Comparative Study of Coarse Head Pose Estimation

Lisa M. Brown and Ying-Li Tian
IBM T.J. Watson Research Center
Hawthorne, NY 10532
{lisabr,yltian}@us.ibm.com

Abstract

For many practical applications, it is sufficient to estimate coarse head to infer gaze direction. Indeed for any application in which the camera is situated unobtrusively in an overhead corner, the only possible inference is coarse pose because of the limitations of the quality and resolution of the incoming data. However, the vast majority of research in head pose estimation deals with tracking full rigid body motion (6 degrees of freedom) for a limited range of motion (typically +/-45 degrees out-of-plane) and relatively high resolution data (usually 64x64 or more.) In this paper, we review the smaller body of research on coarse pose estimation. This work involves image-based learning, estimation of a wide range of pose, and is capable of real-time performance for low-resolution imagery. We evaluate two coarse pose estimation schemes, based on (1) a probabilistic model approach and (2) a neural network approach. We compare the results of the two techniques for varying resolution, head localization accuracy and required pose accuracy. We conclude with details for the implementation specifications for resolution and localization accuracy depending on system accuracy requirements.

Keywords: head pose estimation, gaze direction estimation, head tracking, face tracking.

1. Introduction

Head pose estimation is an important task in human awareness. Examples of applications include dynamic face recognition and facial expression analysis, gaze direction estimation, model based coding for compression and animation, i.e., for low bit rate video teleconferencing and graphical avatars, and hands-free human computer interaction. For many practical applications, it is sufficient to estimate coarse head pose to infer general gaze direction. For most real world applications in which the camera is situated unobtrusively in an overhead corner, the only possible inference is coarse pose because of the limitations of the quality and resolution of the incoming data.

In this paper, we review and compare the work in coarse head pose estimation. Depending on the application, coarse head pose information may be all that is needed. However, in many situations, coarse pose is needed as a prelude to fine pose estimation. We refer to this as “multi-scale” head pose estimation.

2. Background

The majority of work in head pose estimation deals with tracking full rigid body motion (6 degrees of freedom) for a limited range of motion (typically +/-45° out-of-plane) and relatively high resolution data (usually 64x64 or more.) [1,3,4,6,7,15] In addition, such systems typically require initialization to a 3D model. There is a tradeoff between the complexity of this initialization process, the speed of the algorithm and the robustness and accuracy of pose estimation. Although these systems are beginning to achieve real-time computational efficiency, they rely on frame-to-frame estimation and hence are sensitive to drift and require relatively slow and non-jerky motion. All of these systems use relatively high-resolution imagery, measure pose for a limited range (approximately +/- 45° out-of-plane rotations), require some initialization and are sensitive to drift. Because these systems require initialization and failure recovery, coarse pose estimation can play an important role in making these systems robust.

For situations in which the subject and camera are separated by more than a few feet, full rigid body motion tracking of fine head pose is no longer practical. In this case, coarse pose estimation is required. For this type of head pose estimation, systems are needed which will bridge the gap between 2D face tracking and 3D rigid motion head tracking. These systems need to:

(1) determine a wider range of pose beyond +/- 45° out-of-plane rotations,
(2) be insensitive to large motions, slow frame rate, and problems of drift,
(3) not require per person initialization,
(4) be capable of using low resolution imagery,
(5) be insensitive to lighting changes and background clutter,
(6) and to run robustly in faster than real-time.

In order to achieve these goals requires learning pose *a priori* from pose-classified ground truth data so that pose estimation can be performed on a single image at

	Krüger 00	**Niyogi 96**	**Rae 98**	**Wu 00**	**Zhao 99**
Range	+/-20 s.1 X +/-20 s.1 X	+/-50 s20 Y +/-45 s30 X	+/-75 s25 Y +/-45 s30 X	+/-180 s20 Y +/-60 s 10 X +/-20 s 20 Z	+/-90 s 10 Y +/- 90 s 10 X
Accuracy	Physical GT .5-.8° X/Y test/train on same subject	Approx GT Exact 48% Near 87%	Approx GT 11° for subject in training set	Approx GT 19-47° Y (depends on angle) 13° X	Physical GT 9-10° per axis
Method	Gabor Wavelet Network	Tree Structured Vector Quantization	Neural Network	Maximum a Posteriori Estimation	Neural Network
Speed	Estimated 5-10 Hz, 450 Mhz Pentium	11 Hz, SGI Indy	1 Hz includes head detection	3-5 Hz, 450 2-processor Pentium II	15 Hz
Resolution	Not discussed	40x30	80x80	32x32	48x48

