Wide-Range, Person- and Illumination-Insensitive Head Orientation Estimation

Ying Wu
University of Illinois (UIUC)
Urbana, IL 61801
yingwu@ifp.uiuc.edu

Kentaro Toyama
Microsoft Research
Redmond, WA 98052
kentoy@microsoft.com

Abstract

*We present an algorithm for estimation of head orientation, given cropped images of a subject's head from any viewpoint. Our algorithm handles dramatic changes in illumination, applies to many people without per-user initialization, and covers a wider range (*e.g.*, side and back) of head orientations than previous algorithms.*

The algorithm builds an ellipsoidal model of the head, where points on the model maintain probabilistic information about surface edge density. To collect data for each point on the model, edge-density features are extracted from hand-annotated training images and projected onto the model. Each model point learns a probability density function from the training observations. During pose estimation, features are extracted from input images; then, the maximum a posteriori *pose is sought, given the current observation.*

1. Introduction

Facial gaze – the orientation of a person's head – gives cues about a person's intent, emotion, and focus of attention. Thus, head orientation can play an important role in vision-based interfaces, where it can provide evidence of user action and lead to more detailed analysis of the face.

The literature on face tracking confirms this belief. A substantial part of facial image processing is concerned with determination of head pose. There are techniques based on tracking blobs of color [4], tracking particular facial features [12, 18], tracking point features [8, 10, 13, 20], following optic flow [2, 6, 19], and fitting textures [3, 5, 17].

Although many of these algorithms are suited for applications such as graphical avatar puppetteering [19] and hands-free cursor control [20], they have constraints that make them inappropriate for other applications. For example, many algorithms are based on tracking image features or computing dense optic flow, and therefore require high-resolution images of the subject. Some systems also apply restrictions on operation, such as per-user initialization, stable lighting conditions, or near-frontal facial poses.

A few authors have tried alternative approaches to overcome these limitations. Pappu & Beardsley build an ellipsoidal texture model of the head and determine pose by matching model projections to live images [15]. This avoids dependency on high-resolution images and tracks the full range of orientations, but nevertheless requires initialization for each subject and static illumination. Gong *et al.* use Gabor wavelet transforms and construct a linear "eigenface" image space for different poses [9]. PCA-based techniques are not well-suited for capturing pose changes, however; it is not clear whether their algorithm generalizes well to recovering more than the single rotational parameter that they consider. Niyogi & Freeman develop an example-based system which trains a neural network from example poses [14]. Pose estimation is treated as a brute-force recognition task and does not take advantage of known geometry. Lastly, Elagin *et al.* use elastic bunch graphs of wavelet feature vectors to determine head pose [7]. Their technique is relatively insensitive to person and illumination, but depends on good resolution.

Our contribution is an algorithm for coarse head-orientation estimation with the following properties:

- It is insensitive to skin color, to glasses or facial hair, and to other common variations in facial appearance.
- It handles large variations in illumination
- It handles side and back views.
- It works under a significant range of image scales and resolutions.
- It requires no per-user initialization.
- The underlying formalism is Bayesian.

We tradeoff these gains with some loss in precision relative to narrow-range, high-resolution techniques (average error is 19 degrees for near-frontal head poses; allowing user-specific initialization, error drops to 10 degrees), but the loss is acceptable for the kind of applications we envision.

Wide-range, coarse head pose is more useful than range-limited, fine head pose in situations where only the subject's approximate focus of attention is of interest. In intelligent environments, for example, head pose offers evidence of a user's communicative intent ("Computer, please turn *that* TV off."). There is also interest in automated cameramen for taping lectures [1] – these systems could be enhanced by knowledge of the speaker's focus of attention to determine pan and zoom parameters (audiences like to see what the speaker is looking at). Finally, the user-independent quality of our system makes it ideal as an initializer for other pose tracking algorithms that require an initial estimate.

2. Algorithm

Our algorithm builds an ellipsoidal model of points, where each point maintains probability density functions (pdfs) of local image features of the head based on training images. The features capture local edge density, which is independent of person and illumination. Some preprocessing for both training images and input images further diminishes effects of illumination. Lastly, we implement maximum *a posteriori* (MAP) estimation using different priors tailored for the cases of global pose estimation and pose tracking, respectively.

