3-D head pose estimation from video by nonlinear stochastic particle filtering

Bjørn Braathen[*]
Institute for Neural Computation
UCSD
bjorn@inc.ucsd.edu

Marian Stewart Bartlett
Institute for Neural Computation
UCSD
marni@inc.ucsd.edu

Gwen Littlewort-Ford
Institute for Neural Computation
UCSD
gwen@inc.ucsd.edu

Javier Movellan
Institute for Neural Computation
UCSD
movellan@cogsci.ucsd.edu

Abstract

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. We approached this problem by developing a front-end system 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. The system was developed using a non-linear stochastic filtering approach: 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 compare the particle filter approach to deterministic approaches like the orthogonal iteration algorithm [7]. We evaluate the performance of our system as a front-end for an spontaneous expression recognition task.

1 Introduction

Most facial expression recognition work to date has been performed using images collected in controlled environments in which the subjects deliberately face the camera. Since people often nod or turn their heads extensions of this work to spontaneous facial behavior requires a method for handling out-of-plane head rotations. Many approaches to identity recognition including eigenfaces, ICA, and Gabor wavelet analysis also require rotation to alignment of either the faces in the database or the acquired data. We present pilot work on a system

[*] Alternative e-mail address: Bjorn.Braathen@ffi.no

for estimating pose in an image sequence. The pose information is used to warp the image onto a 3-D head model and then rotate the face image to a frontal pose.

We approach 3-D pose estimation as a probabilistic inference problem. 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 (i.e., position and rotation of the face). We represent pose parameters, i.e. rotation and translation, by the sequence $S = (S_1, \cdots, S_t)$. Formally, the estimation of S from O is a probabilistic inference problem known as "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, [6, 5, 2].

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.

The approach proposed here allows to easily incorporate information about spatial constraints between features, and dynamic constraints about the way faces move in 3D space. The current version of the approach relies on knowledge of the position of some facial landmarks in the image plane. However the extension of the approach to non-labeled images is straight-forward.

We tested this tracking method on video sequences of subjects producing spontaneous head motion during discourse, and compared performance with the orthogonal iteration algorithm, known to be one of the most robust algorithms when feature positions are known [7]. We found that the particle filtering approach is relatively fast and more robust to the presence of noise in the feature positions than the deterministic approach. We then evaluated the performance of this system as a front-end for automatic analysis of spontaneous facial behavior using support vector machines. The system successfully classified facial behaviors across significant changes in pose.

2 Particle filters

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. 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 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

$$\lim_{n \to \infty} \frac{n_t(x)}{n} = P(S_t = x | O_1, \cdots, O_t) \tag{1}$$

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.

For this investigation we used $f = 14$ facial feature points: lateral and nasal corners of the eyes, and the centers of the irises, the eyebrows, nostrils, the nose tip, and the base of the upper teeth. In our current prototype ground truth facial feature positions were hand-labeled but the system can be generalized to use feature positions provided by an automatic feature tracker (e.g. [4]) in a straightforward manner. In fact we show that one of the advantages of the particle filtering approach is that its robustness to uncertainty in the feature positions.

3 The Orthogonal Iteration Algorithm

In the OI algorithm [7] 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

$$\mathbf{e}_i = (\mathbf{I} - \mathbf{F}_i)(\mathbf{R}\mathbf{p}_i + \mathbf{t}) \tag{2}$$

where F_i is given by

$$F_i = \frac{\mathbf{v}_i \mathbf{v}_i^T}{\mathbf{v}_i^T \mathbf{v}_i} \tag{3}$$

and $\mathbf{v}_i$ is the projection of the 3D points onto the normalized image plane. In Eq. 2 $\mathbf{p}_i$, $\mathbf{R}$ and $\mathbf{t}$ denote 3D feature positions, the rotation matrix and translation vector, respectively. A minimization of

$$E(\mathbf{R}, \mathbf{t}) = \sum_{i=1}^{n} ||\mathbf{e}_i||^2 \tag{4}$$

is then performed. The algorithm is known to be very robust to the effects of noise [7].

4 Estimation of Face Geometry

The face model was a wire-mesh model with canonical face shape [10]. Because there is variability in head shape among people, the model was modified to fit the specific head-shape of each subject. This was accomplished by an iterative procedure. Ten images were selected from each subject to estimate the the face geometry. An initial estimate of subject pose was obtained using the face model with canonical geometry. The camera's field of view parameter, which affects the perspective geometry, was first estimated using standard values from current imaging devices. Given the camera properties and the pose, we used perspective geometry equations to recover the true position in 3D of the labeled features. This gives us a set of points in 3D which we know are part of the face. Radial basis functions (RBF) are then used 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 compute the displacements for the unconstrained vertices j. We then apply 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

$$f(\mathbf{p}) = \sum_i \mathbf{c}_i \phi(||\mathbf{p} - \mathbf{p}_i||) \tag{5}$$

to all vertices p in the model, where ϕ is an RBFs. 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 $\sum_i \mathbf{c}_i \mathbf{p}_i^T = 0$.

After the geometry of the model is modified, we then re-estimate pose and camera parameters using the new face geometry. Typically 2 or 3 iterations of this process sufficed to converge. After convergence we fix the face geometry model and camera parameters and proceeded to estimate the pose of the entire set of images from a given person.