Table 1. Comparison of research in head pose classification based on learning

any time. A small number of researchers have pursued this type of methodology and their results can be seen in Table 1. These systems are based on either a statistical classification or neural network. The table shows the range of pose that is measured, in degrees, with the associated step size. For example +/-20 s10 Y, indicates a range from –20 degrees to +20 degrees with a step size of 10 degrees, i.e., rotations of -20,-10,0,10, and 20 degrees about the Y (or vertical) axis.

As can be seen from the table, these systems can each address some of the five above-mentioned requirements. However, several still require per person initialization, are not real-time, require relatively high resolution, and cannot deal with the full range of human head pose. The system designed by Wu & Toyama [13] appears to best satisfy the requirements. This Bayesian system models the probability of each pixel based on *a priori* data. For each pixel, a feature vector, based on the edge density at that pixel, is computed for ground truth data. The pose is estimated by maximizing the *a posteriori* probability. The system is near real-time, can run on images whose resolution is as low as 32x32, and can estimate a full range of poses including the back of the head.

In Table 1, we also report the accuracy claims for each project. We refer to physical ground truth (GT) if an external physical sensor was applied such as the electro-magnetic sensor used in [15] or the robotic arm used by [5]. Approximate ground truth (GT) refers to manual human annotation, which, not surprisingly, is not as accurate. Since we are reporting results on classification methods, each method is limited by the pre-defined step size between classes. The method of [5] reports the highest accuracy; this result is achieved by using the same subject (a doll) for training as testing, a very small step size, and a unique representation based on a Gabor wavelet network. However, this method performs fine head tracking based on high-resolution data. The other four methods in the table perform coarse pose estimation.

We did our best to report the accuracy as presented by the investigators. However, since each system classified pose into a different set of ranges, it is very difficult to compare these values. For this study, we would like to systematically make this comparison, quantify the accuracy achievable for very low resolution and the sensitivity to head localization error.

3. Comparative Study

We have chosen to explore the relative merits of two different approaches to coarse head pose estimation: a probabilistic model approach (PM) based on the work of [13] and a neural network approach (NN). In the following two sections of the paper we will describe each of these two methods in detail. We will then compare the performance of these two approaches based on the resolution of the images, the head localization accuracy and required output accuracy.

For both approaches we have used the CMU Pose, Illumination, and Expression (PIE) Database of Human Faces for our ground truth data [11]. This database contains images of 68 people under 13 poses, 43 different illumination conditions and 4 different expressions. In our study, we only use 9 poses of neutral expressions, from –90 to +90 degrees about the vertical axis and natural room lighting.

Different poses were acquired by the simultaneous acquisition of different static cameras positioned around the room. Subjects were asked to look directly at the center camera. Therefore, *frontal pose was defined by the subject.* This clearly introduced some error in absolute pose measurements.

We semi-automatically extracted the rectangular bounding box for each image using normative head size information, skin color detection, and eye/nose positions. The bounding box information can be found on our website[deleted reference for blind review].

4. Probabilistic Model Approach

The probabilistic model approach we used was based on the work of [13]. This method builds a probabilistic model for each pose using several image-based features and determines the pose of an input image by computing the maximum *a posteriori* pose. Their algorithm uses an 3D ellipsoidal model of the head to represent the pose information.

Because our ground truth data is 2D imagery from a small number of poses, we did not use a 3D model to represent the information. Our storage requirements are minimal and since we ultimately determine the maximum *a posteriori* pose using 2D images, a 3D model would only decrease the accuracy.

