First Steps Towards Automatic Recognition of Spontaneuous Facial Action Units

UCSD MPLab TR 2001.07

B. Braathen, M. S. Bartlett, G. Littlewort, J. R. Movellan
Institute for Neural Computation
University of California San Diego

ABSTRACT

We present ongoing work on a project for automatic recognition of spontaneous facial actions (FACs). Current methods for automatic facial expression recognition assume images are collected in controlled environments in which the subjects deliberately face 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. Since our goal is to explore the potential of this approach, we first tried with images with 8 hand-labeled facial landmarks. However the approach can be generalized in a straight-forward manner to work automatically based on the output of automatic feature detectors. A front-end system was developed that jointly estimates camera parameters, head geometry and 3-D head pose across entire sequences of video images. Head geometry and image parameters were assumed constant across images and 3-D head pose is allowed to vary. First a 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 evaluate the performance of the approach as a front-end for an spontaneous expression recognition task.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Input devices and strategies; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*psychology*

PUI 2001 Orlando, FL USA

General Terms

Algorithms, Human Factors, Measurement

Keywords

FACS, Particle Filters, Expression Recognition

1. INTRODUCTION

The Facial Action Coding System (FACS) developed by Ekman and Friesen [7] 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 [5], communication [8], cognition [24], and child development [1]. 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). 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 one hour to score.

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 [16, 23, 21, 9], measurements of the shapes of facial features and their spatial arrangements [13], holistic spatial pattern analysis using techniques based on principal components analysis (PCA) [2, 19, 13] and methods for relating face images to physical models of the facial skin and musculature [16, 22, 14, 9].

Most of the previous work employed datasets of posed expressions collected under controlled image conditions. Subjects deliberately faced the camera and the facial expressions were temporally segmented. Extending these systems to spontaneous facial behavior is a critical step forward for applications of this technology. Psychophysical work has showed that spontaneous facial expressions differ from posed expressions in a number of ways [6]. 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.

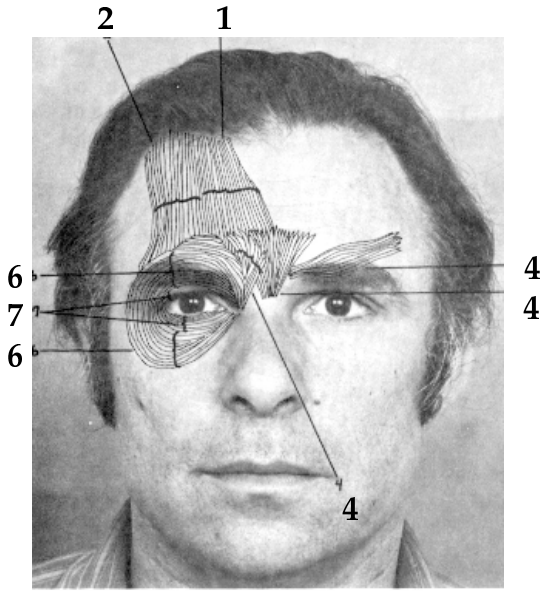

Figure 1: The Facial Action Coding System decomposes facial motion into component actions. The upper facial muscles corresponding to action units 1, 2, 4, 6 and 7 are illustrated. Adapted from Ekman & Friesen (1978).

There are many potentially reasonable approaches to handle 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).

2. ESTIMATION OF FACE GEOMETRY

We start with a canonical wire-mesh face model [20] 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 [20] was then used to modify the canonical face model to fit the 8 known 3D

