An Approach to Automatic Recognition of Spontaneous Facial Actions

B. Braathen, M.S. Bartlett, G. Littlewort, E. Smith, and J.R. Movellan
Institute for Neural Computation
University of California, San Diego
Email: bjorn, marni, gwen, evan, javier @inc.ucsd.edu

Abstract

We present ongoing work on a project for automatic recognition of spontaneous facial actions. Spontaneous facial expressions differ substantially from posed expressions, similar to how spontaneous speech differs from directed speech. Previous methods for automatic facial expression recognition assumed images were collected in controlled environments in which the subjects deliberately faced the camera. Since people often nod or turn their heads, automatic recognition of spontaneous facial behavior requires methods for handling out-of-image-plane head rotations. There are many promising approaches to address the problem of out-of-image plane rotations. In this paper we explore an approach based on 3-D warping of images into canonical views. A front-end system was developed that jointly estimates camera parameters, head geometry and 3-D head pose across entire sequences of video images. First a small set of images was used to estimate camera parameters and 3D face geometry. Markov chain Monte-Carlo methods were then used to recover the most likely sequence of 3D poses given a sequence of video images. Once the 3D pose was known, we warped each image into frontal views with a canonical face geometry. We evaluated the performance of the approach as a front-end for a spontaneous expression recognition system using support vector machines and hidden Markov models. This system employed general purpose learning mechanisms that can be applied to recognition of any facial movement. 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 expression 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.

1. Introduction

The Facial Action Coding System (FACS) developed by Ekman and Friesen [6] 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. FACS has already proven a useful behavioral measure in studies of emotion, communication, cognition, psychopathology, and child development (see [7] for a review). FACS coding is presently performed by trained human observers using visual inspection. The human coders decompose the expression in each video frame into component actions (see Figure 1). A major impediment to the widespread use of FACS is the time required to train human experts and to manually score the video tape. Approximately 300 hours of training are required to achieve minimal competency on FACS, and each minute of video tape takes approximately two hours to score thoroughly.

Much of the early work on computer vision applied to facial expressions focused on recognizing a few prototypical expressions of emotion produced on command (e.g. "smile"). More recently there has been an emergence of groups that analyze facial expressing into elementary movements. For example, Essa and Pentland [8] and Yacoob and Davis [19] proposed methods to analyze expressions into elementary movements using an animation style coding system inspired by FACS. Eric Petajan's group has also worked for many years on methods for automatic coding of facial expressions in the style of MPEG4 [3]. While coding standards like MPEG4 are useful for animating facial avatars, they are of limited use for behavioral research. For example, MPEG4 codes movement of a set of facial feature points, but does not encode some behaviorally relevant facial movements, such as the muscle that circles the eye (orbicularis oculi). It also does not encode the wrinkles and bulges that are critical for distinguishing some facial muscle movements that are difficult to differentiate using motion alone yet can have different behavioral implications (e.g. see Figure 1b, AU 1 vs. 1+4). One other group has focused on automatic FACS recognition as a tool for behavioral research, lead by Jeff Cohn and Takeo Kanade. The work presented here was conducted in collaboration with that group. We explored and compared approaches for automatic FACS coding of spontaneous facial expressions from freely behaving individuals. More details are available in [1, 2, 10]. Here we describe the system developed at UCSD.

The most critical difference between the present work and previous work is the use of spontaneous facial expressions. Most of the previous work employed datasets of posed expressions collected under controlled imaging conditions with subjects deliberately facing the camera. Extending these systems to spontaneous facial behavior is a critical step forward for applications of this technology. Psychophysical work has shown that spontaneous facial expressions differ from posed expressions in a number of ways

[5]. Subjects often contract different facial muscles when asked to pose an emotion such as fear versus when they are actually experiencing fear. In addition, the dynamics are different. Spontaneous expressions have a fast and smooth onset, with apex coordination, in which facial actions in different parts of the face peak at the same time. In posed expressions, the onset tends to be slow and jerky, and the actions typically do not peak simultaneously.

Spontaneous face data brings with it a number of technical issues that need to be addressed for computer recognition of facial actions. One of the most important technical challenges is the presence of out-of-plane rotations due to the fact that people often nod or turn their head as they communicate with others. 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.

