Automatic Analysis of Spontaneous Facial Behavior: A Final Project Report

(UCSD MPLab TR 2001.08, October 31 2001)

Marian S. Bartlett[1], Bjorn Braathen[1], Gwen Littlewort-Ford[1], John Hershey[2], Ian Fasel2, Tim Marks[2], Evan Smith[2], Terrence J. Sejnowski[1,2,3], & Javier R. Movellan[1,2]

[1]Institute for Neural Computation
[2]Department of Cognitive Science
University of California, San Diego

[3]The Salk Institute and Howard Hughes Medical Institute

Abstract

The Facial Action Coding System (FACS) is the leading standard for measuring facial expressions in the behavioral sciences (Ekman & Friesen, 1978). FACS coding is presently performed manually by human experts, it is slow, and requires extensive training. Automating FACS coding could have revolutionary effects in our understanding of human facial expression and on the development of computer systems that understand facial expressions. Two teams, one at University of California San Diego and the Salk Institute, and another at University of Pittsburgh and Carnegie Mellon University, were challenged to develop prototype systems for automatic recognition of spontaneous facial expressions. Working with spontaneous expressions required solving technical and theoretical challenges which had not been previously addressed in the field. This document describes the system developed by the UCSD team. The approach employs 3-D pose estimation and warping techniques to reduce image variability due to general changes in pose. Machine learning techniques are then applied directly on the warped images or on biologically inspired representations of these images. No efforts are made to detect contours or other hand-crafted image features. This system employed general purpose learning mechanisms that can be applied to recognition of any action unit. The approach is parsimonious and does not require defining a different set of feature parameters or image operations for each facial action. The system was tested on a set of eyelid and eyebrow movements and successfully identified these movements in novel subjects. We showed that 3D tracking and warping followed by machine learning techniques directly applied to the warped images, is a viable and promising technology for automatic facial action recognition. One exciting aspect of the approach presented here is that information about movement dynamics emerged out of filters which were derived from the statistics of images. We believe all the pieces of the puzzle are ready for the development of automated systems that recognize spontaneous facial actions at the level of detail required by FACS. The main factor impeding development in this field is the lack of sufficiently large databases for training a greater variety of action units, and which may become a standard for comparison between different approaches. Based on our experience in this project we estimate that a database of 500 subjects, with 1 minute of rich facial behavior per subject, would be sufficient for dramatic improvements in the field.

1 Introduction

The Facial Action Coding System (FACS) developed by Ekman and Friesen (Ekman & Friesen, 1978) provides an objective description of facial behavior from video. It decomposes facial expressions into action units (AUs) that roughly correspond to independent muscle movements in the face (see Figure 1). Measurement of facial behavior at the level of detail of FACS provides information for detection of deceit, including

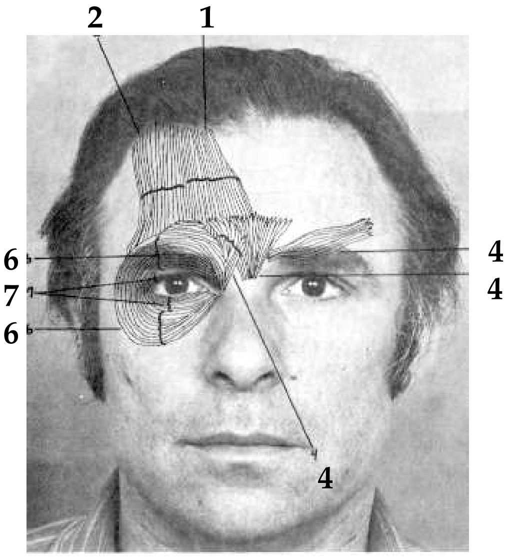


Figure 1: The Facial Action Coding System decomposes facial motion into component actions. The upper facial muscles corresponding to action units 1, 2, 4, 6 and 7 are illustrated. Adapted from Ekman & Friesen (1978).

information about whether an expression is posed or genuine and leakage of emotional signals that an individual is attempting to suppress – See (Ekman, 2001) for a compete discussion.

A major impediment to the widespread use of FACS is the time required to train human experts and to manually score the video tape. FACS coding is presently performed by trained human observers who analyze frame by frame the expression in each video frame into component actions (see Figure 1). Approximately 300 hours of training are required to achieve minimal competency on FACS, and each minute of video tape takes approximately one hour to score. An automated system would make fast, inexpensive, and objective facial expression measurement widely accessible tool for research, clinical, education, and security applications.

A number of ground breaking systems have appeared in the computer vision literature for facial expression recognition. These systems include measurement of facial motion through optic flow (Mase, 1991; Yacoob & Davis, 1994; Rosenblum, Yacoob, & Davis, 1996; Essa & Pentland, 1997) and through tracking of high-level features (Tian, Kanade, & Cohn, 2001), methods for relating face images to physical models of the facial skin and musculature (Mase, 1991; Terzopoulus & Waters, 1993; Li, Roivainen, & Forchheimer, 1993; Essa & Pentland, 1997), methods based on statistical learning of images (Cottrell & Metcalfe, 1991; Padgett & Cottrell, 1997; Lanitis, Taylor, & Cootes, 1997; Bartlett, Donato, Movellan, Hager, Ekman, & Sejnowski, 2000), and methods based on biologically inspired models of human vision (Bartlett, 2001).

Most of the previous work relied on datasets of posed expressions collected under controlled imaging conditions, with subjects deliberately facing the camera. Extending these systems to spontaneous facial behavior is a non-trivial problem of critical importance for realistic application of this technology. Psychophysical work has showed that spontaneous facial expressions differ from posed expressions in their morphology (which muscles are moved), and their dynamics (how the muscles are moved). For example, subjects often contract different facial muscles when asked to pose an emotion such as fear versus when they are actually experiencing fear. Moreover, spontaneous expressions have a fast and smooth onset, while in posed expressions, the onset tends to be slow and jerky, and the actions typically do not peak simultaneously (Ekman, 2001). In addition, spontaneous facial expressions pose a number of technical challenges that are not addressed by the current generation of recognition systems. For example, spontaneous facial expressions often occur in the presence of out-of-plane rotations, due to the fact that people nod or turn their head as they communicate spontaneously with others. This substantially changes the input to the computer vision systems, and produces variations in lighting as the subject alters the orientation of his or her head relative to the lighting source.

Two research teams (UCSD and CMU/Pitt) independently developed systems for automatically measuring facial actions using computer vision techniques (Bartlett, Hager, Ekman, & Sejnowski, 1999; Cohn, Zlochower, Lien, & Kanade, 1999; Donato, Bartlett, Hager, Ekman, & Sejnowski, 1999; Lien, Kanade, J.F., & Li, 2000; Tian et al., 2001). In this project, the two research teams compared performance of their systems on a common dataset. Each developed approaches to address the technical challenges posed by spontaneous facial expressions. Here we present the UCSD approach. Historically the UCSD team has championed methods that merge machine learning techniques and biologically inspired models of human vision while the CMU/Pitt team specializes on contour-based representations of facial features. Both approaches have advantages and disadvantages and it is unclear which approach will scale better when applied to more realistic problems.

2 Previous work at UCSD

Our previous work focused on the used of unsupervised machine learning techniques to find efficient image representations. We compared facial action recognition performance using image filters derived from supervised and unsupervised machine learning techniques. These data-driven filters were compared to Gabor wavelets, in which the image filters are predefined, and closely model the response transfer function of visual cortical receptive fields. These filters can be considered adaptive in a developmental or phylogenic sense. 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; Bartlett, 2001).

Image database: We used a database of directed facial actions collected in a previous collaboration with Paul Ekman at the University of California, San Francisco. The full database consists of 1100 sequences containing over 150 distinct actions and action combinations, and 24 subjects. A sample of three facial actions is shown in Figure 2. Our initial analysis addressed 12 facial actions, 6 in the upper face and 6 in the lower face, performed by 20 subjects.

Adaptive methods: We compared four techniques for developing image filters adapted to the statistical structure of face images. The techniques were Principal Component Analysis (PCA), Local Feature Analysis (LFA) (Penev & Atick, 1996), Fisher linear discriminants (FLD), and Independent Component Analysis (ICA). 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 sensitive to high-order dependencies, not just covariances in the data. 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). The PCA and ICA representations are described in more detail here.

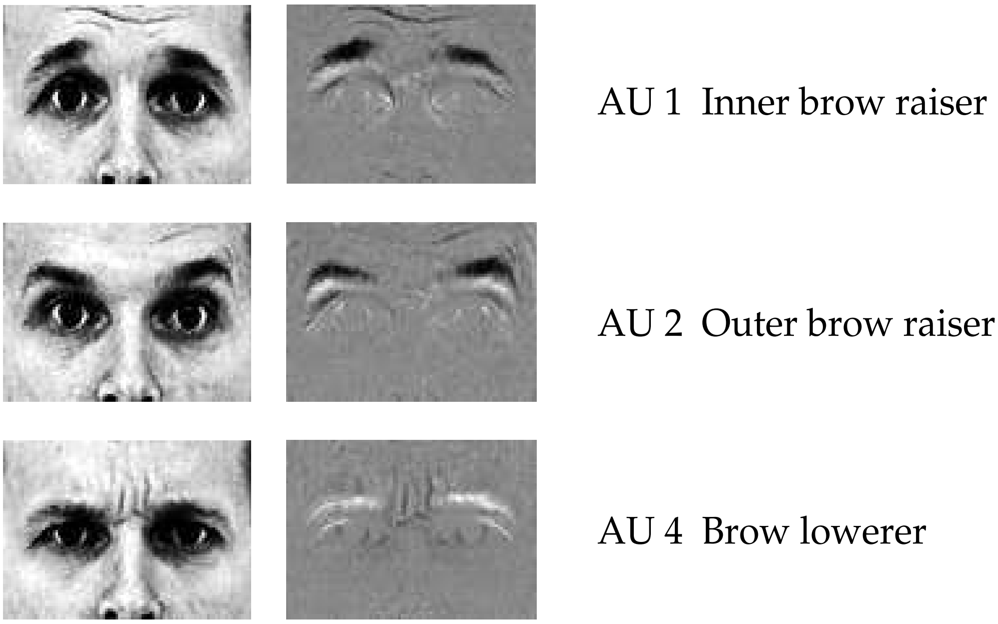


Figure 2: Sample of three upper facial actions at high intensity muscular contraction. On the right are the difference image obtained by subtracting the pixel gray-values of a neutral expression image. Gray is zero, negative values are dark and positive values are light. AU: Action Unit.

PCA: The PCA representation is also known as eigenfaces (Turk & Pentland, 1991). We performed PCA on the dataset of difference images, where each difference image comprised a point in R^n given by the brightness of the n pixels. The PCA basis images were the eigenvectors of the pixel by pixel covariance matrix (see Figure 4a), and the first p coefficients with respect to the new basis comprised the representation. Multiple ranges of components were tested, from $p = 10$ to $p = 200$, and performance was also tested excluding the first 1-3 components. Best performance of 79.3% correct was obtained with the first 30 principal components.

ICA: Representations such as PCA (eigenfaces), are insensitive to the high-order dependencies among the pixels, i.e., the obtained eigenfaces depend only on the pair-wise correlations between image pixels in the training database. Independent component analysis (ICA) is a generalization of PCA that is sensitive to the high-order dependencies between image pixels, not just pair-wise linear dependencies. We obtained an ICA representation for facial expression images using Bell & Sejnowski's infomax algorithm (Bell & Sejnowski, 1995, 1997) algorithm. Independent component analysis developed very different image representations from the other image processing techniques. The ICA representations were local and feature-like (see Figure 4b) while the PCA representations where more holistic (see Figure 4a). Unlike PCA, there is no inherent ordering to the independent components of the dataset. We therefore selected as an ordering parameter the class discriminability of each component, defined as the ratio of between-class to within-class variance. Best performance of 96% was obtained with the first 75 components selected by class discriminability. Class discriminability analysis of a PCA representation was previously found to have little effect on classification performance with PCA (Bartlett, Movellan, & Sejnowski, 2001). Of all the adaptive methods ICA gave the highest performance of 96% correct generalization to novel faces, whereas PCA, LFA, and FLD gave 79%

and 81%, and 76 % accuracy respectively.