2.1. Head Model

In its most general form, our model is an ellipsoid with a set of points on the surface. Each point, indexed by i, is represented by its coordinates, $\mathbf{q}_i$ (lying on the ellipsoid surface), and a pdf representing the belief probability, $p_i(\mathbf{z}|\boldsymbol{\theta})$ – the belief that given a particular 3D pose, the point i will project observation $\mathbf{z}$. "Observations" are local feature vectors extracted from the image. The model itself does not specify what features should be used, although in the next section, we describe our choice for estimation of head pose.

We tried two different placements of the model points. In the first, the points are placed at the intersections of regularly-spaced latitudinal and longitudinal lines. "North pole" coincides with the top of the head. Longitudinal lines are drawn every 10 degrees and latitudes at roughly every 11 degrees, for a total of 562 points. The second point distribution is the same except that the point positions are rotated 90 degrees such that north pole coincides with the nose.

Empirical evidence indicates that the second option captures head information better for the purposes of pose estimation, because it has the greatest concentration of points at the front of the face (see Figure 1). We hypothesize that a point distribution specifically tailored for the texture landscape of heads would fare even better.

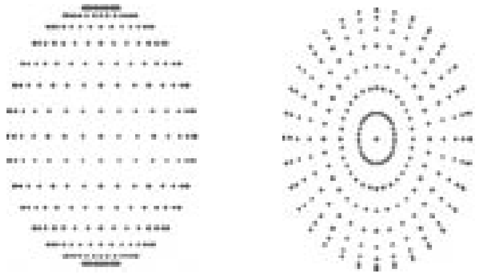

Figure 1. Two point distributions. The one on the right was used in all experiments.

2.2. Features

Nothing in our methodology requires the use of any particular feature, but careful selection of features was nevertheless a critical component of the design. We wanted features which would be nearly universal (for a range of appearances), insensitive to illumination, and still able to distinguish among different orientations. At the same time, these features could not be those that depended on high-resolution imagery or nearly frontal poses.

The first set of constraints rules out the use of local color or brightness information, which varies significantly depending on the person (compare Figure 8(a) with Figure 8(c)). Color is also highly variable under dramatic changes in illumination, in spite of the persistent belief otherwise (see the variation in Figure 7). Skin-color models which purport to capture the racial spectrum of human skin only model skin under a particular lighting condition; and color constancy remains a difficult open problem, especially when illumination within an image itself varies (as in Figure 7(d)). The second set of constraints eliminates the use of precise facial-feature detectors. In severe lighting conditions and in low-resolution imagery, the information to reliably detect an eye, as such, is simply absent (see Figure 8).

The alternative is to use features that are sensitive to local texture. We tried several options, each of which result in feature vectors from the convolution of the following templates applied at each pixel:

1. Gabor wavelets at 4 orientations and at 4 scales each.
2. Rotation-invariant Gabor "wavelets" at 5 different scales. Each template is the sum of 10 Gabor wavelets at 10 orientations such that the result is approximately rotationally symmetric.
3. A Gaussian at a coarse scale, and rotation-invariant Gabor templates at 4 scales.
4. One Gaussian, and Laplacians at 4 scales.

Option 1 was quickly discarded, because convolution coefficients for a given model point changed with orientation. The remaining options were deliberately designed to be rotationally invariant. Options 2 through 4 all worked fairly well, but of these, Option 3 worked the best. The rotation-invariant Gabor template appears to detect high-frequency texture, or *edge density,* as opposed to the Laplacian's tendency to emphasize existence of a single edge. We use Option 3 for the remainder of this paper.

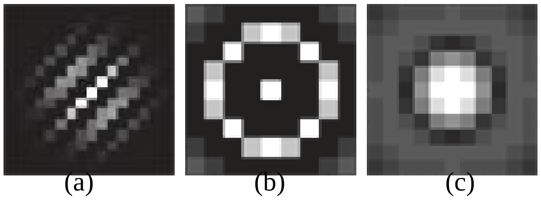

Figure 2. Convolution templates for detecting high textural content: (a) directed Gabor wavelet; (b) rotation-invariant Gabor wavelet; (c) Laplacian.