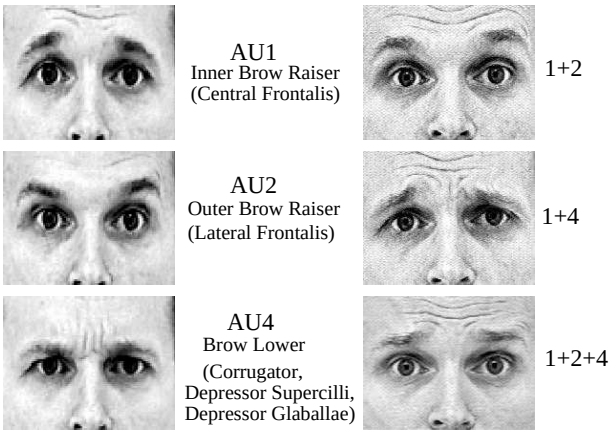

Figure 1. The Facial Action Coding System decomposes facial motion into component actions. The three individual brow region actions and selected combinations are illustrated.

There are a number of possible approaches to handling head rotations. In this paper we explore an approach based on 3D pose estimation and warping of face images into canonical poses (e.g., frontal views). Since our goal is to explore the potential of this approach as a front-end to facial expression recognition, we first tested it using 8 hand-labeled facial landmarks. However the approach can be generalized in a straightforward and principled manner to work with automatic feature detectors.

2. Estimation of Face Geometry

We start with a canonical wire-mesh face model [18] 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 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 images we recovered, the 3D positions of the 8 tracked features in object coordinates. A scattered data interpolation technique [18] was 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 $\mathbf{u_i} = \mathbf{f}(\mathbf{p_i})$ from which we can compute $\mathbf{u_j} = \mathbf{f}(\mathbf{p_j})$. Interpolation then consists of applying

$$f(\mathbf{p}) = \sum_{\mathbf{i}} \mathbf{c_i} \phi(||\mathbf{p} - \mathbf{p_i}||) \quad (1)$$

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 $\mathbf{u_i} = \mathbf{f}(\mathbf{p_i})$ and the constraints $\sum_i \mathbf{c_i} = \mathbf{0}$ and $\sum_i \mathbf{c_i} \mathbf{p_i^T} = \mathbf{0}$.

3. 3D pose estimation

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". Here we explored a solution to this problem using Markov Chain Monte-Carlo methods, also known as condensation algorithms or particle filtering methods, [12, 11, 4].

3.1. Particle filters

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.

Since our main goal was to explore the use of 3D models to handle out-of-plane rotations in expression recognition problems, our first version of the system, which is the one presented here, relies on knowledge of the position of facial landmarks in the image plane. We are currently working on extensions of the approach to rely on the output of automatic feature detectors, instead of hand-labeled features. In the current version of the system we used the 8 landmarks mentioned Section 2.

Our approach works 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 f 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 f features on the image plane and the positions hypothesized by the particle. In future versions this likelihood function will be based on the output of automatic feature detectors. 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 [12] 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

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

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 OI algorithm [15], which is known to be very robust to the effects of noise.

3.2. The Orthogonal Iteration Algorithm

In the OI algorithm [15] the pose estimation problem is formulated as that of minimizing an error metric based on collinearity in object space. The method is iterative and directly computes orthogonal rotation matrices which are globally convergent. The error metric is

$$\mathbf{e_i} = (\mathbf{I} - \mathbf{F_i})(\mathbf{R}\mathbf{p_i} + \mathbf{t}) \tag{3}$$

where F_i is given by

$$F_i = \frac{\mathbf{v_i}\mathbf{v_i^T}}{\mathbf{v_i^T}\mathbf{v_i}} \tag{4}$$

and $\mathbf{v_i}$ is the projection of the 3D points onto the normalized image plane. In Eq. 3 $\mathbf{p_i}$, $\mathbf{R}$ and $\mathbf{t}$ denote 3D feature positions, the rotation matrix and translation vector, respectively. A minimization of

$$E(\mathbf{R}, \mathbf{t}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} ||\mathbf{e_i}||^2 \tag{5}$$

is then performed. The algorithm is known to be very robust to the effects of noise [15].

3.3. Results

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 2a 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).