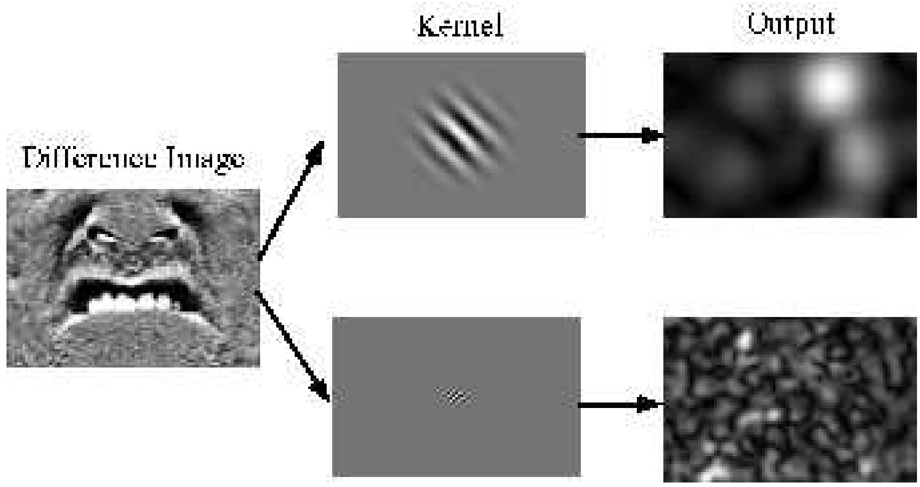


Figure 3: Example image decomposition. Here, a difference image of a lower face action is convolved with a family of Gabor wavelets. The output of the decomposition is channeled to the classifier.

Gabor wavelets: Gabor wavelets are 2-D sine waves modulated by a Gaussian envelope. They are shown to be good models of the receptive fields found in simple cells of the primary visual cortex (Daugman, 1988). We tested a representation which employed a family of Gabor wavelets at 5 spatial frequencies and 8 orientations (see Figure 4c). To provide robustness to lighting conditions and to image shifts we employed a representation in which the outputs of two Gabor filters in quadrature are squared and then summed. This representation is known as Gabor energy filters and it models complex cells of the primary visual cortex. Classification performance with the Gabor representation was 96%, matched only by the ICA representation.

Optic flow: Motion is an important source of information for facial expression recognition. We compared the image decomposition representations above to a motion- based representation. Facial motion was extracted using a correlation-based optic flow algorithm with sub-pixel accuracy (Singh, 1991). Recognition accuracy using motion alone was 86% correct.

Explicit feature measurement: We also examined a feature-based representation that measured facial wrinkles and the degree of eye opening. Wrinkles were measured using image gradients in specific locations, and eye opening was measured as the area of visible sclera lateral to the iris. Recognition with explicit feature measurements attained only 57% accuracy.

Overall Findings: Image decomposition with gray-level image filters outperformed explicit extraction of facial wrinkles or motion flow fields. Best performance was obtained with 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 agreements rates of expert human subjects on this set of images. The Gabor and ICA representations employed local filters, which supports recent findings that local filters are important for face image analysis (Padgett & Cottrell, 1997; Gray, Movellan, & Sejnowski, 1997;

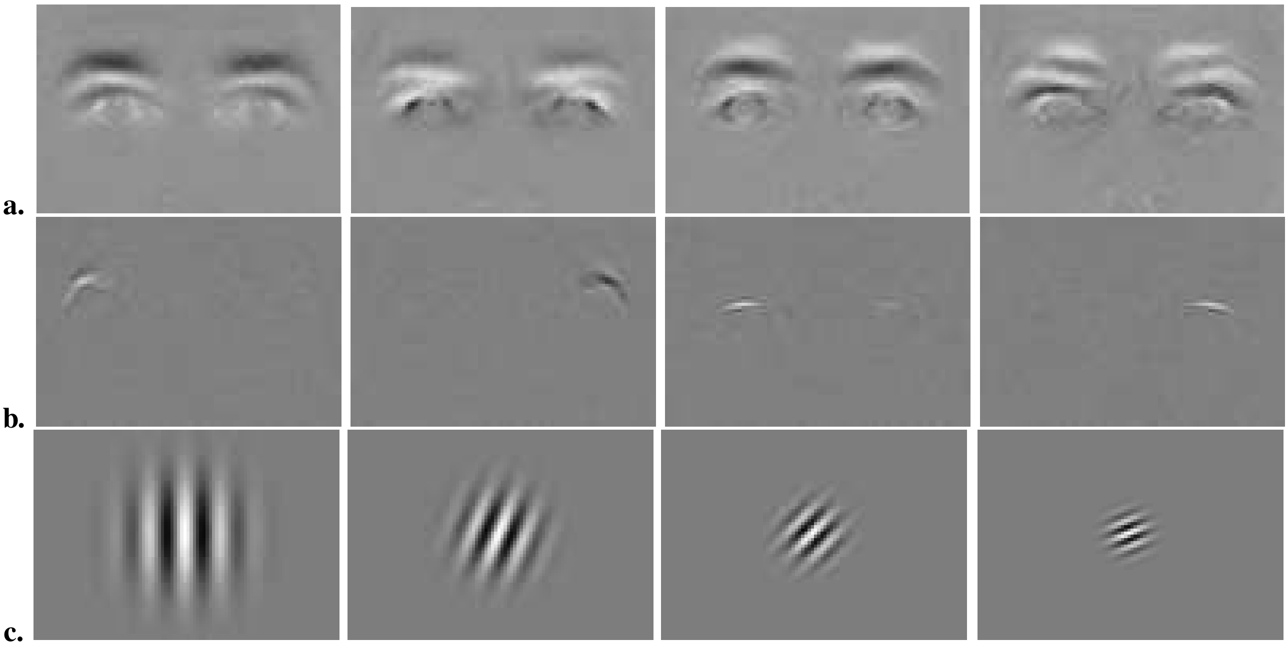


Figure 4: Sample image filters for the upper face. a. First four eigenfaces. b. Four independent component filters. The ICA filters are local, spatially opponent, and adapted to the image ensemble. c. Gabor wavelets.

Computational Analysis	Eigenfaces	79.3% ± 4
	Local Feature Analysis	81.1% ± 4
	Independent Component Analysis	95.5% ± 2
	Fisher's Linear Discriminant	75.7% ± 4
	Gabor Wavelet Decomposition	95.5% ± 2
	Optic Flow	85.6% ± 3
	Explicit Features	57.1% ± 6
Human Subjects	Naive	77.9% ± 3
	Expert	94.1% ±2

Table 1: Action unit recognition in novel (i.e., cross-validation) face images. Values are percent agreement with FACS labels on the database.

Lee & Seung, 1999). Yet, the local property alone was insufficient, as local filters based on second-order statistics (LFA and local implementations of PCA), did not perform significantly better than global PCA. Other properties shared by Gabor and ICA filters include sensitivity to high-order dependencies among the pixels (Field, 1994; Simoncelli, 1997), and relationships to visual cortical neurons (Daugman, 1988; Bell & Sejnowski, 1997). See Bartlett (2001) for a more detailed discussion.

3 Project Description

3.1 Image data

In this project, the two teams (UCSD and Pitt/CMU) attempted to recognize facial actions in spontaneous facial behavior. Paul Ekman and Mark Frank provided video tapes of 20 subjects from a deception study which included two conditions, named "Opinion" and "Theft". In the "Opinion" condition, subjects either had to lie or tell the truth about a strongly held opinion. This condition was not analyzed in this project. The "Theft" condition, which was the focus of this project, employed the following experimental paradigm: For half the subjects, a drawer contained $50, for the other half the drawer was empty. If the drawer contained money, subjects were told that they had the choice of taking it or not. They were told that afterwards they would have to convince an experimenter that either the drawer had no money or that they did not take it. They were told that if they managed to convince the experimenter, then they could keep the money. They were also told that if the experimenter thought they were lying, then they would have to be subjected to a very uncomfortable loud noise for 1 minute. Subjects were given a sample of the noise at the beginning of the experiment. Subjects caught lying were not actually punished.

Frank and Ekman (unpublished) found that deception could be reliably detected from facial actions scored by human FACS coders. The detection rate based on FACS codes was significantly higher than the detection rate of both naive human subjects and police officers watching the video.

Videotapes were digitized by Yaser Yacoob at the University of Maryland, and the image data was distributed to the two teams on hard disks. The image data consisted of 300 gigabytes of 640 x 480 color images. The video was digitized at 30 frames per second with 2:1 interlacing. The quality of the digitized images was relatively poor for current standards.

Approximately one minute of video was FACS coded for each subject. FACS codes were initially provided only for 10 of the 20 subjects, where the second 10 subjects were reserved for testing. Of the 10 subjects originally designated for training the computers, 7 were Caucasian, 2 were African American, and 1 was Asian. Three subjects wore glasses, one had facial hair, and one had another occluder (bandaid) on his nose. FACS codes were later provided for 7 additional subjects, consisting of 4 Caucasians, 1 African American, and 2 Asians. Two had facial hair and none wore glasses.

3.2 Technical challenges

This project is the first serious attempt to automatically code facial movement in spontaneous facial expressions. The previous work of both teams employed datasets of posed expressions collected under very controlled imaging conditions, with subjects deliberately facing the camera. Extending our systems to spontaneous facial behavior is a critical step forward, and a major contribution of this project to basic research.

As mentioned earlier in this document, spontaneous face data brings with it a number of technical issues that need to be addressed for computer recognition of facial actions. Perhaps the most important issues are: (1) The presence of out-of-plane head rotations as subjects nod or turn their heads as they interact with the interviewer or respond to a stimulus. This substantially changes the input to the computer vision systems, and it also produces variations in lighting as the subject alters the orientation of his or her head relative to the lighting source; (2) The presence of occlusions caused by out-of-plane head rotation, glasses and beards; (3) In spontaneous behavior, the facial actions are not temporally segmented, and may not begin with a neutral expression; (4) The low amplitude of spontaneous facial actions; (5) Coarticulation effects between facial actions, and between facial actions and speech related movements.

Very small sample size: An unexpected issue that compounded these technical challenges in the current project was the very small sample sizes available to the research teams. The machine learning algorithms employed by both the UCSD and Pittsburgh teams require a large number of examples of each facial action in order to accurately recognize them in novel subjects. Large sample sizes are particularly important when there is significant image variability such as that described in this section and in Section 3.1. Tables indicating the numbers of examples of each AU in the FACS codes provided by the Rutgers team are given in Appendix B. The numbers are very small. Hence the two teams agreed with Kathy Miritello and the two technical reviewers (Yaser Yacoob and Pietro Perona) to use the data from all 20 subjects[1] for development, and test performance using leave-one-out cross-validation (Tukey, 1958). This procedure maximizes the available data for training system parameters. In this procedure, data from all but one subject is used to estimate system parameters, and the remaining subject is used for testing. The parameters are deleted and re-estimated multiple times, each time leaving out a different subject for testing. Mean test performance provides an estimate of generalization performance on novel subjects. For technical reasons, data from 17, not 20 subjects was employed.

Even with leave-one-out cross-validation, small sample sizes continued to be an overriding issue. The two teams, in conjuction with the two technical reviewers, conferred in June to define some recognition tests for which there was sufficient data to train and test their systems. These tests are described in Section 4.

4 Comparison tests

The Pittsburgh and UCSD teams selected two tasks to test their systems. The tasks involve detection and discrimination of of action unit 45 (blinks), action units 1+2 (brow raise) and action unit 4 (brow lower). Figure 5 shows examples of each of these action units. The main criteria for selecting these tasks was the presence of a minimally sufficient number of examples for training the computer and the relevance to detecting psychological states such as alertness, anxiety, and surprise. It is important to emphasize that while these tests evaluate only a few basic facial movements, the goal of this project was not to develop systems that performed well on these specific tests. Both teams attempted to develop general purpose approaches to FACS recognition that could generalize to other facial actions, provided sufficient training data were available. It may be possible to develop ad-hoc procedures that capitalize on the specifics of detection of blinks and brow raises on this database. However, this was not the purpose of our work, for such ad-hoc procedures may not generalize well to other action units and other databases.

Examples in which the Rutgers coder and the Pittsburgh coder disagreed on the presence of the action unit were excluded from the tests. The tables in Appendix A give the quantities of each action unit coded by the Rutgers coder. The table shows, for example, that the Rutgers coder detected a viable number of

[1]For technical reasons, data from 17, not 20 subjects was provided. The amount of data could be more than doubled simply by providing FACS codes for the additional 3 subjects, plus the Opinion condition for all 20.

Figure 5: Example of action unit 45 (left), 1+2 (center) and 4 (right).

examples of action unit 7. The Pittsburgh coder, however, disagreed with most of the examples of action unit 7, and hence Action Unit 7 was not considered for the preliminary tests.

For each action unit, the Pittsburgh coder defined a sequence window containing the beginning and end of the action, where the first and last frame of the sequence was as close to neutral as possible. Often, this was one frame on either side of the beginning and end of the AU, but sometimes it was not as simple as that. Please refer to the report by the Pittsburgh group for more information. The Pittsburgh and UCSD teams agreed to use the first frame of each sequence if their systems required a neutral frame.