Finally, because convolution output is strongly correlated with image brightness and contrast, we perform a feature vector normalization which effectively reduces that dependency. This eliminates one parameter from each feature vector but improves performance.

2.3. Image Preprocessing

All of our training images and input images undergo a preprocessing phase that scales the image, eliminates the background, and enhances contrast.

First, given a cropped, rectangular image of a face, we rescale it to a 32×32 image using bilinear interpolation. This step performs the scaling necessary for scaled orthographic registration of the head model with the image.

Next, we apply a circular mask to the image so that any background, non-head portions of the image are ignored. The mask is designed to be conservative: its diameter is 0.7×32 pixels. This reduces the possibility of including background pixels and also effectively bars those parts of the model which would undergo the most foreshortening from contributing to the training and estimation processes. Extreme foreshortening is a problem since it changes the aspect ratio of textures during projection.

Finally, the masked image is histogram equalized.

Our preprocessing phase comprises a subset of the steps performed for face detection by Rowley *et al.* [16]. We avoid elimination of a linear brightness component from the image, since brightness variation is already handled by the normalization of model feature vectors.

2.4. Model Training

We assume we are given a set of annotated training data. The annotation consists of a tight bounding box for the head and an estimate of the rotation matrix, $\mathbf{R}$, that maps the head coordinate system to the camera coordinate system.

Given a set of annotated images, training proceeds as follows: First, the image within the bounding box is preprocessed as described in Section 2.3. Then, normalized feature vectors are computed for each pixel as outlined in Section 2.2. Let $\mathbf{Z}$ be the concatenation of feature vector observations, $\mathbf{z}_j$, which occur at coordinates $\mathbf{p}_j = [x_j \; y_j]^T$ in the image. The feature vector $\mathbf{z}_j$ contributes to the data collected for the model point i_j^* that is the nearest neighbor to $\mathbf{p}_j$ after orthographic projection:

$$i_j^* = \arg\min_{i \in \mathcal{I}} \|\mathbf{p}_j - \mathbf{O}\,\mathbf{R}\,\mathbf{q}_i\|_2, \tag{1}$$

where $\mathbf{O}$ is the matrix that projects $[x \; y \; z]^T$ to $[x \; y]^T$, and $\mathcal{I}$ is the set of model points, $\{i : (\mathbf{R}\,\mathbf{q}_i) \cdot \hat{\mathbf{k}} < 0\}$, that could actually project to the image, assuming model opacity ($\hat{\mathbf{k}}$ is a unit vector along the camera's optical axis, and recall that $\mathbf{q}_i$ is the coordinate of model point i in the model frame).

Once all of the data is collected for each model point, i, we estimate the pdf for that point. In our implementation, we approximate the pdf with a single Gaussian whose mean and covariance coincide with the data. More precisely,

$$p_{i_j^*}(\mathbf{z}_j|\boldsymbol{\theta}) \quad = \quad e^{-(1/2)(\mathbf{z}_j - \bar{\mathbf{z}}_{i_j^*})^T \Sigma_{i_j^*}^{-1} (\mathbf{z}_j - \bar{\mathbf{z}}_{i_j^*})}, \tag{2}$$

where

$$\bar{\mathbf{z}}_{i_j^*} \quad = \quad \frac{\sum_{k \in \mathcal{D}_{i_j^*}} \mathbf{z}^k}{|\mathcal{D}_{i_j^*}|},$$

$$\Sigma_{i_j^*} \quad = \quad \frac{\sum_{k \in \mathcal{D}_{i_j^*}} (\mathbf{z}^k - \bar{\mathbf{z}}_{i_j^*})(\mathbf{z}^k - \bar{\mathbf{z}}_{i_j^*})^T}{|\mathcal{D}_{i_j^*}|},$$

and k indexes the available data $\mathcal{D}_{i_j^*}$ for point i_j^*. This is consistent with a Bayesian approximation of the pdf with a low-information prior.