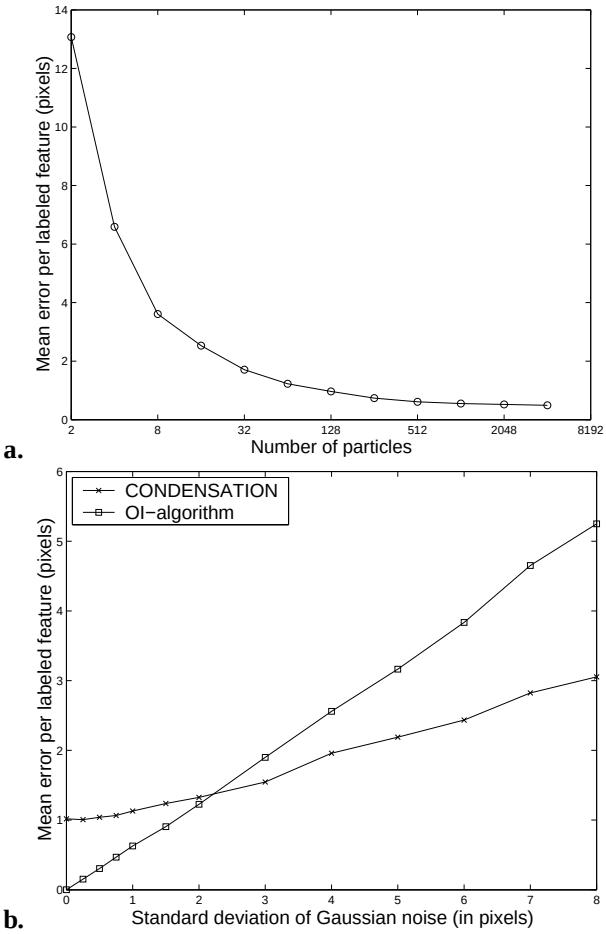

Figure 2. a. Performance of the particle filter is shown as a function of the number of particles used. b. Performance of the particle filter and the OI algorithm as a function of noise added to the true positions of features.

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 2b 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.

4. Automatic FACS recognition

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

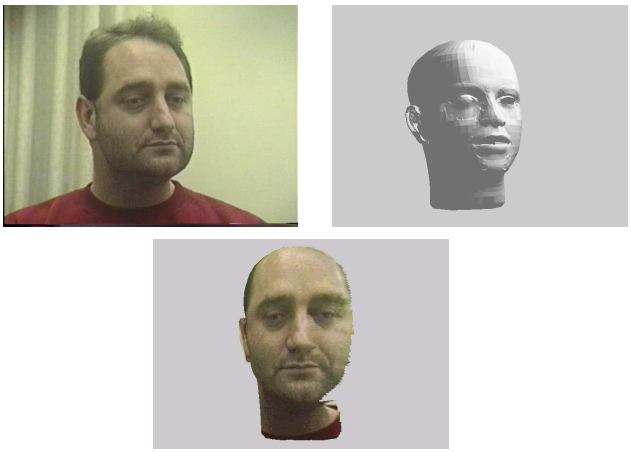

Figure 3. Original image, model in estimated pose and warped image.

As a preliminary test of the ability to classify facial movements in the rotated face data, three facial behaviors were classified in the video sequences: Blink (AU 45 in the FACS system), brow raise (joint expression of AU 1+2), and brow lower (AU 4)[1]. (See Figure 1.) These facial actions were chosen for their frequency in the database, such that sufficient training samples would be available, and also for their relevance to applications such as monitoring of alertness, anxiety, and confusion. Twelve subjects provided spontaneous examples of brow raises, and nine subjects provided spontaneous examples of movements from the brow lower category. Data from ten subjects was used to train and test blinks. A fourth category consisted of randomly selected image sequences from the experimental session, matched by subject and sequence length.

Head pose was estimated in the video sequences using a particle filter with 100 particles. Face images were then warped onto a face model with canonical face geometry, rotated to frontal, and then projected back into the image plane, as illustrated in Figure 3. This alignment was used to define and crop a subregion of the face image containing the eyes and brows. The vertical position of the eyes was 0.67 of the window height. There were 105 pixels between the eyes and 120 pixels from eyes to mouth. Pixel brightnesses were linearly rescaled to [0,255]. Soft histogram equalization was then performed on the image gray-levels by applying a logistic filter with parameters chosen to match the mean and variance of the gray-levels in the neutral frame [16]. The resulting images were then convolved with a bank of Gabor kernels at 5 spatial frequencies and 8 orientations. Output magnitudes were normalized to unit length and then downsampled by a factor of 4.

[1] To increase the number of training samples, also included in this category was AU 9 (nose wrinkle) which also lowers the brows, and AU 1+4

Blinks: SVM's were first trained to discriminate images of the peak of blink sequences (as labeled by FACS coders) from randomly selected images containing no blinks. Generalization to novel subjects was tested using leave-one-out cross-validation. A nonlinear SVM applied to the Gabor representations obtained 95.9% correct 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 [14], 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.