Before computing image-based features, the head is located. The images are converted to grey-scale, histogram equalized and reduced to the same resolution. Each pixel in each image-based feature, is assumed to be independent and normally distributed. The mean and covariance is computed based on the training data.

Wu & Toyama use 4 image-based features: convolution with a coarse scale Gaussian and convolution with rotation-invariant Gabor templates at four scales. They experimented with other sets of features based on Laplacians, with and without the Gaussian and this was found to work the best. We used a similar set. For our rotation-invariant Gabor templates, we used the sum of 4 orientations (0,45,90,135 degrees). We found the most effective set of features to be composed of convolution with a 3x3 Gaussian mask, and convolution with 3 rotation-invariant Gabor templates with frequencies 0.5,0.25, 0.0125 and scales of 1,2, and 4 respectively.

The first four images of Figure 1 show the average image for each of the four features based on the first 34 faces in the PIE database for a pose of 22.5 degrees (frontal is 0 degrees.) The rightmost four images show the respective standard deviation images.

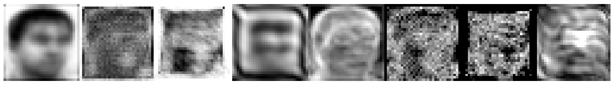

Figure 1. Average images (left) and standard deviation images (right) for each of 4 features for 34 faces in the PIE database, near frontal view

To determine pose, we compute the maximum a posteriori pose $\boldsymbol{\theta}^*$ given the observation **Z**, using Bayes rule:

$$\theta^* = \arg\max_{\theta}\ p(\theta \mid Z) = \arg\max_{\theta}\ \frac{p(Z \mid \theta)\, p(\theta)}{p(Z)}.$$

For the tests performed here in which only static images are used, we assume $p(\theta)$ is constant. Since $p(Z)$ is also constant, the MAP estimation reduces to maximum likelihood estimation,

$$\theta^* = \arg\max_{\theta}\ p(Z \mid \theta) \approx \prod_j \prod_i p_j(z_i \mid \theta),$$

in which we take the product over all features **j** and all image locations **i**. This assumes feature and pixel independence, which of course, is not valid.

Since we assume each pixel/feature is normally distributed, the above equation can be simplified by taking the logarithm and finding the pose that has the minimum value of the expression,

$$\sum_j \sum_i \left(\frac{(x_{j,i} - \mu^{\theta}_{j,i})}{\sigma^{\theta}_{j,i}} \right)^2,$$

given the mean **μ** and standard deviation **σ** of the ith pixel and the jth feature for each pose. However, since the underlying distributions are clearly *not* normally distributed, we can improve the accuracy of our estimation using a so-called *robust* statistic.

The choice of M-estimate depends on the distribution of the errors; in our case, this is the scaled difference between the measurement and mean. We experimented with double or two-sided exponentially distributed errors which results in using an M-estimate based on minimizing the mean absolute deviation (rather than the mean square deviation) and Lortentzian distributed errors. Our best results were achieved using a mean absolute deviation without scaling, with a fixed cap on error size. The fixed cap used (120) was the same for all tests. We also found it useful to re-scale each feature image to capture the most relevant information. A fixed re-scaling was performed for both training and testing.

5. Neural Network Approach

Neural networks (NN) have proven to be a useful tool for face localization, face detection, facial expression recognition, hand posture recognition, head orientation estimation etc. [2,8,9,10,12,14]. Rae and Ritter [9] used three networks to do color segmentation, face localization, and head orientation estimation respectively. The inputs of their neural network for head orientation estimation are a set of heuristically parameterized Gabor filters extracted from the head region (80x80). Their system is user-dependent – it works well for a person included in the training data but performance degrades for unseen persons. Zhao & Pingali [14] also presented a head orientation estimation system using neural networks. They used two NNs to determine pan and tilt angles separately. Our system is most similar to Zhao's system. We also histogram equalize to reduce the effects of variable lighting conditions.