Figure 3 shows an example of a trained model, using data from 10 subjects. Edge density consistently distinguishes the eyes, nose, and mouth from the rest of the face.

2.5. Pose Estimation

We seek the maximum *a posteriori* pose, given the observation:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{Z}) = \arg\max_{\boldsymbol{\theta}} \frac{p(\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{Z})}, \tag{3}$$

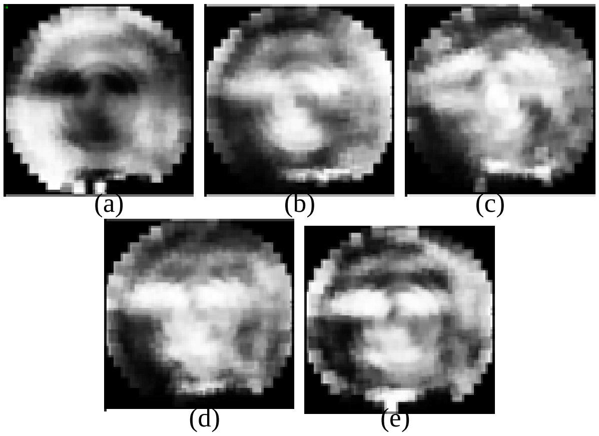

Figure 3. Trained model at a particular pose, after applying (a) a Gaussian, and (b-e) rotation-invariant Gabor masks, at 4 different scales. High intensity corresponds to high mean for the distribution at a model point.

using Bayes' Rule. Since $p(\mathbf{Z})$ is constant, we can ignore the denominator in the right hand side when computing $\boldsymbol{\theta}^*$.

For global pose estimation, our prior is constant, further eliminating $p(\boldsymbol{\theta})$ from Equation 3. In this case, MAP estimation reduces to maximum likelihood estimation. Specifically, we wish to find the pose, $\boldsymbol{\theta}^*$, that maximizes

$$p(\mathbf{Z}|\boldsymbol{\theta}) \quad \approx \quad \prod_j p_{i_j^*}(\mathbf{z}_j|\boldsymbol{\theta}), \qquad (4)$$

with the terms on the right hand side given by the trained model as in Equation 2. The product is only valid if these terms are independent, which they are not – the results nevertheless bear out use of this approximation.

2.6. Tracking Variants

For pose estimation in an online, continuous tracking situation, additional constraints force us to modify this general approach. In particular, tracking often imposes a limit on computational processing. Below, we consider a decimation of the search space. Tracking also provides us with additional information due to spatio-temporal continuity. We incorporate this into the MAP framework by applying priors on $\boldsymbol{\theta}$ that depend on earlier processing.

First, if computational resources are limited, we cannot afford to compute Equation 4 over the entire pose space. Because we are only interested in approximate pose, anyway, we coarsely discretize the search space (equivalent to setting the prior in Equation 4 to be zero or a normalized constant, depending on $\boldsymbol{\theta}$). We therefore only consider poses every 20 degrees of rotation about the vertical axis of the head (corresponding to yaw), every 10 degrees about the axis through the ears (pitch), and +/-20 degrees or 0 degrees about the axis through the nose (roll).

Next, we tried different priors for Equation 3. We examined three: (a) a constant prior ($p(\boldsymbol{\theta}) = c$)); (b) a Gaussian prior, with constant variance and mean at the previous pose estimate ($p(\boldsymbol{\theta}_t) = N(\boldsymbol{\theta}_{t-1}, \bar{\Sigma})$); (c) the previous posterior as prior ($p(\boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t-1}|\mathbf{Z}_{t-1})$). The last alternative is equivalent to the prior used in CONDENSATION [11] with a trivial motion model; it differs also in that we take advantage of our small state space and decouple the sampling scheme from estimates of the posterior probability.