Blink trajectories: Figure 4a shows the time course of SVM outputs for Blinks. The SVM output was the margin (distance along the normal to the class partition). Although the SVM was not trained to measure the amount of eye opening, it is an emergent property. In all time courses shown, the SVM outputs are test outputs (the SVM was not trained on the subject shown). Figure **??** shows the SVM trajectory when tested on a sequence with multiple peaks.

Classifying full sequences. A better test of action unit recognition is for the case in which the location of the peak frame is unknown. Hidden Markov Models (HMM's) were trained to classify action units from the trajectories of SVM outputs. The HMM's were trained on the outputs of the SVM. 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 a set of "test" output sequences that were then used to train the HMM's. Two hidden Markov models, one for Blinks and one for random matched sequences, 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 98.2% correct was obtained using 6 states and 7 Gaussians.

Brows: The goal was to discriminate three action units localized around the eyebrows. Since this is a 3-category task and SVMs are originally designed for binary classification

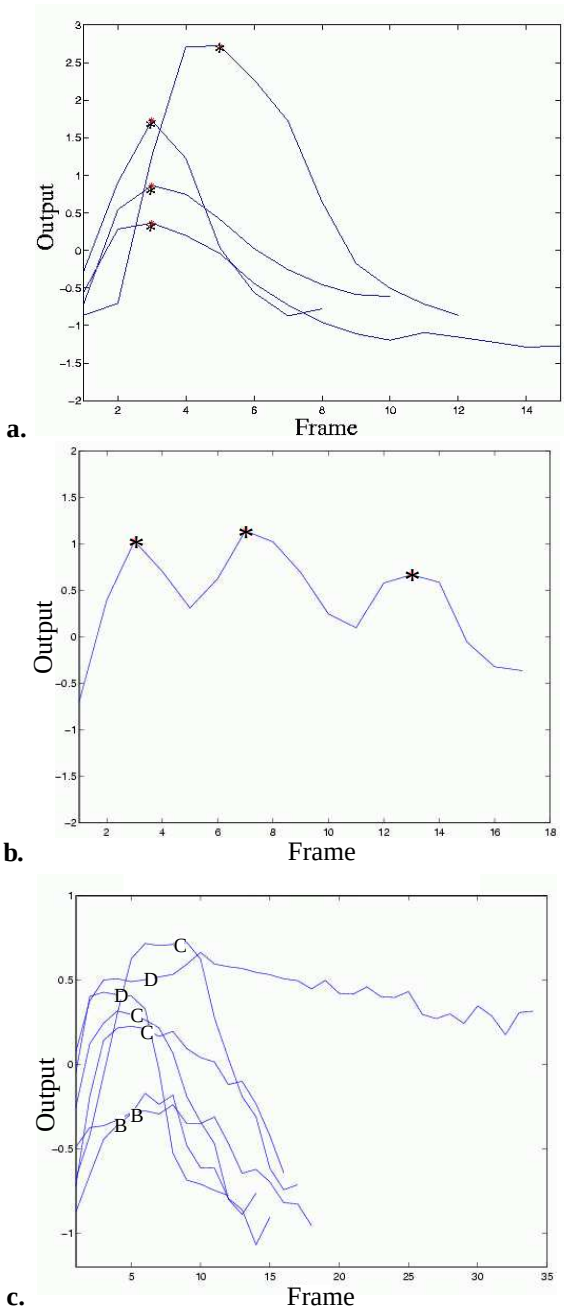

Figure 4. a. Blink trajectories of SVM outputs for four different subjects. Star indicates the location of the AU peak as coded by the human FACS expert. b. SVM output trajectory for a blink with multiple peaks (flutter). c. Brow raise trajectories of SVM outputs for one subject. Letters A-D indicate the intensity of the AU as coded by the human FACS expert, and are placed at the peak frame.

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

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

Contribution of 3D alignment to recognition: In order to examine 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. The aspect ratio was adjusted so that there were 105 pixels between the eyes and 120 pixels from eyes to mouth. A nonlinear SVM obtained 95.5% accuracy for detecting blinks from non-blinks in individual 2D aligned images. This was identical to performance using the 3-D rotations. In contrast, the 3D rotations appear to have aided the detection of brow raises. Performance of a nonlinear SVM dropped from 88.5% to 83.3% when using the 2D aligned images.

Action	**% Correct (HMM)**	**N**
Blink vs. Non-blink	98.2	336
Brow Raise vs. Random	90.6	96
Brow Lower vs. Random	75.0	28
Brow Raise vs. Brow Lower	93.5	62
Brow Raise vs. Lower vs. Random	78.2	124

Table 1. Summary of results. All performances are for generalization to novel subjects. N: Total number of positive and negative examples.