1. Blinks

   Blink detection (AU 45) was selected as a basic test of the ability of the computer systems to classify a simple facial movement in this highly variable imaging environment. Blinks were chosen because of their relevance to applications such as monitoring of alertness and anxiety, and because there was significantly more training data for blinks than for any other facial action. The positive examples consisted of every sequence containing a blink for which the Rutgers and Pittsburgh coders agreed a blink was present. Some of these sequences contained multiple blinks or “flutters” without a return to neutral. There were 184 blinks occurring in 168 sequences. The negative samples for this task consisted of 168 randomly selected sequences matched by subject and length. The only criteria was that the negative sequence did not contain a blink. Mean length of the blink sequences was 12 frames with the peak occuring on frame 4. The human subject composition of this dataset consisted of 7 Caucasians, 2 African Americans, 1 Asian, 3 subjects with glasses, one with facial hair, and one with a band-aid on his nose.

2. Brows

   The brow region contained contained facial actions of importance for monitoring psychological states, and for which we had a reasonable amount of training data. The two teams converged on a three category tests related to the brow region:

   (A) The first category, which we named “brow raise” contained combinations of AU1 and AU2. There were 48 examples of brow raises from 12 subjects for which the Rutgers and Pittsburgh coders agreed. Included in this category is any sequence containing a AU1+2, regardless of co-occurring actions. There were 38 examples of AU1+2 alone performed by 10 subjects, 9 examples of AU1+2+5 performed by 4 subjects, and 1 example of AU1+2+5+7. Mean length of the brow raise sequences was 26 frames, with the peak occuring on frame 9. The human subject composition of this dataset consisted of 7 Caucasians, 3 African Americans, 2 Asians, 4 subjects with glasses, and 2 with facial hair.

(B) The second category was called "brow lower" and consisted of sequences containing AU4 (brow furrow) and or strong AU9 (nose wrinkle, which also lowers the brows). There was a total of 14 examples of AU 4 and/or strong AU9 for which both coders agreed, and these were performed by 9 subjects. There were 8 examples of AU4 alone, 1 example of AU9 alone, 1 example of AU4+9, 2 examples of AU4+1, 1 example of AU4+5, and 2 examples of AU4+7. There were no examples of AU1+2+4 on which the Rutgers and Pittsburgh coders agreed. Mean length of the brow lower sequences was 28 frames, with the peak occuring on frame 14. The human subject composition of this dataset consisted of 5 Caucasians, 3 African Americans, 1 Asian, 1 subject with glasses, and 2 with facial hair.

(C) The third category consisted of randomly selected sequences which did not contain images of categories A and B. These sequences were matched by subject and length to the sequences in categories A and B. A total of 62 random sequences were selected.

5 Technical Approach

We identified out-of-plane rotations as the most important technical challenge for application of our previous research methods to this database. Our approach applies statistical methods directly to images or filter bank image representations. While in principle such methods may be able to learn the invariances underlying out-of-plane rotations, in practice the amount of data needed to learn such invariances would be beyond the scope of this project. Instead, we decided to reduce the effect of this source of variability by using 3D models. The idea is to fit 3D models to the image plane, texture those models and warp them into canonical views (e.g., frontal views) and face geometries. The approach can be characterized as 3D face warping, and it is a generalization of the 2D warping approach we used in our previous work (Bartlett, Viola, Sejnowski, Larsen, Hager, & Ekman, 1996; Movellan, 1995).

While we believed this approach had good chances for success, we also identified reasons for its potential failure: (1) Automatic estimation of 3D pose from 2D images may not feasible with today's technology; (2) Even if 3D pose estimation is possible, the distortions introduced by the warping process may still overwhelm the statistical classifiers; (3) Even if the warping process does not introduce major distortion, using high dimensional image data, instead of parameterized contour models, may overwhelm the statistical classifiers, specially for spontaneous facial actions which are known to have a relatively low signal to noise ratio.

Based on the fact that we had limited time and human resources we decided to concentrate most of our work on answering questions (2) and (3). Once it became clear that the approach was promising, one member of our group started working on the problem of real time, automatic estimation of 3D pose.

5.1 3D pose estimation

First we developed a system for estimating face geometry and 3D pose from hand-labeled feature points. This allowed us to test the feasibility of our approach before allocating the resources required to tackle the problem of fully automated 3D pose estimation. In addition the labeled images were used to provide ground truth for training automatic feature tracking and automatic 3D pose tracking algorithms. In parallel, one member of our group is on an extended visit to Matthew Brand at MERL, who recently developed a state of the art method for recovering 3D models from 2D image sequences (Brand, 2001).

When landmark positions in the image plane are known, the problem of 3D pose estimation is relatively easy to solve. While deterministic algorithms are possible, we decided to use a stochastic filtering approach

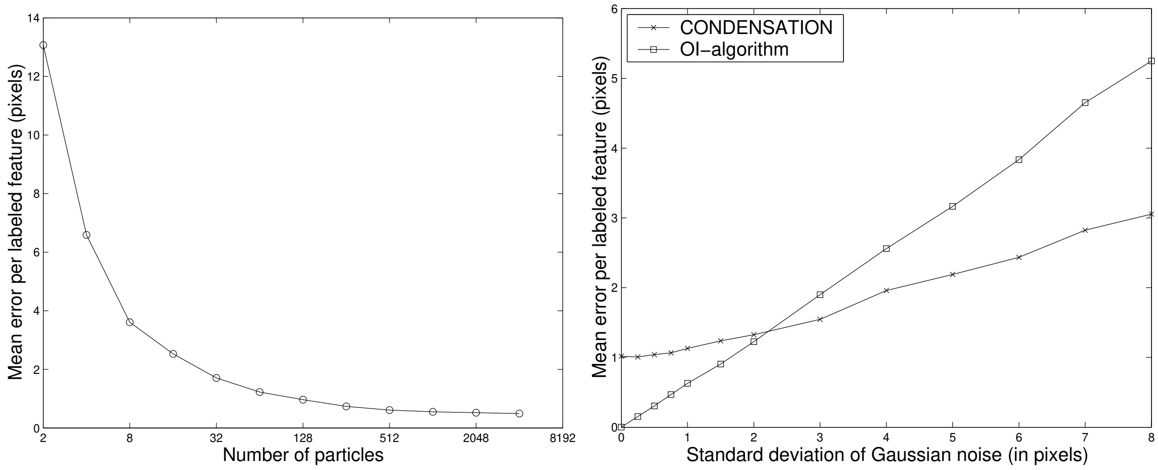


Figure 6: On the left, the performance of the particle filter is shown as a function of the number of particles used. On the right the performance of the particle filter and the OI algorithm as a function of noise added to the true positions of features.

that could be easily modified for the case in which landmark positions are unknown, and are estimated simultaneously. We did some preliminary studies and found that when uncertainty is introduced in the feature positions, the stochastic filtering approach performed significantly better than the most robust deterministic algorithms (Figure 6).

5.1.1 Estimation of Face Geometry

We start with a canonical wire-mesh face model (Pighin, D, Szeliski, & Salesin, 1998) which is then modified to fit the specific head-shape of each subject. To this effect 30 images are selected from each subject to estimate the face geometry and the position of 8 features on these images is labeled by hand (ear lobes, lateral and nasal corners of the eyes, nose tip, and base of the center upper teeth). Based on those 30 images the 3D positions of the 8 tracked features is recovered using standard computer vision techniques. A scattered data interpolation technique (Pighin et al., 1998) is then used to modify the canonical face model to fit the 8 known 3D points and to interpolate the positions of all the other vertices in the face model whose positions are unknown. In particular, given a set of known displacements $\mathbf{u}_i = \mathbf{p}_i - \mathbf{p}_i^0$ away from the generic model feature positions $\mathbf{p}_i^0$, we computed the displacements for the unconstrained vertices j. We then applied a smooth vector-valued function $f(\mathbf{p})$ that we fit to the known vertices $\mathbf{u}_i = f(\mathbf{p}_i)$ from which we can compute $\mathbf{u}_j = f(\mathbf{p}_j)$. Interpolation then consists of applying

$$f(\mathbf{p}) = \sum_i \mathbf{c}_i \phi(||\mathbf{p} - \mathbf{p}_i||) \tag{1}$$

to all vertices p in the model, where ϕ is a radial basis function. The coefficients $\mathbf{c}_i$ are found by solving a set of linear equations that includes the interpolation constraints $\mathbf{u}_i = f(\mathbf{p}_i)$ and the constraints $\sum_i \mathbf{c}_i = 0$ and $\sum_i \mathbf{c}_i \mathbf{p}_i^T = 0$.

5.1.2 Particle filters

3-D pose estimation can be addressed from the point of view of statistical inference. Given a sequence of image measurements $O = (O_1, \cdots, O_t)$, a fixed face geometry and camera parameters, the goal is to find the most probable sequence of pose parameters $S = (S_1, \cdots, S_t)$ representing the rotation, scale and translation of the face on each image frame. In probability theory the estimation of S from O is a known "stochastic filtering". The main advantage of probabilistic inference methods is that they provide a principled approach to combine multiple sources of information, and to handle uncertainty due to noise, clutter and occlusion. Markov Chain Monte-Carlo methods provide approximate solutions to probabilistic inference problems which are analytically intractable.

We explored a solution to this problem using Markov Chain Monte-Carlo methods, also known as condensation algorithms or particle filtering methods, (Kitagawa, 1996; Isard & Blake, 1998). Our approach worked as follows. First the system is initialized with a set of n particles. Each particle is parameterized using 7 numbers representing a hypothesis about the position and orientation of a fixed 3D face model: 3 numbers describing translation along the X, Y, and Z axes and 4 numbers describing a quaternion, which gives the angle of rotation and the 3D vector around which the rotation is performed. Since each particle has an associated 3D face model, we can then compute the projection of 8 facial feature points in that model onto the image plane. The likelihood of the particle given an image is assumed to be an exponential function of the sum of squared differences between the actual position of the 8 features on the image plane and the positions hypothesized by the particle. In future versions this likelihood function may be based on the output of feature detectors and/or optic flow algorithms. At each time step each particle "reproduces" with probability proportional to the degree of fit to the image. After reproduction the particle changes probabilistically in accordance to a face dynamics model, and the likelihood of each particle given the image is computed again. It can be shown (Kitagawa, 1996) that as $n \to \infty$ the proportion of particles in a particular states at a particular time converges in distribution to the posterior probability of the state given the image sequence up to that time

$$\lim_{n \to \infty} \frac{n_t(x)}{n} = P(S_t = x \mid O_1, \cdots, O_t) \tag{2}$$

where $n_t(x)$ represents the number of particles in state x at time t. The estimate of the pose at time t is obtained using a weighted average of the positions hypothesized by the n particles.

We compared the particle filtering approach to pose estimation with a recent deterministic approach, known as the orthogonal iteration (OI) algorithm (Lu, Hager, & Mjolsness,), which is known to be very robust to the effects of noise.

Performance of the particle filter was evaluated as a function of the number of particles used. Error was calculated as the mean distance between the projected positions of the 8 facial features back into the image plane and ground truth positions obtained with manual feature labels. Figure 6 (Left) shows mean error in facial feature positions as a function of the number of particles used. Error decreases exponentially, and 100 particles were sufficient to achieve 1-pixel accuracy (similar accuracy to that achieved by human coders).

A particle filter with 100 particles was tested for robustness to noise, and compared to the OI algorithm. Gaussian noise was added to the positions of the 8 facial features. Figure 6 (Right) gives error rates for both pose estimation algorithms as a function of the variance of the Gaussian noise. While the OI algorithm performed better when the uncertainty about feature positions was very small (less than 2 pixels per feature). The particle filter algorithm performed significantly better than OI for more realistic feature uncertainty levels.

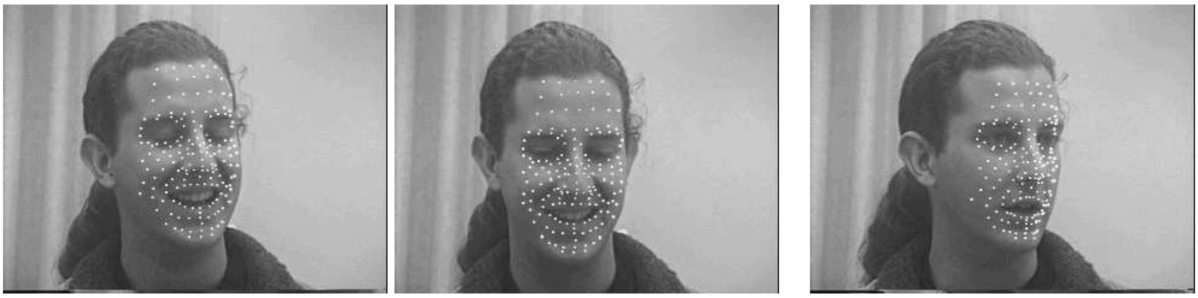


Figure 7: A demonstration of Brand's real time 3D face tracking method at work on Subject 19.