After the head is located, the head image is converted to gray-scale, histogram equalized and resized to the estimated resolution. Then we employ a three layer of NN to estimate the head pose. The inputs to the network are the processed head image. The outputs are the head pose angles. We trained one NN to 9 pan angles from –90° to +90° in steps of 22.5.°

6. Results

We compare the results of the two techniques for varying resolution, specified pose accuracy, and head localization accuracy. We also tested the sensitivity to head tilt and the generalizability to different data. We conclude with details for the implementation specifications for resolution and localization accuracy depending on system accuracy requirements.

6.1 Data

From the CMU PIE database, we use 9 poses of neutral expressions, from –90 to +90 degrees about the vertical axis with natural room lighting. Of the total of 68 subjects, the first 34 subjects were used for training (306 images). The remaining 34 subjects were used for testing (306 images). Therefore, no subject appears in both training and testing sets. Figure 2 shows an example of the 9 poses in CMU PIE database.

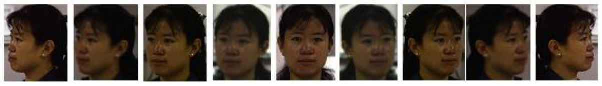

Figure 2. Nine head poses in CMU PIE database from -90° to +90° in steps of 22.5°

6.2 Sensitivity to Different Resolutions

To analyze the sensitivity of the head pose estimation to the image resolution, we down sampled the head region from the original image to six different resolutions: 64X64, 32X32, 24X24, 16X16, 12X12, and 8X8. We did not test images less than 8X8 because it becomes impractical to detect the head when the head size is too small. We first tested both approaches on 9 poses from –90° to +90° in step of 22.5°. The recognition rates of the two approaches for different resolutions are shown in Table 2.

The probabilistic model approach achieved the same level of recognition performance as the neural network approach when the head is 16X16 pixels or larger. When the resolution is lower than that, the recognition rates of the neural network approach are kept at the same level but they decreased rapidly for the probabilistic model approach.

Table 2: Recognition rates for 9 poses for different resolutions by the probabilistic model (PM) approach and the neural network (NN)

Resolution	64	32	24	16	12	8
PM	88%	91%	91%	85%	82%	75%
NN	89%	91%	88%	87%	87%	88%

In order to get the best results, the different window sizes of Gabor templates are required for different resolutions in the probabilistic model approach. We used 3X3 window for 8X8 head image, 5X5 window for 12X12 head image, 7X7 window for 16X16 head image, and 11X11 window size when the head size is larger. For the neural network method, we tested various numbers of hidden units to obtain the best performance. We found that 8 hidden units are enough.

The average accuracy for the probabilistic model approach and the neural network approach for 32X32 head resolution is 3.6° and 4.6° respectively. However, since this measure is dependent on the discretization of classification space (number of poses and step size between poses) it is difficult to compare the performance with the results from previous investigations.

We tested both approaches on a 770MHz, single-processor Pentium III PC. The procedure of head pose estimation (including resizing the image, conversion to gray scale, and intensity normalization) runs at 31 Hz for the probabilistic model approach and 399 Hz for the neural network.

6.3 Sensitivity to Specified Pose Accuracy

In this experiment, we tested the two methods on 5 poses from –90° to +90° in steps of 45°. The results for the two approaches are shown in Table 3. The recognition rates for 5 poses are much higher, as expected, than the accuracy for 9 poses (Table 2) for both the probabilistic model approach and the neural network method.

Table 3: Recognition rate for 5 poses for different resolutions by PM and NN approaches

Resolution	64	32	24	16	12	8
PM	95%	97%	95%	96%	89%	89%
NN	95%	96%	95%	97%	96%	96%

6.4 Sensitivity to Head Localization Accuracy

As shown by [14] head pose estimation is sensitive to head localization. In general, head pose estimation proceeds after the head is found using a head finder algorithm. In our case, we apply head pose estimation on live video after background subtraction, and silhouette extraction. This is followed by finding extremities and ultimately the head using curvature information. If the head region is segmented incorrectly head pose estimation degrades. It was conjectured that this degradation is particularly sensitive to asymmetric cropping, i.e., when the face was no longer centered in the image.

