eds.) (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System.* (pp 395-397). New York: Oxford.

Fasel, I., Fortenberry, B., and Movellan, J. (submitted). Real time detection of face, eyes, and blinks in video using gently boosted classifiers.

Frank, M.G. (2002). Facial expressions. In N. Eisenberg (Ed.) *International Encyclopedia of the Social and Behavioral Sciences*. (in press). Oxford: Elsevier.

Frank, M.G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high stake lies. *Journal of Personality and Social Psychology, 72,* 1429-1439.

Freund, Y., and Schapire, R.E. (1996). Experiments with a new Boosting algorithm. *Proc. 13th International Conference on Machine Learning*. Morgan Kaufmann, p. 148-146.

Gottman, J. (1994). *Why Marriages Succeed or Fail.* New York: Fireside.

Heller, M. and Haynal, V. (1994). The faces of suicidal depression (Translation). Les visages de la depression de suicide. *Kahiers Psychiatriques Genevois (Medecine et Hygiene Editors*) V. 16, p. 107-117.

Ishiguro, H., Ono, T., Imai, M., Maeda, T., Kanda T., and R. Nakatsu, R. (2001). Robovie: an interactive humanoid robot. *International Journal of Industrial Robotics* 28(6):498–503.

Kanade, T., J.F. Cohn, J.F., and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings of the fourth IEEE International conference on automatic face and gesture recognition (FG'00)*, pages 46–53, Grenoble, France, 2000.

Katsikitis, M., & Pilowsky, I. (1991).A controlled quantitative study of facial expression in Parkinson's disease and depression. *Journal of Nervous and Mental Disease, 179*, 683-688.

Keltner, D., Moffit, T., & Stouthamer-Loeber, M. (1995). Facial expression and psychopathology in adolescent boys. *Journal of Abnormal Psychology, 104*, 644-652.

Krause, R., Steimer, E., Sanger-Alt, C., & Wagner, G. (1989). Facial expressions of schizophrenic patients and their interaction partners. *Psychiatry, 52*, 1-12.

Kring, A. M., & Neale, J. M. (1996). Do schizophrenics show a disjunctive relationship among expressive, experiential, and psychophysiological components of emotion? *Journal of Abnormal Psychology 105*, 249-257.

Littlewort, G., Bartlett, M.S., Chenu, J., Fasel, I., Kanda, T., Ishiguro, H., & Movellan, J.R.(in press). Towards Social Robots: Automatic evaluation of human-robot interaction by face detection and expression classification. *Advances in Neural Information Processing Systems*, Vol 16. MIT Press.

Lyons, M., J. Budynek, A. Plante, and S. Akamatsu. (2000). Classifying facial attributes using a 2-d gabor wavelet representation and discriminant analysis. In *Proceedings of the 4th international conference on automatic face and gesture recognition*, pages 202–207, 2000.

Marks, T.K., Roddey, J.C., Hershey, J., and Movellan, J.R. (2003). Determining 3D face structure from video images using G-Flow. Demonstration, *Advances in Neural Information Processing Systems*.

Pantic, M., & Rothcrantz, L.J.M. (2003). Toward an Affect-Sensitive Multimodal Human-Computer Interaction. Proceedings of the IEEE 91(9) p. 1370-1390.

Picard, R.W. (1997). *Affective Computing*. Cambridge: MIT Press.

Rosenberg, E. L; Ekman, P, & Blumenthal, J.A. (1998). Facial expression and the affective component of cynical hostility in male coronary heart disease patients. *Health Psychology*, 17, 376-380.

Rosenberg, E.L.. Ekman, P., Jiang,W., Babyak, M., and others (2001). Linkages between facial expressions of anger and transient myocardial ischemia in men with coronary artery disease. *American Psychological Assn, US. Emotion* 1(2) p. 107-115.

Smith, E., Bartlett, M.S., amd Movellan, J.R. (2001). Computer recognition of facial actions: An approach to co-articulation effects. *Proceedings of the 8th Joint Symposium on Neural Computation*.

Steiner, F. (1986). Differentiating smiles. In E. Branniger-Huber & F. Steiner (Eds.) *FACS in psychotherapy Research*, p. 139-148. Zurich: Department of Clinical Psychology.

Thoits, P. A. (1985). Self-labeling processes in mental illness: The role of emotional deviance. *American Journal of Sociology, 91*, 221-249.

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

Viola, P. and Jones, M. (2001). Robust real-time object detection. *Second International Workshop no Statistical and Conceptual Theories of Vision. International Conference on Computer Vision*.