5.1.3 Fully automated real time 3D pose tracking

The FACS classification results presented here are based on the output of the 3D pose estimation system which employs hand-labeled feature points. However the particle filtering approach extends in a principled way for easy integration with state of the art methods for real time 3D tracking. In particular, one member of our team, is making an extended visit to Matthew Brand's laboratory at MERL with the goal of merging the particle filtering approach with his recently published method for real time 3D face tracking (Brand, 2001). Figure 7 shows an example of the current prototype at work on one of Subject 19 from the current database.

5.2 Generation of image representation.

Once 3D pose was estimated, faces were rotated to frontal and warped to a canonical face geometry. Image warping was performed using the interpolation technique described in Section 5.1.1. It has been shown that for face recognition, methods such as eigenfaces give superior performance when faces are warped to a canonical geometry (Wechsler, Phillips, Bruce, Fogelman-Soulie, & Huang, 1998). For expression recognition, it may be even more important to remove variations in face shape. An example output of the face rotation and warping system is shown in Figure 8.

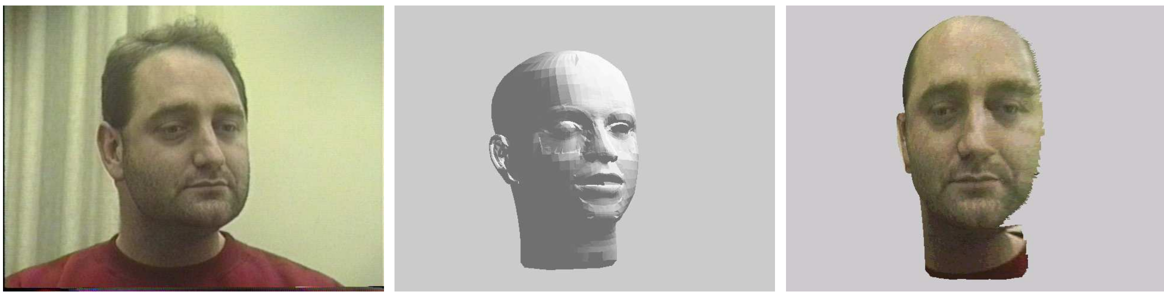


Figure 8: Original image, model in estimated pose, and warped image.

Images were scaled and cropped automatically from the output of the pose estimation and warping system. Images were cropped to 192 x 132 with 105 pixels between the eyes. The vertical position of the eyes was 0.67 of the window height.[2] Pixel brightnesses were linearly rescaled to [0,255], and passed

[2]The resulting window contained both the left and right sides of the upper face. Lateralized FACS codes could be obtained by

through a soft histogram equalization performed using a logistic filter with parameters chosen to match the mean and variance of the pixel values in the neutral frame. Difference images were obtained by subtracting a neutral expression frame from the subsequent frames in each sequence.

Our previous work demonstrated that Gabor wavelet filters and independent component analysis (ICA) were the most effective forms of image representation among those we tested for facial action recognition. The difference images were passed through the Gabor filters described in Section 2. A bank of Gabor filters at 5 spatial frequencies and 8 orientations was employed. The amplitude of the output was calculated and passed to the classifier. Figure 9 shows the user interface of our current system. The system works on MSWindows and Linux environments.

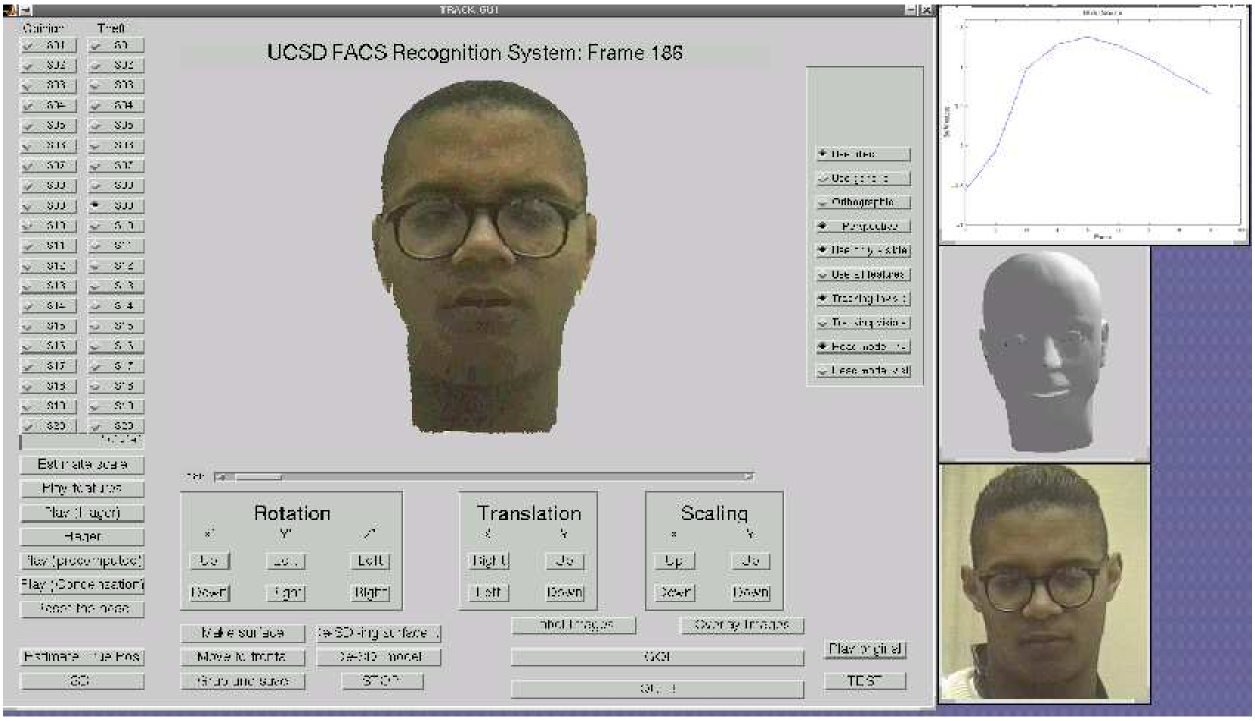


Figure 9: User Interface for the UCSD FACS Recognition System. The curve shows the output for the blink detector during the relaxation phase of a blink. The image shown was 5 frames past the peak.

5.3 Classification.

Once the image representation is obtained it is channeled to a statistical classifier. In the current version we are using support vector machines. Support vector machines, introduced by V. Vapnik in 1992, are proving to be excellent classifiers for a very wide variety of problems. By effectively embedding vectors in a higher dimensional space, where they may be linearly separable, SVM's can solve highly non-linear problems. The objective of support vector machines is to find a hyperplane that separates the data correctly into classes with as much distance from that plane as possible. The training vectors closest to the optimal hyperplane are called the support vectors. SVMs tend to achieve very good generalization rates when compared to other classifiers because they focus on these maximally informative exemplars, the support vectors. The

cropping quadrant windows instead.

computational complexity of an SVM depends on the number of training examples, but not on the dimension of the embedding space or on the dimensions of the original vectors. This is well suited to many expression recognition problems, which tend to involve relatively small sets of training samples (order hundred), while each mage representation vector may be quite long (order million dimensions). The dimension of the Gabor representation, for example, is the number of pixels times the number of filters.

Support vector machines were trained and tested on the AU peak frames[3], as identified by the Pittsburgh FACS coder. Gabor representations comprised the input to the SVM's. We also examined performance of the SVM's applied directly to the difference images, without the Gabor filters.

The output of the classifiers was then fed to a bank of hidden Markov models, to take advantage of dynamic information in the output of the classifier. Hidden Markov models were trained and tested on full sequences, without information about the location of the AU peak. We compared performance of HMM's trained on the outputs of the SVM's to HMM's trained directly on Gabor representations.

6 Research Questions

We designed a series of tests to explore the following research questions:

1. **Image alignment:** It was an open question whether 3D pose estimation and warping would in fact improve recognition performance over 2D alignment methods, or whether noise introduced in the process would interfere with recognition. We tested : (1) A 3D face model with geometry adapted to each subject based on 8 anchor points; (2) A plane model, which is equivalent to the 2D warping approach used in our previous research. Future experiments will test other models, such as cylinders, or ellipsoids.

2. **Image representation:** We compared representations based on: (1) Raw pixel levels; (2) Gabor energy filters. We also experimented with different spatial frequency ranges for the Gabor filters. Future experiments will test alternative representations, such as ICA and PCA.

3. **Classifier:** Here we examined HMM trained on the outputs of SVM's and trained on Gabor filter bank representations. We also examined HMM's trained directly on dimensionality-reduced Gabor representations rather than SVM outputs. Future tests will include diffusion networks (a recent competitor of HMMs), Bayes point Machines (a recent competitor of SVMs), multilayer neural networks, and standard nearest-neighbor classifiers.

7 Results

All results reported in this section are for generalization to novel subjects, tested using leave-one-out cross-validation. The results are summarized in Table 4.

7.1 Blinks

Action unit dynamics were incorporated using Hidden Markov Models. Hidden Markov models were applied in two ways: (1) Taking Gabor representations as input, and (2) Taking SVM outputs as input.

[3]Note that this task is different from discrimination when the position of the peak frames is unknown, which is the task solved by the HMMs.

HMMs on Gabors: Hidden Markov models were trained on Gabor representations of the Blink and Non-Blink sequences, using methods similar to those in Bartlett et al. (2000). All difference images in each sequence were convolved with Gabor filters at 5 spatial frequencies and 8 orientations. Because the dimension of the Gabor output is too great for training HMM's, the Gabor output was first reduced to 100 dimensions per image using PCA. Six sequences per subject, 3 blinks and 3 non-blinks, were employed for generating the principal component eigenvectors. PCA was performed for each of the 5 spatial frequencies separately, which has been shown to be more effective for facial expression analysis than performing PCA on all of the outputs together (Cottrell, Dailey, Padgett, & R, 2000). The first 20 principal components for each spatial frequency were retained, producing a 100-dimensional representation vector for each image.

Two hidden Markov models, one for Blinks and one for NonBlinks, were trained and tested using leave-one-out cross-validation. A mixture of Gaussians model was assumed. Test sequences were assigned to the category for which the probability of the sequence given the model was greatest. The number of states was varied from 1-10, and the number of Gaussians was varied from 1-7. Best performance of 95.7% correct was obtained using 5 states and 3 Gaussians.

Including the first derivative of the observations in the input to the HMM's has been shown to improve recognition performance for speech (Rabiner, 1989) and lipreading (Movellan, 1995). The HMM's were trained again, this time with 200 dimensions in the input consisting of the 100 Gabor dimensions and their first derivative. The first derivative was approximated by $d(t) = (O(t+1) - O(t-1))/2$. Classification performance on this observation set was 89.9% using 1 state and 1 Gaussian. The slight reduction in performance may be due to the large number of parameters to be estimated for an HMM with 200 input dimensions.

HMMs on SVM outputs: SVM's were first trained to discriminate blinks from non-blinks in individual frames. A nonlinear SVM applied to the Gabor representations 94.3% correct generalization performance for discriminating blinks from non-blinks when using the peak frames. The nonlinear kernel was of the form $\frac{1}{k+d^2}$ where d is Euclidean distance, and k is a constant. Here $k = 4$. Consistent with our previous findings (Littlewort-Ford, Bartlett, & Movellan, 2001), Gabor filters made the space more linearly separable than the raw difference images. A linear SVM on the Gabors performed significantly better (93.5%) than a linear SVM applied directly to difference images (78.3%). Nevertheless, a nonlinear SVM applied directly to the difference images performed similarly (95.9%) to the nonlinear SVM that took Gabor representations as input.

Figure 10 shows the time course of SVM outputs for Blinks and Non Blinks. The SVM output was the margin (distance along the normal to the class partition). All time courses shown are "test" outputs. The SVM was trained on the other 9 subjects, and the output shown is that for the test subject. A sample distribution from a single subject is shown on the top of Figure 10, and a sample distribution from multiple subjects is shown on the bottom. Although the SVM was not trained to measure the amount of eye opening, it is an emergent property. Figure 11 shows the SVM trajectory when tested on a sequence with multiple peaks.

For each example from the test subject in the leave-one-out cross-validation, the output of the SVM was obtained for the complete sequence. This produced SVM output sequences for 10 "test" subjects. HMM's were then trained and tested using leave-one-out cross validation on these "test" outputs. The HMM's gave 98.2% generalization performance for discriminating blinks from non-blinks, using 6 states and 7 Gaussians. These results were obtained using the outputs of a nonlinear SVM trained on difference images. We are presently testing whether performance will improve when the SVM takes Gabors as input.