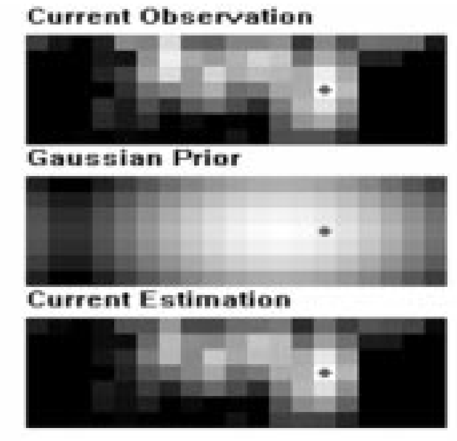

Figure 4. The likelihood outputs shown for pitch and yaw of the head, where whiter values correspond to greater likelihoods.

3. Experimental Results

We collected data from various sources. Some data was captured by a digital video recorder and downloaded onto disk. Others are from movie files of old lectures, where we know little about the camera. Altogether, we used 16 sequences of 11 different people, under different illumination conditions, and at varying distances from the camera.

"Ground truth" pose was determined by hand because many of the data sequences were from prerecorded video. This introduced minor errors during both training and testing. For training, model points tend to accumulate some data meant for their neighbors, which results in blurring of the model. For testing purposes, all errors of the algorithm are measured with respect to the annotation. On one sequence in which true pose was measured by a Polhemus device, true errors (Polhemus pose - hand annotation) and reported errors (estimated pose - hand annotation) were within 13 degrees 90% of the time.

All algorithms were implemented on a 450MHz, double-

Annotation within:		$0° - 45°$		$45° - 90°$		$90° - 135°$		$135° - 180°$	
Training	Testing	Rot Y	Rot X	Rot Y	Rot X	Rot Y	Rot X	Rot Y	Rot X
1 person	same person	10.4	5.7	14.8	6.8	16.9	5.9	28.5	8.7
10 people	different person	19.2	12.0	33.6	16.3	38.0	15.7	47.5	13.2
1 person	different person	21.2	13.7	35.1	17.2	50.6	12.7	70.5	10.9

Figure 5. Average estimation errors.

processor Pentium II PC. Our initial implementation runs at 3 to 5 Hertz, depending on the predictive scheme used.

We tried three types of experiments. In the first, we train the model on a single person and test on the same person. In the second, we train a "generic" model using data from many people, and test on individuals. And, in the third set, we train the model on one user and test on another person.

Error results are shown in Figure 5 for results with a weak Gaussian prior. The first row shows results for training and testing on the same person; results are averaged over 11 people. The second row shows the averaged result for 11 instances of training on 10 people and testing on the remaining person. The final row shows results for training on one person and testing on a different person; results are averaged over 11 trials, where for each trained model, one test subject was chosen at random.

Because texture is more stable on the face than in hair, results were far more accurate when all or part of the face was actually visible. Thus, it made little sense to report average errors over an arbitrary sequence that might consist of part frontal-face views and part back-of-the-head views. Instead, we average over four regions of the pose space. The columns in Figure 5 show the range for which errors were averaged. These numbers indicate the angle between the annotated face normal and the camera's optical axis (inverted). This data is plotted for a single subject in Figure 6.

We note several qualitative trends in our results.

Most significant is that we are able to track side and back views at all, though with decreased accuracy. In particular, for the case in which we are most interested (training on many subjects and estimating pose for an individual), we have an average error of less than 20 degrees for near-frontal poses, approximately 36 degrees for side views, and 47.5 degrees when the subject is facing away from the camera. This is as expected, since there is less textural structure in the back of the head.

The best results come when the model is trained using data for the same person who is tracked. Second-best results are observed when many subjects are used to train the model, and the worst results are for training the model using a single person and tracking a different person altogether. This confirms the intuition that two individuals chosen at random tend to differ more than either to the mean.

Errors in rotation about the x-axis are generally lower than those about the y-axis. This is a representational artifact, based on the wider range and coarser discretization for y-axis rotations.