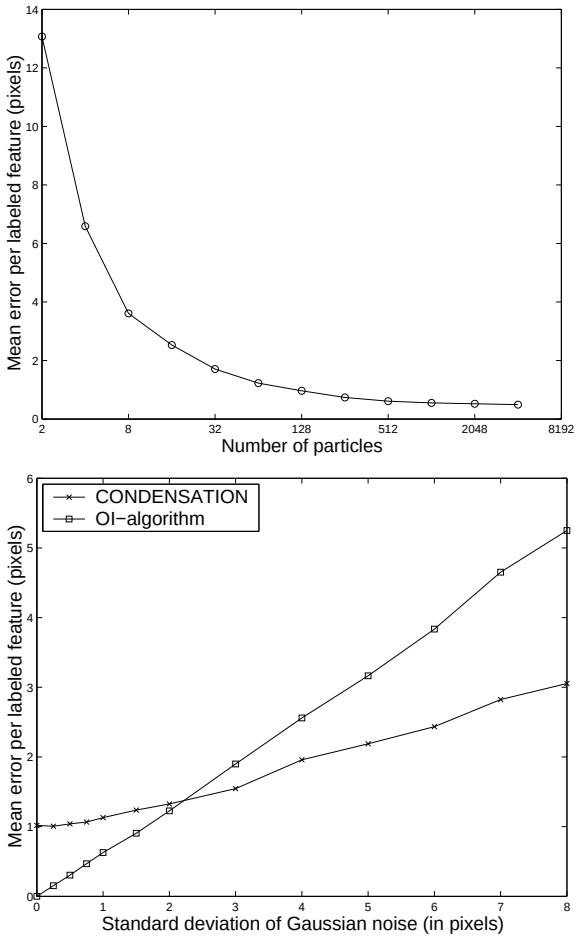

Figure 2: On the left, the performance of the particle filter is shown as a function of the number of particles used. On the right the performance of the particle filter and the OI algorithm as a function of noise added to the true positions of features.

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

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

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 $\mathbf{u}_i = f(\mathbf{p}_i)$ and the constraints $\sum_i \mathbf{c}_i = 0$ and $\sum_i \mathbf{c}_i \mathbf{p}_i^T = 0$.

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 explore 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 $n \rightarrow \infty$ 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

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

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

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

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 2 (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 2 (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.

4. AUTOMATIC FACS RECOGNITION

4.1 Database

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 2 Asian, and 1 African American, and 7 Caucasian subjects. 3 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 CMU.

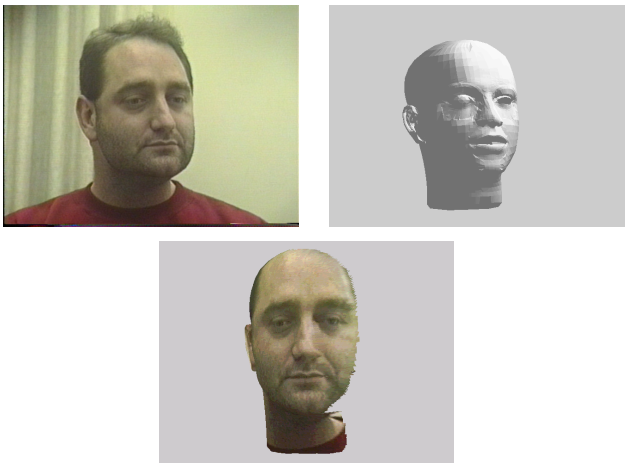

Figure 3: Original image, model in estimated pose and warped image.

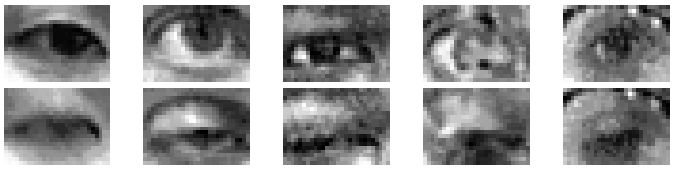

Figure 4: Examples of blink (lower row) and non-blinks (upper row) images after warping. The first three subjects (left 3 columns) had no glasses. The last 2 columns show blinks and non-blinks for 2 subjects with glasses. The prior rotation of the images allowed the same pixel numbers to be used to locate the eyes in every example.

As a preliminary test of the ability to classify facial movements in the rotated face data, two facial behaviors were classified in the video sequences: Blink (AU 45 in the FACS system) and brow raise (joint expression of AU 1 and AU 2). These facial actions were chosen for their well known relevance to applications such as monitoring of alertness and anxiety.

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 two subregions in the face images, one centered on the eyes (20x40), and the other centered on the brows (20x80). Soft histogram equalization was performed on the image gray-levels by applying a logistic filter with parameters chosen to match the mean and variance of the gray-levels of each image sequence [18]. Difference images were obtained by subtracting a neutral expression image from images containing the facial behaviors.