The HMM trained on these trajectories missed only one blink. The peak image of the miss is shown

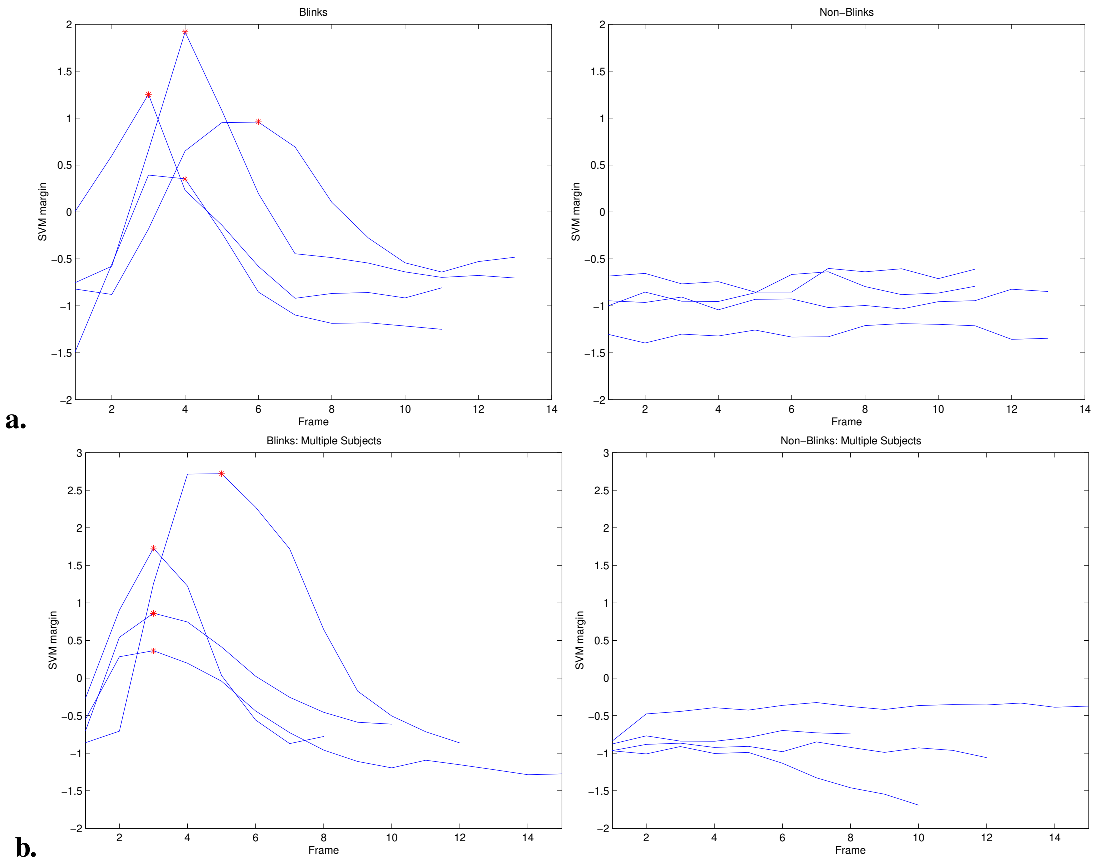


Figure 10: a. Blink and non-blink trajectories of SVM outputs for one subject (Subject 6). These are outputs of a nonlinear SVM trained on difference images. Star indicates the location of the AU peak as coded by the human FACS expert. b. Blink and non-blink trajectories of SVM outputs for four different subjects.

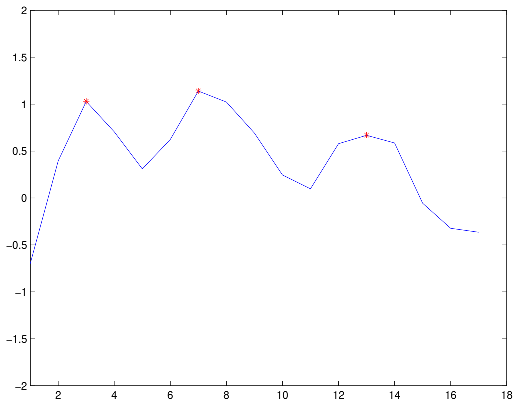


Figure 11: SVM output trajectory of a blink with multiple peaks (flutter).

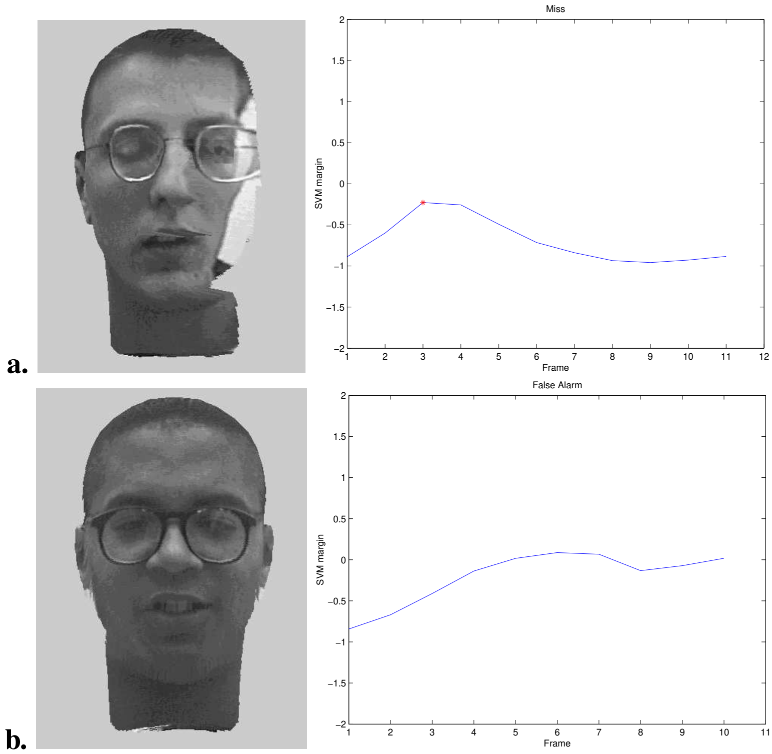


Figure 12: a. A blink incorrectly classified as a non-blink by the HMM. The image shown is the peak frame. The star on the SVM trajectory indicates the location of the peak as coded by the human FACS expert. b. A non-blink incorrectly classified as a blink by the HMM. The image shown is the 6th frame in the sequence.

in Figure 12a. Note that one eye is open. In the present implementation, the system tested both eyes simultaneously. By altering the cropping, it is possible to train and test on each eye individually. For this particular database, it may have been possible to obtain better performance by using the subjects' right eye only, as the camera was situated to the subjects' right. However, we chose not to do that as our goal is to develop and test general systems that can be applied to any database. An example false positive is shown in Figure 12b. Figure 12 also provides an example of an image rotation with a poor (a) and good (b) fit to the geometric face model.

7.2 Brows

HMMs on Gabors: HMM's were trained on the 3-category decision (Brow Raise vs. Brow Lower vs. Random matched sequences) taking Gabor representations of the brow images as input. All difference images in each sequence were convolved with Gabor filters at 5 spatial frequencies and 8 orientations. Dimensionality was then reduced to 100 by taking the first 20 principal components of each spatial frequency. The PCA-reduced Gabors comprised the input to 3 HMM's, one for each of the three brow categories. Best performance of 80.6% was obtained using 2 states, 2 Gaussians, and first derivatives. Inspection of the confusion matrix, however, revealed that very few images were classified in the Brow Lower category. We therefore adjusted the decision criteria to increase the likelihood that the system would output Brow Lower. This was performed by creating bias weights for the likelihoods of the three models. The bias weights were chosen to maximize the minimum accuracy over the three categories. This reduced the performance accuracy over all examples down to 70.16%. The confusion matrix for these HMM's is given in Table 3. If the Brow Lower category was omitted, performance raised to 90.91%.

HMMs on SVM outputs: 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 versus matched random sequences (A vs. C), Brows Lower versus another set of matched random sequences (B vs. C), and Brow Raise versus Brows Lower(A vs. B). The output of these three SVM's was then fed to a bank of HMMs for classification. SVM theory can be extended in various ways to perform multiclass decisions (e.g. (Lee, Lin, & Wahba, 2001)). In future work, a multiclass SVM will be included as input to dynamic models as well.

Three hidden Markov models were trained, one for each of the three classes. The input to each HMM consisted of three values: the outputs of one SVM for each of the three 2-category decisions described above. For A vs. C we used the nonlinear SVM trained on difference images; For B vs. C we used the nonlinear SVM trained on Gabors; and For A vs. B we used the linear SVM trained on difference images. At the time, these were the best performing SVM's for each task. As before, 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 and first derivatives. Inspection of the confusion matrix, however, revealed again that very few images were classified in the Brow Lower category. With bias weights, the performance accuracy over all examples was 66.94%. The confusion matrix is given in Table 2. If the Brow Lower category was omitted, performance raised to 89.53%.

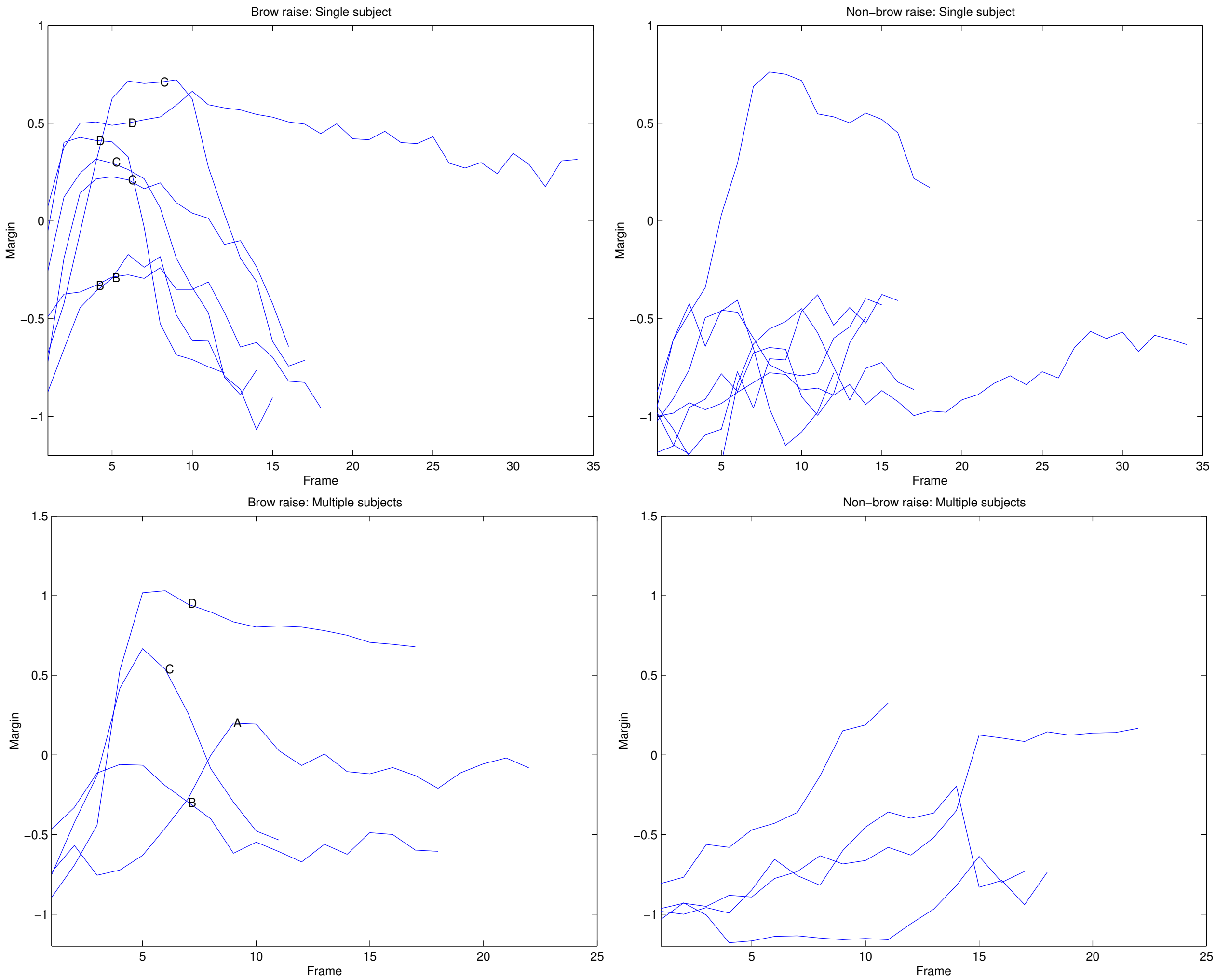
	Automated System		
Human Coder	Brow Raise	Brows Lower	Matched Random Sequences
Brow Raise	39	5	4
Brow Lower	2	6	6
Matched Random Sequences	5	19	38

Table 2: Confusion matrix for HMM trained on SVM outputs. Overall agreement = 66.94 %. Agreement omitting Brow Lower = 89.53 %.

	Automated System		
Human Coder	Brow Raise	Brow Lower	Matched Random Sequences
Brow Raise	35	6	7
Brow Lower	0	7	7
Matched Random Sequences	1	16	45

Table 3: Confusion matrix for HMM trained on PCA-reduced Gabors. Overall agreement = 70.16 %. Agreement omitting Brow Lower = 90.91 %.