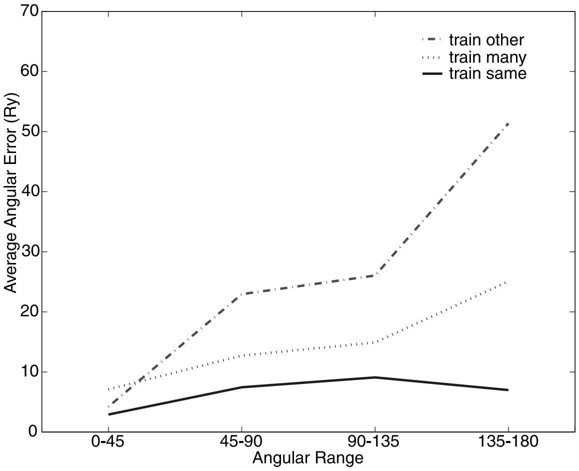

Figure 6. Differences in estimation errors due to training set, for a typical subject.

Finally, we add that the results shown here are numerically close to results when we use a constant prior. That is, the predictive Gaussian prior adds very little information. In part, our sequences are responsible for this result, since many contain frames that are up to 1 second apart, and thus lose spatio-temporal smoothness. On the other hand, global estimation is fairly reliable in itself, testifying to the soundness of our underlying model.

To give a better feel for the performance of our algorithm under some extreme situations, we offer several images of successful pose estimation under a wide range of circumstances. In Figure 7, we show the same speaker tracked under very different illumination conditions – of course, with no model adaptation. Figure 8 shows some difficult cases on other subjects. In each figure, the overlaid ellipse and line indicate the coronal plane of the head that passes through each ear and the normal to that plane. The model used in all of these images was trained on multiple subjects.

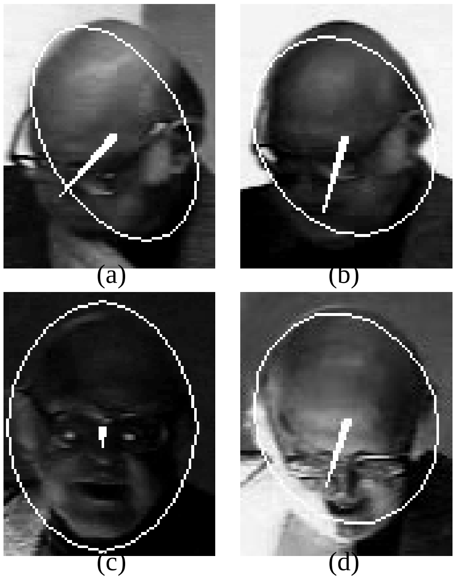

Figure 7. Pose tracking under large variation in lighting conditions (cropped images shown).

4. Conclusion

We have presented an algorithm for performing wide-range, person- and illumination-insensitive head pose estimation. Our results show considerable robustness to real-world visual perturbations. Estimation errors decrease when the subject is facing the camera. Good results are obtained for a generically-trained model; the best results come from a model trained specifically for the target.

In ongoing work, we allow the generic model to provide noisy supervision for learning a user-specific model. We are also considering applying our framework to non-ellipsoidal shapes for coarse pose estimation of other objects.

Acknowledgments: We thank our reviewers and John Krumm for helpful feedback.

References

[1] G. D. Abowd. Classroom 2000: An experiment with the instrumentation of a living educational environment. *IBM Systems Journal*, November 1999.

[2] S. Basu, I. Essa, and A. Pentland. Motion regularization for model-based head tracking. In *Proc. Int'l Conf. on Patt. Recog.*, 1996.

[3] M. L. Cascia, J. Isidoro, and S. Sclaroff. Head tracking via robust registration in texture map images. In *Proc. CVPR*, 1998.

[4] Q. Chen, H. Wu, T. Fukumoto, and M. Yachida. 3D head pose estimation without feature tracking. In *Proc. Int'l Conf. on Autom. Face and Gesture Recog.*, pages 88–93, 1998.

[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *Proc. European Conf. on Computer Vision*, pages 484–498, 1998.

[6] D. DeCarlo and D. Metaxas. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *Proc. CVPR*, pages 231–238, 1996.