[2] SVM theory can be extended in various ways to perform multiclass decisions (e.g. [13]). In future work, a multiclass SVM will be included as input to dynamic models as well.

5. Conclusions

We explored an approach for handling out-of-plane head rotations in automatic recognition of spontaneous facial expressions from freely behaving individuals. The approach fits a 3D model of the face and rotates it back to a canonical pose (e.g., frontal view). We found that machine learning techniques applied directly to the warped images is a promising technology for automatic coding of spontaneous facial expressions. This approach employed general purpose learning mechanisms that can be applied to the recognition of any facial action. The approach is parsimonious and does not require defining a different set of feature parameters or image operations for each facial action. One exciting finding is the observation that important measurements emerged out of filters derived from the statistics of the images. For example, the output of the SVM filter matched to the blink detector could be potentially used to measure the dynamics of eyelid closure, even though the system was not designed to explicitly detect the contours of the eyelid and measure the closure. (See Figure 4.)

We found a particle filtering approach to 3D pose estimation was more robust to noise than the IO algorithm, one of the most robust deterministic pose estimation algorithms [15]. Most importantly, generalization of the particle filtering approach to use automatic feature detectors instead of hand-labeled features is relatively straightforward. We are presently developing automatic feature detectors [9] to be integrated with this system. We are also combining the particle filtering approach with a system for developed by Matthew Brand for automatic real-time head pose estimation based on optic flow [**?**]. The particle filters presented here use very simple (zero drift) face dynamics. Another advancement underway is to train diffusion networks [17] to develop more realistic face dynamics models. The optic flow measurements will be input to the face dynamics model to refine the distribution of head poses at the next time step.

While the database we used was rather large for current digital video storage standards, in practice the number of examples of each action unit in the database was relatively small. This was the primary reason why we could only prototype the system on the three actions which had the most examples: Blinks (168 examples), Brow raise (48 examples), and Brow lower (14 examples). Inspection of the performance of our system shows that 14 examples was sufficient to successfully learn an action, an order of 50 examples was sufficient to achieve performance over 90%, and an order of 150 examples was sufficient to achieve over 98% accuracy and learn smooth trajectories.

All of 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. Collection of a much larger, realistic database to be shared by the research community is a critical next step.

References

[1] M. Bartlett, B. Braathen, G. Littlewort-Ford, J. Hershey, I. Fasel, T. Marks, E. Smith, T. Sejnowski, and J. Movellan. Automatic analysis of of spontaneous facial behavior: A final project report. Technical Report UCSD MPLab TR 2001.08, University of California, San Diego, 2001.

[2] J. Cohn, T. Kanade, T. Moriyama, Z. Ambadar, J. Xiao, J. Gao, and H. Imamura. A comparative study of alternative FACS coding algorithms. Technical Report CMU-RI-TR-02-06, Robotics Institute, Carnegie-Mellon Univerisity, 2001.

[3] P. Doenges, F. Lavagetto, J. Ostermann, I. Pandzic, and E. Petajan. Mpeg-4: Audio/video and synthetic graphics/audio for real-time, interactive media delivery. *Image Communications Journal*, 5(4), 1997.

[4] A. Doucet, N. de Freitas, and N. Gordon. An introduction to sequential monte carlo methods. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[5] P. Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. W.W. Norton, New York, 2nd edition, 1991.

[6] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

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

[8] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–63, 1997.

[9] I. Fasel, M. Bartlett, and J. Movellan. A comparison of gabor filter methods for automatic detection of facial landmarks. In *Proceedings of the 5th International Conference on Face and Gesture Recognition*, 2002. Accepted.

[10] M. Frank, P. Perona, and Y. Yacoob. Automatic extraction of facial action codes. final report and panel recommendations for automatic facial action coding. Technical report, Investment Program Office, Central Intelligence Agency, 2001.

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

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

[13] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. Technical Report TR 1040, U. Wisconsin, Madison, Dept. of Statistics, 2001.

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

[15] C.-P. Lu, D. Hager, and E. Mjolsness. Object pose from video images. Accepted to appear in IEEE PAMI.

[16] J. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 851–858. MIT Press, Cambridge, MA, 1995.

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

[18] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin. Synthesizing realistic facial expressions from photographs. *Computer Graphics*, 32(Annual Conference Series):75–84, 1998.

[19] Y. Yacoob and L. Davis. Recognizing human facial expressions from long image sequences using optical fbw. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.