Brow movement trajectories: Figure 13 shows example output trajectories for the SVM trained to discriminate Brow Raise from Random matched sequences. Trajectories for an individual subject are shown in (a) and for multiple subjects in (b). The output trajectories for the brows were noisier than for the blinks. Nevertheless 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 ($r = .43, t(42) = 3.1, p = 0.0017$). Figure 13a shows an example false positive trajectory. This false positive was due to vertical alignment error resulting from human error in marking feature positions. When a difference image is generated, a vertical alignment error will produce an image very similar to one produced by a brow raise.



a.

b.

Figure 13: a. Brow raise and non-brow raise trajectories of SVM outputs for one subject (Subject 20). These are outputs of a nonlinear SVM trained on difference images. Letters A-D indicate the intensity of the AU as coded by the human FACS expert, and are placed at the peak frame. Note the false positive. b. Brow raise and non-brow raise trajectories of SVM outputs for four different subjects.

Figure 14 shows sample output trajectories for an SVM trained to discriminate Brow Lower from Random matched sequences. The output trajectories are noisier for Brows Lower than for the Brow Raise category. The SVM outputs are also less predictive of AU intensity. There may have been insufficient data to learn this property given the very small sample size (14 sequences), relative to the amount of variability in the data due to factors such as the presence of glasses, variety of races, variety of facial actions (4, 1+4, and 9), variety of action intensities, and alignment error.

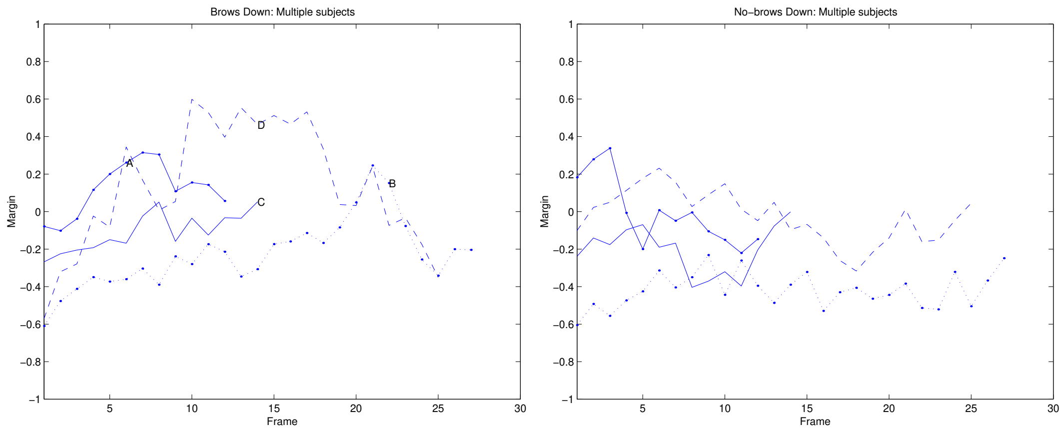


Figure 14: a. Brows lowered and non-brow lowered trajectories of SVM outputs for four subjects. These are outputs of a nonlinear SVM trained on Gabors. Letters A-D indicate the intensity of the AU as coded by the human FACS expert, and are placed at the peak frame.

Discriminating Brow Raise from Random matched sequences: A system was specifically trained to discriminate Brow Raise from Non-brow motion. HMM's were trained on the outputs of a nonlinear SVM applied to Gabor images. Best performance off 90.6% correct was obtained with 3 states and 1 Gaussian, and including first derivatives of the observation sequence.

Discriminating Brow Lower from Random matched sequences: A second system was specifically trained to discriminate Brow Lower from Non-brow motion. HMM's were trained on the outputs of a nonlinear SVM applied to Gabor images. Best performance off 75.0% correct was obtained using 1 state, 4 Gaussians, and first derivatives.

Discriminating Brow Raise from Brow Lower We also tested a system specifically trained on discriminating Brow Raise (AU1+2) from Brow Lower (AU4). HMM's were trained on the outputs of a linear SVM applied to difference images. Best performance of 93.5% correct was obtained using 9 states, 6 Gaussians and first derivatives. Given knowledge that a brow movement has occurred, it appears to be relatively easy to decide whether the movement was up or down.

7.3 SVM's on 2D aligned images

In order to evaluate the benefit, or cost, of the 3D rotation, 2-D aligned images were also generated. These images were rotated in the plane so that the eyes were horizontal, and then cropped and scaled identically to the 3D rotated images. In the 2D aligned images, the aspect ratio was adjusted so that there were 105 pixels between the eyes and 120 pixels from eyes to mouth.

	HMM		N
	on Gabor	on SVM	
Blink vs. Non-blink	95.7	98.2	336
Brow Raise vs. Matched Random Sequences	-	90.6	96
Brow Lower vs. Matched Random Sequences	-	75.0	28
Brow Raise vs. Brow Lower	-	93.5	62
Brow Raise vs. Lower vs. Random	80.6	78.2	124
Brow Raise vs. Lower vs. Random, with bias	70.2	66.9	124

Table 4: Summary of results. All performances are for generalization to novel subjects. N: Total number of positive and negative examples.

A nonlinear SVM taking difference images as input obtained 95.5% accuracy for detecting blinks from non-blinks. This was identical to performance using the 3-D rotations. For blink detection using SVM's, nothing appears to have been gained or lost by performing 3D alignment.

Detection performance for Brow Raise (A vs. C) using 2D aligned images was 83.3%. This performance was obtained using a nonlinear SVM applied to difference images. The same SVM applied to 3D rotated images performed at 88.5% correct. The 3D rotation and warping significantly improved the detection of brow raises.

7.4 Spatial frequency

There appeared to be a range of spatial frequencies for the Gabor filters that gave best performance with the SVM's. We experimented with sets of 5 spatial frequencies separated by 1/2 octave steps. For the brow raises, the most robust performance was obtained using 6-24 cycles per facewidth, where facewidth is measured temple to temple. Performance was the same for 3-12 cycles per facewidth using nonlinear SVM's, but dropped for the linear SVM's. Performance for both linear and nonlinear SVM's dropped slightly for the lowest spatial frequency range of 1.5-6 cycles per facewidth, and dropped significantly for the higher spatial frequency range of 12-48 cycles per facewidth. For the blinks, performance peaked at 3-12 cycles per facewidth.

7.5 Precision of the feature position labels

The results presented here rely on a 3D pose estimation stage, which required knowledge of the position 8 facial features. As mentioned earlier the technology already exists for automatic real-time 3D pose estimation, and we are in the process of integrating one such tracker into our system. One might question whether the 3D alignment for the tests presented here is unrealistically precise because of the use of hand labels. However it is of interest to note that the accuracy of the hand labeled points was in fact very poor. Our undergraduate assistants often failed to move the mouse when the head moved, resulting for example, in eye corners marked on the subject's cheek. The assistants waited until finals week to do most of their feature marking, and it is well known in experimental psychology that human data collected under such conditions tends to be unreliable. We assessed the reliability of these labels by carefully labeling a sample of images. The mean deviation between the two sets of labels was 4 pixels, with a standard deviation of 8.7 pixels. Note the high standard deviation. We attribute this lack of reliability to human inattention rather than human error, as the variability in feature positions when humans are attending to the task is certainly much smaller. Based on preliminary work we anticipate performance to improve when the automatic 3D tracker is used.

8 Discussion

Earlier in this document we identified the following issues that needed to be addressed by the next generation of automatic FACS recognition systems: (1) The presence of out-of-plane head rotations; (2) The presence of occlusions caused by out-of-plane head rotation, glasses and beards; (3) Video segmentation. In spontaneous behavior, the facial actions are not temporally segmented, and may not begin with a neutral expression; (4) The low amplitude of spontaneous facial actions; (5) Coarticulation effects between facial actions, and between facial actions and speech related movements.

The results presented here are an important step towards the solution of these problems. We showed that 3D tracking and warping followed by machine learning techniques directly applied to the warped images, is a viable and promising technology that may be capable of addressing all these issues. This system employed general purpose learning mechanisms that can be applied to recognition of any action unit. The approach is parsimonious and does not require defining a different set of feature parameters or image operations for each facial action.

On tests for recognition of some basic facial movements, this system attained 98% accuracy for detecting blinks in novel subjects, 91% accuracy for brow raises, 94% accuracy for discriminating brow raises from lowering of the brows, and 80% accuracy for a 3-alternative decision between brow raise, brow lower, and randomly selected sequences containing neither.

One exciting aspect of the approach presented here is the fact that important measurements emerged out of filters which were derived from the statistics of images. For example, the output of our blink detector could be potentially used to measure the amount of eyelid closure, even though the system was not designed to explicitly detect the contours of the eyelid and measure the closure (Figure 11).

While the database we used was rather large for current digital video storage standards, in practice the number of examples for each action unit in the the database was relatively small. This was the main reason why we could only prototype the system on the 3 action units which had a reasonable number of examples: AU 45, for which we had 168 examples, AU1+2, for which we had 48 examples, and AU4, for which we had 14 examples. Inspection of the performance of our system shows that 14 examples was sufficient for the system to successfully learn an action, and an order of 50 examples was sufficient to achieve performance over 90%. Appendix B examines the AU frequency of 10 subjects from the Theft database, one minute of coded video each. Based on these frequency tables, a database 50 times larger (e.g. 500 subjects, one minute per subject) would provide at least 50 examples of all upper face AU's, all eyelid actions, and the majority of the lower face actions. While progress can certainly be made with smaller databases (e.g. 100 subjects), the history of automated speech recognition, handwriting recognition, and face recognition has shown that large, shared databases can catalyze the field and lead to major advances in technology (Cole, 2000).

Due to time and funding limitations there are still important issues that we did not address in this project. In particular we tested our system on segmented action units rather than unsegmented video sequences; and we tested our system on upper face action units, thus avoiding issues of coarticulation with speech. However the outputs of the filters developed here appear to provide a very good signal that could be used for unsegmented video sequences. We do not know yet how difficult it will be to address the effects of speech movements on lower face action units. This may be a problem for which the automatic lip-reading community (for which visual speech is signal and expressions are noise) and the expression recognition community (for which visual speech is noise and expressions are signal) may be fruitfully collaborate.

We believe all the pieces of the puzzle are ready for the development of automated systems that recognize spontaneous facial actions at the level of detail required by FACS. One important lesson learned from the speech recognition community is the need for shared and realistic databases. 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 from the research community with very high certainty. Development of these databases is an important priority that will require a joint effort from the computer vision, machine learning and psychology communities.

9 Appendix A: Additional Studies

In this section we present additional research and development which was directly related to this project but was not part of the comparative tests described in Section 4 or used to generate the results in Section 7. Several of these studies were conducted while we were waiting for the FACS codes on the Theft database to be finalized, and were carried out on different image databases. This section summarizes work described in the following papers. These papers are available on request, or can be downloaded from http://mplab.ucsd.edu/tr.html.

- Fasel, I.R., Bartlett, M.S., and Movellan, J.R. A Comparison of Gabor Filter Methods for Automatic Detection of Facial Landmarks. UCSD MPLab TR 2001.04, June 2001.
- Smith, E. A SNoW-Based Automatic Facial Feature Detector. UCSD MPLab TR 2001.06, June 2001.
- Braathen, B., Bartlett, M.S., Littlewort-Ford, G., and Movellan, J.R. 3-D head pose estimation from video by nonlinear stochastic particle filtering UCSD MPLab TR 2001.05, July 2000.
- Littlewort-Ford, 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 Joint Symposium on Neural Computation.
- Bartlett, M.S., Donato, G.L., Movellan, J.R., Hager, J.C., Ekman, P., and Sejnowski, T.J. (2000). Image representations for facial expression coding. In S. Solla, T. Leen, & K. Mueller, Eds. Advances in Neural Information Processing Systems 12, Cambridge, MA: MIT Press, p. 886-892.
- E. Smith, E., Bartlett, M.S. and Movellan, J.R. (2001). Computer Recognition of Facial Actions: A Study of Co-Articulation Proceedings of the 8th Joint Symposium on Neural Computation.
- Tim K. Marks and Javier R. Movellan. Diffusion Networks, Products of Experts, and Factor Analysis UCSD MPLab TR 2001.02, June 2001.

9.1 Automatic Feature Detection

Our facial feature detectors are statistical models based on (Wiskott, Fellous, Krüger, & von der Malsburg, 1999) work. First we apply Gabor filter banks on the desired face landmarks of a set of training face images. The response of the Gabor filters at the desired landmarks is then stored in a matrix array. When detecting whether a pixel in a new image contains a desired fiducial point, we compute the Euclidean distance from the output of the filter bank at that pixel to each of the filter bank outputs in the stored matrix array. The shortest Euclidean distance to one of the training samples is the degree of response of the feature detector. (See Figure 15). When integrated with the pose estimation system, these distances will be added over the f feature points. The overall distance represents the degree of match between the image and the particle. This degree of match will be used by the particle filter in a similar manner as we used the sum of squared error when the images are hand-labeled.