Separate support vector machines (SVM's) were trained for blink versus non-blink, and brow raise versus no brow raise. The peak frames of each action, as coded by the human FACS coders, were used to train and test the support vector machines. A sample of images from the blink versus no-blink task is presented in Figure 4. The task is quite challenging due to variance in race, the presence of glasses, and noise in the human FACS coding. Note in Figure 4 that the eyes are not always fully closed in the peak frames. Generalization to novel subjects was tested using leave-one-out cross-validation. Linear SVM's taking difference images performed in the low 80%'s. Non-linear SVM's improved performance by up to 10%. Specifically, the Gaussian radial basis function SVM based on the Euclidean distances between difference images performed as follows: 90.5% for blinks for all subjects, 94.2% for blinks without glasses and 84.5% on brow raises.

Performance depended on the the goodness of fit to the head model of the subject's facial geometry. We are presently making the model more robust to variations in face shape by adding more feature points and experimenting with different feature points. Reduced performance on subjects with glasses is being addressed by including information on the position of the frames in the face model. Support vector machines are presently being trained taking Gabor wavelet representations as input. Our previous work demonstrated that Gabor wavelet representations are highly effective as input for facial expression classification [3].

5. CONCLUSIONS

We explored an approach for handling out-of-plane head rotations in automatic recognition of spontaneous facial expressions. The approach fits a 3D model of the face and rotates it back to a canonical pose (e.g., frontal view). The output of the images warped into frontal views, were then used to recognize blinks (AU 45) and brow raises (AU 1+2). The results were very promising and serve as a starting point with respect to which future systems may be evaluated.

We found a particle filtering approaches to 3D pose estimation were also very promising, significantly outperforming some of the most robust deterministic pose estimation algorithms, like the OI algorithm [15]. Most importantly, generalization of the particle filtering approach to use automatic feature detectors instead of hand-labeled features is relatively straight forward.

We presented work in progress and significant improvements of the system are occurring as we write this report. The particle filters presented use very simple (zero drift) face dynamics. We are in the process of training diffusion networks [17] to develop more realistic face dynamics models. Such models may significantly reduce the number of particles needed to achieve a desired accuracy level. We are also developing automatic feature detectors [10] to be integrated with the particle filtering approach for fully automatic 3D tracking. We are also developing methods to estimate face geometry more accurately and to take into account special conditions, like the presence of glasses.

6. REFERENCES

[1] L.A. Camras. Facial expressions used by children in conflict situations. *Child Development*, 48:1431–35, 1977.

[2] G. Cottrell and J. Metcalfe. Face, gender and emotion recognition using holons. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 3, pages 564–571, San Mateo, CA, 1991. Morgan Kaufmann.

[3] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE PAMI*, 21(10):974–989, 1999.

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

[5] P. Ekman. Expression and the nature of emotion. In K.R. Scherer and P. Ekman, editors, *Approaches to Emotion*, pages 319–343. Lawrence Erlbaum, Hillsdale, NJ, 1984.

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

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

[8] P. Ekman and H. Oster. Facial expressions of emotion. *Annual Review of Psychology*, 30:527–554, 1979.

[9] 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.

[10] Ian R. Fasel, Evan C. Smith, Marian S. Bartlett, and Javier Movellan. A comparison of methods for automatic detection of facial landmarks. In *Proc. VII Joint Symp. Neural Computation*, San Diego, CA, 2000.

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

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

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

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

[17] J. R. Movellan, P. Mineiro, and R. J. Williams. Partially observable SDE models for image sequence recognition tasks. In T. Dietterich, editor, *Advances in Neural Information Processing Systems*, number 13. MIT Press, Cambridge, Massachusetts, In Press.

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

[19] C. Padgett and G. Cottrell. Representing face images for emotion classification. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, Cambridge, MA, 1997. MIT Press.

[20] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin. Synthesizing realistic facial expressions from photographs. *Computer Graphics*, 32(Annual Conference Series):75–84, 1998.

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

[22] D. Terzopoulus and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.

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

[24] R.B. Zajonc. The interaction of affect and cognition. In K. Scherer and P. Ekman, editors, *Approaches to Emotion*, pages 239–246. Lawrence Erlbaum, Hillsdale, NJ, 1984.