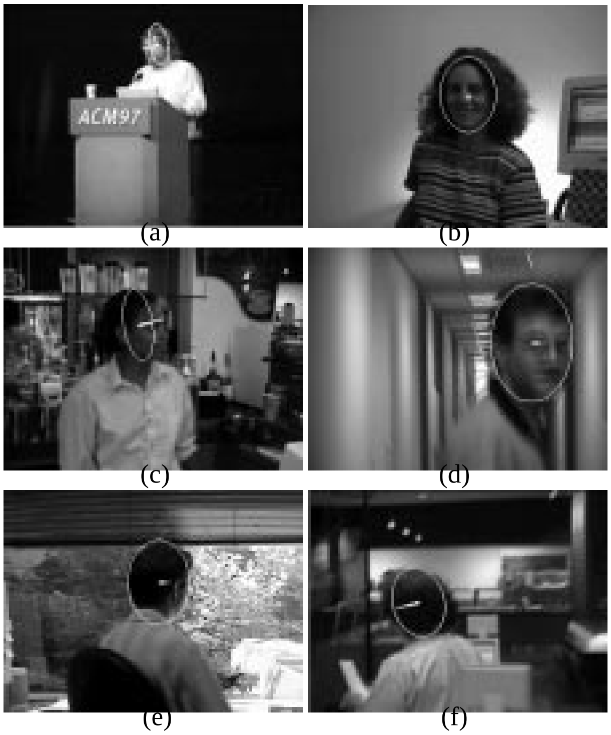

Figure 8. Variation in appearance. (a) facial hair, small scale; (b) back lighting; (c) dark skin, blurring; (d) glasses, ghosting; (e) back lighting, facing away; (f) ghosting, facing away, and with some luck. In (e) and (f) the normals are pointing into the image.

[7] E. Elagin, J. Steffens, and H. Neven. Automatic pose estimation system for human faces based on bunch graph matching technology. In *Proc. Int'l Conf. on Autom. Face and Gesture Recog.*, pages 136–141, 1998.

[8] A. Gee and R. Cipolla. Fast visual tracking by temporal consensus. *Image and Vision Computing*, 14(2):105–114, March 1996.

[9] S. Gong, S. McKenna, and J. J. Collins. An investigation into face pose distributions. In *Proc. Int'l Conf. on Autom. Face and Gesture Recog.*, pages 265–270, 1996.

[10] J. Heinzmann and A. Zelinsky. Robust real-time face tracking and gesture recognition. In *IJCAI97*, pages 1525–1530, 1997.

[11] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. ECCV*, pages I:343–356, 1996.

[12] T. S. Jebara and A. Pentland. Parametrized structure from motion for 3D adaptive feedback tracking of faces. In *Proc. CVPR*, 1997.

[13] T. Maurer and C. von der Malsburg. Tracking and learning graphs and pose on image sequences of faces. In *Proc. FG96*, pages 176–181, 1996.

[14] S. Niyogi and W. T. Freeman. Example-based head tracking. In *Proc. FG96*, pages 374–377, 1996.

[15] R. Pappu and P. Beardsley. A qualitative approach to classifying gaze direction. In *Proc. FG98*, pages 160–165, 1998.

[16] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. Patt. Anal. and Mach. Intel.*, 20(1):23–38, 1998.

[17] A. Schoedl, A. Haro, and I. A. Essa. Head tracking using a textured polygonal model. In *Proc. Wkshp on Perceptual UI*, pages 43–48, 1998.

[18] R. Stiefelhagen, J. Yang, and A. Waibel. Tracking eyes and monitoring eye gaze. In *Proc. Wkshp on Perceptual UI*, Banff, Canada, 1997.

[19] H. Tao and T. S. Huang. Bezier volume deformation model for facial animation and video tracking,. In *Proc. IFIP Workshop on Modeling and Motion Capture Techniques for Virtual Environments (CAPTECH'98)*, November 1998.

[20] K. Toyama. 'Look Ma, no hands!' Hands-free cursor control with real-time 3D face tracking. In *Workshop on Perceptual User Interfaces*, 1998.