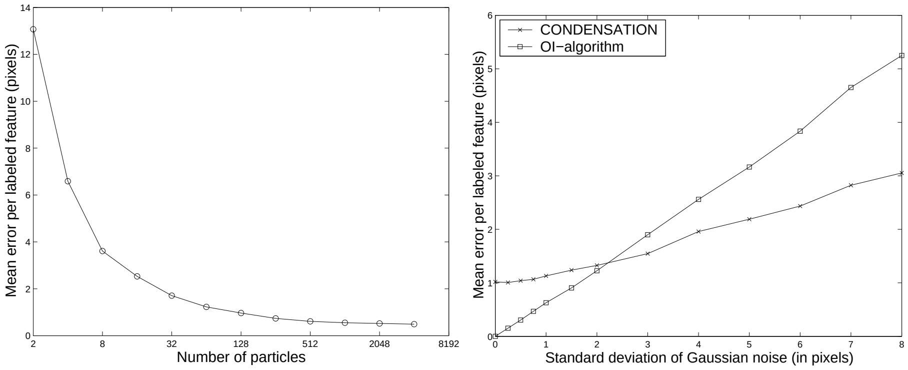

Figure 1: On the left, the performance of the particle filter is shown as a function of the number of particles used. On the right the performance of the particle filter and the OI algorithm as a function of noise added to the true positions of features.

5 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 14 facial features back into the image plane and ground truth positions obtained with manual feature labels. Figure 1 (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 14 facial features. Figure 1 (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.

6 Application: Recognition of facial movements

Figure 2 shows a sample image, the pose estimated by the particle filter, and a resulting image that was warped to a frontal view and to canonical face geometry. With 100 particles, the system works in real time (30 frames per second) on a 1.1 GHz AMD K-7 CPU Linux system running OpenGL on a GeForce2 NVIDIA graphics card.

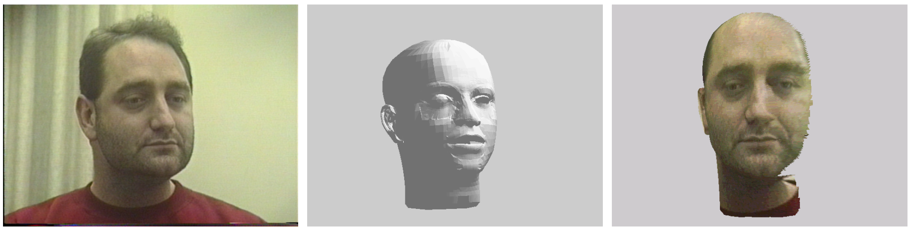

Figure 2: Original image, model in estimated pose and warped image.

3-D pose estimation and warping was applied as a front-end to a system for recognizing spontaneous facial movements [1]. The goal of this system is to recognize each of the 46 facial movements defined in the Facial Action Coding System (FACS) [3] during spontaneous facial behavior. The dataset of our preliminary test consisted of 300 Gigabytes of digitized video from 10 male college students engaged in spontaneous discourse. 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 by human facial expression experts using the Facial Action Coding System.

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 and brow raise. 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 2. 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 [9]. 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 3. The task is quite challenging due to variance in race, the presence of glasses, and noise in the human FACS coding. Note in Figure 3 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 [1].

7 Conclusions

We presented an approach for recognition of spontaneous facial expressions. The main focus of the paper was the exploration of a potential front end to the system to handle in-plane and out-of-plane head movements. We proposed a probabilistic inference approach in which head geometry, camera properties, and 3D pose is simultaneously estimated. Due to the analytical intractability of the resulting stochastic filtering equations, inference is approximated via Markov Chain Monte-Carlo methods (particle filtering). The approach is very promising. First we found that particle filters significantly outperformed some of the most robust deterministic pose estimation algorithms, like the OI algorithm [7]. Second we found that 100 particles were sufficient to achieve accuracies similar to that of human coders. With this number of particles, the system can run in real time in a high end PC. Most importantly, generalization of the particle filtering approach to use automatic feature

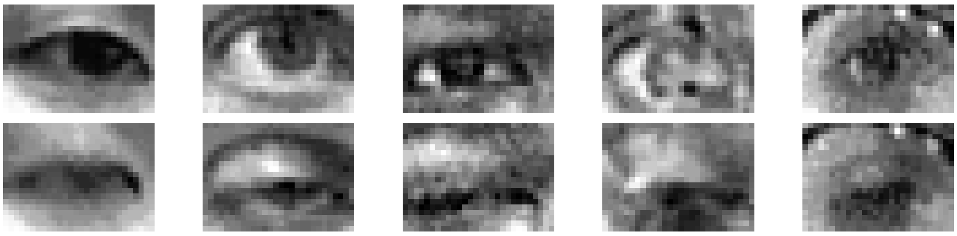

Figure 3: Examples of blink (lower row) and non-blinks (upper row) images after warping. The first three subjects (left 3 columns) had no glasses. The last 2 columns show blinks and non-blinks for 2 subjects with glasses. The prior rotation of the images allowed the same pixel numbers to be used to locate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

detectors instead of hand-labeled features is straight forward. The fact that the particle filtering approach is very robust to uncertainty in feature positions is very encouraging.

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 [8] 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 [4] 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. In spite of the current limitations, the approach presented here is a promising starting point with respect to which future systems may be evaluated.

References

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

[2] 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.

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

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.