Our pilot work on facial feature detection is described in Fasel, Bartlett, and Movellan (2001) and Smith (2001). We used 446 frontal view images from the FERET face database (Phillips, Wechsler, Juang, & Rauss, 1998) to train our feature detectors. In our pilot experiments we used pupils and philtrum as the features of interest. We employed a standard recognition engine (nearest neighbor) and estimated sensitivity of our detectors using the A' statistic, a non-parametric measure of sensitivity commonly used in the

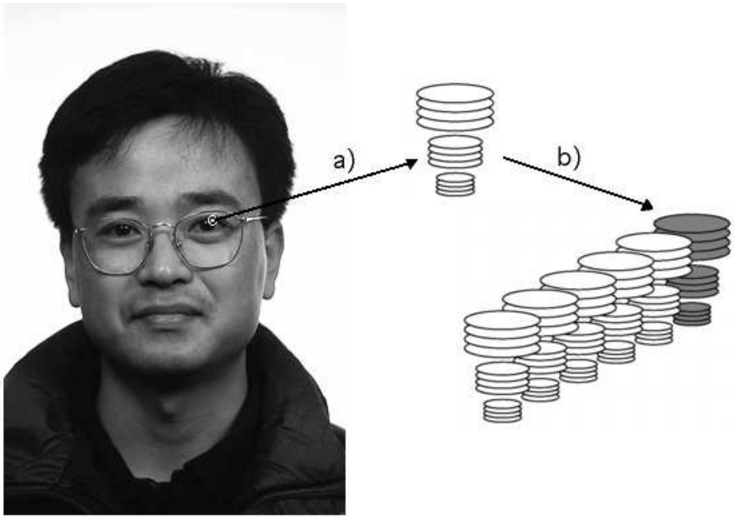


Figure 15: The process of matching features. First, a feature is convolved with a set of kernels to make a jet (a). Then that jet is compared with a collection of jets taken from training images, and the similarity value for the closest one is taken (b).

psychophysical literature. Fasel et al. (2001) tested a very wide variety of architectures for the Gabor filter banks and obtained peak performance for filter banks with eight orientations and carriers of 32 pixels/cycle (about 5.5 iris widths per cycle) for eyes and 44 pixels/cycle (about 7.5 iris widths per cycle) for philtrum. The best A's were 0.89 and 0.87 respectively. This performance is significantly better than that obtained using the original approach described in Wiskott et al. (1999).

A second feature detector developed in our lab employed a neural network architecture based on a Sparse Network of Winnows (SNoW) (Smith, 2001) The choice of algorithm was motivated by the exceptional success of a SnoW-based face detector (Yang, Roth, & Ahuja, 1999). SNoW feature detector obtained A's of .98 for both the pupil and philtrum. A description of this feature detector is also included in the Appendix (Smith, 2001).

9.2 Integration: Fully automated 3D pose estimation using particle filters.

The particle filtering approach to pose estimation requires a measure of the degree of match between a particle (i.e., a hypothesis about the 3D pose of the face) and the image. The current version of our pose estimator employs hand-labeled feature positions for this degree of match. In the next version of the system, this degree of match will be obtained from automatic feature detectors as follows: First using the pose model hypothesized by each particle, we warp the image into a frontal view. Then we apply feature detectors to the warped images in the positions specified by the particle. For example, if the particle says that there should be an eye on pixel 17 we apply an eye detector on that pixel. The sum of the degree of response of the different feature detectors is then taken as an overall measure of goodness of fit. The approach is similar to the one used by (Wiskott et al., 1999) only that we generalize it to handle out-of-image-plane rotations. A strength of this approach is that information about facial geometry and the dynamics of head movements constrains the estimates of facial feature positions.

9.3 Automatic 3D Pose Tracking via optic flow

Matthew Brand has recently developed a real time system for tracking flexible objects in 3D as they move and flex before a video camera (Brand, 2001). A member of our team, is making an extended visit to Matthew Brand's laboratory at MERL with the goal of adapting his general purpose approach to the specific problem of real time 3D face tracking. The approach works as follows: A 3D model consisting of a base shape and several morph bases is used to track the face at a number of points (see Figure 7). Morph bases describe shape changes that occur during speech. Point locations are projected onto successive pairs of video frames, and the surrounding local patches of image pixels are analyzed for flow information. This information includes an estimate of the 2D displacement for each point along with measurements of the uncertainty related to the image texture associated with each estimate. These comprise a Gaussian pdf describing the motion of each point from one frame to the next. The maximum likelihood translation of the model is estimated and removed from the flow. The remaining flow along with its uncertainty measurements is used to compute a maximum likelihood estimate of the motion to be explained by rotation and morph weights. This motion estimate is then factored using an "orthonormal decomposition" into estimates of rotation and morph weights. Model parameters are refined by repeating the operation using the translation, rotation, and morph estimates to sample from patches closer to the final destination of each point in the second frame. A particle filtering approach is also under development.

9.4 Gabor Filters and Support Vector Machines

The support vector machines employed in section 7 were further analyzed for the contribution of Gabor filters, using linear, and nonlinear kernels. The four 2-category classification tasks were trained (Blink vs. non-blink, Brow Raise vs. Non-brow motion, Brow Lower vs. Non-brow motion, and Brow Raise vs. Brow Lower). SVM's were trained and tested on the peak frames, frames which the human coder identified as the peak of the action unit. The results are given in Table 5. The main effect of the Gabor filtering step was to make the space more linearly separable. Linear SVM's applied to Gabor representations performed significantly better than linear SVM's applied to difference images on three of the four classification tasks. Performance was the same on the fourth. The success of multiscale Gabor representations as a preprocessing step may be understood in the context of kernel methods that have been discussed in relation to SVM's. See (Littlewort-Ford et al., 2001) for further discussion.

	SVM			
	Diff Images		Gabor	
	Linear	Nonlinear	Linear	Nonlinear
Blink vs. Non-blink	78.3	95.9	94.8	95.9
Brow Raise vs. Non-brow	65.6	88.5	91.7	91.7
Brow Lower vs. Non-brow	64.3	67.9	75.0	82.1
Brow Raise vs. Lower	90.3	90.3	90.3	90.3

Table 5: Summary of SVM results taking either difference images or Gabor representations as input. Linear and nonlinear kernels were tested. Accuracy is for classifying peak frames. All performances are for generalization to novel subjects.

9.5 Duchenne vs. non-Duchenne smiles

We investigated the problem of computer recognition of "Duchenne" vs. non-Duchenne smiles using support vector machines (Littlewort-Ford et al., 2001). Duchenne smiles include the contraction of the orbicularis oculi, the sphincter muscles that circle the eyes. In the Facial Action Coding System, contraction of this muscle is coded as AU 6. Genuine, happy smiles can be differentiated from posed, or social smiles by the contraction of this muscle (Ekman, Friesen, & O'Sullivan, 1988). This is a difficult visual discrimination task. Previously published performance of computer vision systems on this task is in the low 80%'s for a two-alternative forced choice (e.g. (Cohn et al., 1999)).

We investigated the performance of SVM's taking on Gabor representations as input on this task. Gabor wavelet representations have been found to be highly effective for both identity recognition (Lades, Vorbrüggen, Buhmann, Lange, Konen, von der Malsburg, & Würtz, 1993) and facial expression analysis (Donato et al., 1999; Zhang, Lyons, Schuster, & Akamatsu, 1998; Cottrell et al., 2000; Fasel et al., 2001). Aside from the similarity of Gabors to the transfer functions of visual cortical cells, the reason for the success of this representation is unclear. One hypothesis is that the bank of Gabor filters projects the images into a high dimensional space where the classes are linearly separable in a manner analogous to SVM kernels. If that is the case, then a linear classifier should perform as well as a nonlinear classifier, each taking Gabor representations as input.

This study was carried out using three image databases: The Ekman-Hager database of directed facial actions, the Cohn-Kanade database of posed expressions (Kanade, J.F., & Tian, 2000), and a database of spontaneous and posed smiles collected by the BBC. The SVMs discriminated Duchenne from non-Duchenne smiles with 87% accuracy. The Gabor kernels made the task more linearly separable, as linear kernels applied to the Gabors performed significantly better than linear kernels applied directly to the difference images. The task was not entirely linearly separable in the Gabor space, however, as linear kernels performed 5-10 percentage points lower than polynomial or Gaussian kernels applied to the Gabor outputs. SVM's on unfiltered difference images typically performed only 1-3 percentage points lower than SVM's on Gabors, but required more complex kernels and a more extensive search for the right kernel. The Gabor filters also appeared to minimize the need for taking difference images. SVM's applied directly to the graylevel images, without subtracting the neutral expression, were near chance, whereas after passing the images through Gabor filters, SVM's performed almost as well for the graylevel images as for the difference images. Performance on the individual graylevel images was typically only 1-2 percentage points lower than performance on the difference images.

9.6 Handling co-articulation effects

Co-articulation effects is a term borrowed from the speech recognition literature (Rabiner & Juang, 1993), and refers to the observation that phonemes can have very different waveforms when produced in context with other phonemes. A related problem exists in facial expression recognition. Facial actions can have a very different appearance when produced in combination with other actions. A standard solution to the co-articulation problem in speech is to extend the units of analysis to include context. For example, instead of developing phoneme models one may develop triphone models which include previous and posterior context. In our case we could develop models for combinations of 2 and 3 FACS units. While the number of possible combinations grows exponentially, in speech only a small percentage of such combinations appears in practice. The problem with this approach is that the amount of data available to teach combination models decreases dramatically with the number of combinations and thus the models become less reliable. This is known in the statistics literature as the bias/variance dilemma (Geman, Bienenstock, & Doursat, 1992). Simple models that do not take into account context tend to be more robust, while context-dependent models tend to be more precise.

An approach used in the speech recognition community to address this problem is to combine context independent models and context dependent models. We will illustrate the approach with an example. Suppose we have 10 examples of Action unit 1 in isolation (AU1), 10 examples of AU2 in isolation, and 10 examples of AU1+AU2 (a combination of AU1 and AU2). First we create two "context independent" models, one trained with the 20 examples of AU1 (alone and in AU1+AU2) and another one trained with the 20 examples of AU2. Second, we create 3 "context dependent" models, each of which use 10 examples. The context independent models will in general be more robust but less precise, while the context dependent HMMs will be more precise but less robust. Finally the two sets of models are combined by interpolation: If λ_i and λ_d are he parameters of the context independent and context dependent models for AU1, the interpolated model would have parameters $\epsilon\lambda + (1 - \epsilon)(1 - \lambda)$. The value of ϵ is set using a validation set to maximize generalization performance.

This technique has been successfully applied to several problems in speech recognition using HMM's (Rabiner & Juang, 1993). Interpolation models were found to be effective when there is insufficient data to reliably train models for all possible combinations of phonemes. A similar situation occurs in facial expression, where there is insufficient data to train all possible combinations of the 46 facial actions defined in FACS.

Smith, Bartlett, and Movellan (2001) presents a neural network analog of the interpolation models discussed in the speech recognition community. This network was applied to the problem of recognizing facial action combinations in the upper face.This study was carried out using two image databases: The Ekman-Hager database of directed facial actions and the Pittsburgh database of posed expressions. The network demonstrated robust recognition for the six upper facial actions, whether they occurred individually or in combination. Mean recognition accuracy over the six action units was 98%. This figure is the detection accuracy for each facial action regardless of whether it occurred individually or in combination. A second way to evaluate performance is to measure the joint accuracy for all six action units. Here all six outputs must be correct in order for the network output to be scored as correct. The joint accuracy was 93%. Similar results were obtained by Jeff Cohn's group using a different form of input representation (Tian et al., 2001). These results show that neural networks may provide a reasonable approach to handling coarticulation effects in automatic recognition of facial actions.

9.7 Incorporating facial dynamics with HMM's.

We applied hidden Markov models to the task of recognizing 6 upper and 6 lower facial actions in the Ekman-Hager database of directed facial actions. HMM performance was compared taking PCA, ICA, and PCA-reduced Gabor representations as input. This study is described in (Bartlett et al., 2000).

Hidden Markov models were trained on 80 sequences of 12 directed facial actions performed by 20 subjects. The input to the HMM consisted of the principal component (PCA), independent component (ICA), or Gabor representation of each frame. One model was trained for each facial action. Test sequences were classified by calculating the probability of the test sequence given each model and assigning the class label for which this conditional probability was greatest. The Gabor representation contained 1600 outputs for each image. Because estimating the parameters for hidden Markov models becomes intractable with large numbers of observations, the Gabor representation was reduced to 75 dimensions using PCA before training the HMM. The PCA and ICA representations were identical to those in our previously published work, with 75 ICA dimensions and 30 PCA dimensions.

The number of states in the HMM was varied from 1-5, and best performance was obtained with a 3-state model. Classification performance with the HMM was compared to the classifier used for our previously published results: nearest neighbor. The HMM improved classification performance with ICA from 95.5% to 96.3%, and gave similar percent improvements to the PCA and PCA-reduced Gabor representations over their nearest neighbor performances. Thus a recognition engine that is sensitive to the dynamics of facial expressions improved classification performance over a static classifier for all three representations. In addition, the dynamic recognition engine did not alter our previous findings about which image representation techniques were most effective for expression analysis. Performance improved for all three representations by employing the HMM, but the relative order of performances remained the same. The ICA and Gabor representations continued to outperform eigenfaces, and performed equally well to each other.

The dynamics in the directed facial action database were limited because the sequences were time-warped by hand prior to digitization. Each sequence contained seven frames beginning with a neutral expression and ending with a high-intensity facial action. The intervening frames were hand-selected to contain two low, two medium, and two high intensity samples of the action. There was little difference between the performance of a one-state HMM and multi-state HMM's, suggesting that the contribution of the dynamics to recognition was limited. This may be due to the manual time-warping of the image sequences in this database. The next step in this work is to apply these methods to the databases of spontaneous facial behavior provided by this project. This database has not been time-warped, and because the behavior is spontaneous, the facial movements may be faster and smoother than the directed facial actions used in the previous study.

9.8 Handling occlusions using products of experts

An issue that arises in the image representation step is that faces containing out-of-plane head rotations often have missing data. Portions of the face are occluded when the face rotates away from a frontal view. Some images in the Theft database used in this project also contain occlusions from hands in front of the face. For the results presented in Section 7, such missing valued were left untreated. However, we recently developed an approach to the missing data problem in which the missing regions are estimated from the statistics of the visible regions. The technique reconstructs missing portions of the image using products of experts and is described in (Marks & Movellan, 2001).

Hinton (Hinton, in press) recently proposed a learning algorithm for a class of probabilistic models called product of experts. Whereas in standard mixture models the "beliefs" of individual experts are averaged, in products of experts the "beliefs" are multiplied together and then renormalized. One advantage of this approach is that the combined beliefs can be much sharper than the individual beliefs of each expert, potentially avoiding the problems of standard mixture models when applied to high dimensional problems. It has been shown that a restricted version of the Boltzmann machine, in which there are no lateral connections between hidden units or between observation units, performs products of experts. The paper in the Appendix by Marks and Movellan (2001) generalizes these results to diffusion networks, a continuous-time, continuous-state version of the Boltzmann machine. Diffusion networks may be better suited to continuous data such as images or sounds than Boltzmann machines which use binary-state units.

In feedback models such as diffusion networks, one can solve inference problems such as pattern completion for occluded images using the same architecture and algorithm that are used for estimating generative models. Given an image with known values O_k, reconstructing the values of the occluded units O_u, involves finding the posterior distribution $p(O_u|O_k = o_k)$. The feedback architecture of diffusion networks provides a natural way to find such a posterior distribution: clamp the known observable units to the values O_k, and let the network settle to equilibrium. The equilibrium distribution of the network will then be the posterior distribution $p(O_u|O_k = o_k)$. When the unit activation functions are linear, this product of expert architecture is equivalent to a factor analyzer. The pattern completion is related to that obtained using PCA. Most importantly, diffusion networks can also be defined with nonlinear activation functions (Movellan, Mineiro, & Williams, in press), which leads to nonlinear generalizations of factor analysis. We have begun exploring nonlinear extensions of the linear case outlined in Marks and Movellan (2001).

10 Appendix B: Action unit tables

The following tables summarize the action unit frequency in the Theft database for 10 subjects, as coded by the Rutgers team.

	Brow			Eye							
AU / **Subject**	1	2	4	5	6	7	1+2	1+2+4	1+2+4+5	1+2+5	1+2+5+7
4						6	1				
6	1						1				
7						3	1				
9		1	1	2	1	8	4	1			
11						1	1				
14		1	1	2	1		3	2	2		
15				1				6	1	1	
16					1					2	
17		1				2	3			4	
18				2	2	1	4			1	1
No. Subjects	1	3	2	4	4	6	8	3	2	4	1
No. Examples	1	3	2	7	5	21	18	9	3	8	1

Table 6: Upper Face Actions.

AU / **Subject**	1+5	2+5	4+5	4+6	4+7	5+7	6+7	Total No. Actions
4					2	2		11
6								2
7								4
9				1				19
11								2
14	1							13
15					1			10
16			2					5
17		1				2	1	14
18								11
No. Subjects	1	1	1	1	2	2	1	
No. Examples	1	1	2	1	3	4	1	92

Table 7: Upper Face Actions (Cont.)

AU Subject	Droop 41	Slit 42	Closed 43	Squint 44	Blink 45	Wink 46
4	2	4			8	
6	1	2			10	
7	11	5	1		29	
9	5	2			40	
11		3		26		
14	2	4	2		12	
15		1			14	
16			1		10	
17	3	2	3	1	8	
18	3	5	4		15	
# Subjects	7	9	5	1	10	0
# Examples	27	28	11	1	172	0

Table 8: Eyelid Actions

AU	8	10	12	14	15	17	20	23	24	25	26	30	32	38
Subject														
4			3	1		3			2		9			
6				2					2	1	5			
7				2	1					1			1	
9	1		9	1	3	1			1					
11			2		1		1				7	1		2
14														
15		1		2							2			
16										1				
17		2		1				3	1		1			
18		7	2		1						2			
# Subjects	1	3	4	6	4	2	1	1	4	3	6	1	1	1
# Examples	1	10	16	9	6	4	1	3	6	3	26	1	1	2

Table 9: Lower Face Actions: Individual

AU / Subject	Lower Face Actions: Combos x 2							
	10+12	10+14	10+17	12+15	12+17	12+25	12+26	13+26
4					1		3	
6								
7								
9				1				
11								
14						1	1	1
15								
16						1		
17								
18	2	1	2					
No. Subjects	1	1	1	1	1	2	2	1
No. Examples	2	1	2	1	1	2	4	1

AU / Subject	Lower Face Actions: Combos x 2 (Cont.)								
	14+20	14+23	14+24	15+17	15+20	17+26	19+23	19+37	26+37
4			1					1	
6		1	2						
7	1	1	2		1				1
9									
11				1					
14			1		1				
15			1	1					
16						1			
17							1		
18									
No. Subjects	1	2	5	2	2	1	1	1	1
No. Examples	1	2	6	2	2	1	1	1	1

AU	10 12 26	10 15 17	10 17 24	14 15 24	14 17 32	14 20 24	14 23 37
Subject							
4							
6						2	
7				1	1		1
9							
11							
14							
15		1					
16							
17		1					
18	1	2	3				
# Subjects	1	3	1	1	1	1	1
# Examples	1	4	3	1	1	2	1

Table 10: Lower Face Actions: Combos $\times$ 3

Subject	Upper Face	Lower Face	Overall
4	11	24	35
6	2	14	16
7	4	14	18
9	19	17	36
11	2	15	17
14	13	5	18
15	10	8	18
16	5	3	8
17	14	10	24
18	11	23	34
Total Examples	92	133	225

Table 11: Total Number of Actions

11 References

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. (2001). Image representations for facial expression recognition. *IEEE transactions on neural networks*. Submitted.

Bartlett, M., Viola, P., Sejnowski, T., Larsen, J., Hager, J., & Ekman, P. (1996). Classifying facial action. In D. Touretski, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems*, Vol. 8 (pp. 823–829). San Mateo, CA: Morgan Kaufmann.

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.

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.

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

Cohn, J., Zlochower, A., Lien, J., & Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. *Psychophysiology*, *36*, 35–43.

Cole, R. (2000). Creating the next generation of automated intelligent systems. *Neural Information Processing Systems Postconference Workshop on Affective Computing*.

Cottrell, G., Dailey, M., Padgett, C., & R, A. (2000). *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges*, Chap. Is all face processing holistic? The view from UCSD. Erlbaum.

Cottrell, G., & Metcalfe, J. (1991). Face, gender and emotion recognition using holons. In D. Touretzky (Ed.), *Advances in Neural Information Processing Systems*, Vol. 3 (pp. 564–571). San Mateo, CA: Morgan Kaufmann.

Daugman, J. (1988). Complete discrete 2d gabor transform by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *36*, 1169–1179.

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.

Ekman, P. (2001). *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. New York: W.W. Norton, 3rd edition.

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., Friesen, W., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, *545*, 414 – 420.

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., Bartlett, M., & Movellan, J. (2001). *A comparison of gabor filter methods for automatic detection of facial landmarks* (Technical Report MPLab TR 2001.04). UCSD Dept. of Cognitive Science.

Field, D. (1994). What is the goal of sensory coding? *Neural Computation*, *6*, 559–601.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

Gray, M. S., Movellan, J. R., & Sejnowski, T. J. (1997). Dynamic features for visual speechreading: A systematic comparison. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems*, Vol. 9. Cambridge, MA: MIT Press.

Hinton, G. (in press). Training products of experts by minimizing contrastive divergence. *Neural Computation*.

Isard, M., & Blake, A. (1998). Condensation: conditional density propagation for visual tracking. *Int. J. Computer Vision*, *29*(1), 5–28.

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

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.

Lades, M., Vorbrüggen, J., Buhmann, J., Lange, J., Konen, W., von der Malsburg, C., & Würtz, R. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, *42*(3), 300–311.

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

Lee, D., & Seung, S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

Lee, Y., Lin, Y., & Wahba, G. (2001). *Multicategory support vector machines* (Technical Report TR 1040). U. Wisconsin, Madison, Dept. of Statistics.

Li, H., Roivainen, P., & Forchheimer, R. (1993). 3-d motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(6), 545–555.

Lien, J., Kanade, T., J.F., C., & Li, C. (2000). Detection, tracking, and classification of action units in subtle changes of facial expression. *Journal of Robotics and Autonomous Systems*, *31*(3), 131–46.

Littlewort-Ford, G., Bartlett, M., & Movellan, J. (2001). Are your eyes smiling? detecting genuine smiles with support vector machines and gabor wavelets. *Proceedings of the 8th Joint Symposium on Neural Computation*.

Lu, C.-P., Hager, D., & Mjolsness, E. *Object pose from video images*. To appear in IEEE PAMI.

Marks, T. K., & Movellan, J. R. (2001). *Diffusion networks, products of experts, and factor analysis* (Technical Report MPLab TR 2001.02). UCSD Dept. of Cognitive Science.

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

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.

Movellan, J. R., Mineiro, P., & Williams, R. J. (in press). A Monte-Carlo EM approach for partially observable diffusion processes: Theory and applications to neural networks. *Neural Computation*.

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*, Vol. 9. Cambridge, MA: MIT Press.

Penev, P., & Atick, J. (1996). Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, *7*(3), 477–500.

Phillips, P., Wechsler, H., Juang, J., & Rauss, P. (1998). The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing Journal*, *16*(5), 295–306.

Pighin, F., D, H., Szeliski, R., & Salesin, D. (1998). Synthesizing realistic facial expressions from photographs. *Proc SIGGRAPH*.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.

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.

Simoncelli, E. P. (1997, November 2-5). Statistical models for images: Compression, restoration and synthesis. *31st Asilomar Conference on Signals, Systems and Computers*. Pacific Grove, CA.

Singh, A. (1991). *Optic flow computation*. Los Alamitos, CA: IEEE Computer Society Press.

Smith, E. (2001). *A snow-based automatic facial feture detector* (Technical Report MPLab TR2001.06). UCSD Department of Cognitive Science.

Smith, E., Bartlett, M., & Movellan, J. (2001). Computer recognition of facial actions: A study of co-articulation effects. *Proceedings of the 8th Joint Symposium on Neural Computation*.

Terzopoulus, 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), 569–579.

Tian, Y., Kanade, T., & Cohn, J. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 97–116.

Tukey, J. (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.

Wechsler, H., Phillips, P., Bruce, V., Fogelman-Soulie, F., & Huang, T. (Eds.). (1998). *Face recognition: From theory to applications*. NATO ASI Series F. Springer-Verlag.

Wiskott, L., Fellous, J.-M., Krüger, N., & von der Malsburg, C. (1999). Face recognition by elastic bunch graph matching. In L. C. Jain, U. Halici, I. Hayashi, & S. B. Lee (Eds.), *Intelligent biometric techniques in fingerprint and face recognition* (Chap. 11, pp. 355–396). CRC Press.

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.

Yang, M., Roth, D., & Ahuja, N. (1999). A snow-based face detector. *Advances in nueral information processing systems*, Vol. 12.

Zhang, Z., Lyons, M., Schuster, M., & Akamatsu, S. (1998). Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. *Proceedings, Third IEEE International Conference on Automatic Faceand Gesture Recognition* (pp. 454–459). IEEE Computer Society.