Therefore, we decided to evaluate the sensitivity of the two systems to different head localization errors. All the tests are based on the 32X32 resolution images for 9 head poses. As shown in the Figure 3, five types of the head localization errors are tested: (a) asymmetric width error (nose side only) (b) symmetric width error (c) asymmetric height error (chin side only) (d) symmetric height error (e) symmetric width and height error. The learning procedure is based on the original head image (the solid rectangular). The performance of the two different approaches are shown in Figure 4a-e corresponding to the 5 types of head localization errors. Along the x-axis is the width/height error ranging from –20% to 20%; negative error indicates a smaller head region. We found that:

(1) Both methods were more sensitive to a smaller head region than a larger head region.

(2) Both methods were more sensitive to the width change than the height change.

(3) The NN method is more sensitive to the head localization error.

6.5 Sensitivity to Head Tilt

When we examined the results of our tests, we observed that all the head pose results for a particular subject (subject id 34) in the CMU PIE database were incorrect. This subject appeared to be looking down rather than straight at the camera, leading us to believe that both methods were sensitive to head tilt. We evaluated both methods using additional data from the PIE database acquired of each subject from cameras placed above and below the center (frontal). In each group, there are 68 images from different subjects. We found that both methods were very sensitive to head tilt – recognition drops by nearly half. The recognition rates decreases even more dramatically for head tilt down.

6.6 Generalizability to Different Data

We also performed preliminary tests of both approaches on live video taken in our laboratory. Background subtraction was performed, followed by head region detection using silhouette information. The original head region was on average 30x30 pixels. The training data used was the same 34 images from the PIE database. Example frames are shown in Figure 5. Pose is shown in the dial at the top of each frame. In Figure 6, a plot of the pose estimation results of the two approaches is shown. The PM method appears to favor non-frontal poses. Despite considerable localization inaccuracy (see frame in upper right of Fig 5) and different lighting conditions, the PM method is able to estimate pose accurately in most cases. The NN method more accurately estimates the frontal pose (frames 130-160) but is noisier; we believe this is because it is more sensitive to the head localization error.

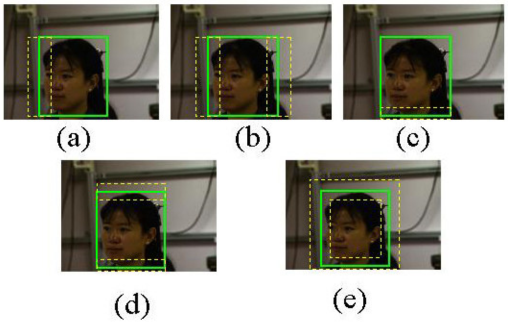

Figure 3. Five different head localization errors

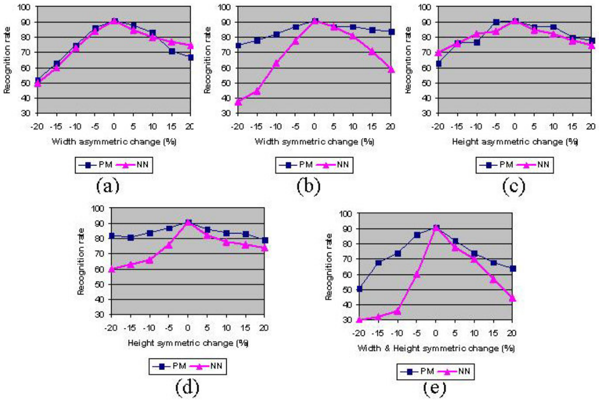

Figure 4. Recognition rate for different head localization errors

7. Conclusions

In this paper, we reviewed coarse pose estimation techniques and compared a probabilistic model approach and a neural network method. In order to compare results, researchers need to train and evaluate their methods on standardized data with the same discretization of classification space. “Average accuracy” is difficult to interpret without this information. We analyzed the different accuracies and sensitivities of two approaches using the CMU PIE database.

The probabilistic model approach was more robust to head localization accuracy but did not perform as well on very low-resolution head images. The neural network method was able to perform pose estimation even for images as small as 8X8 pixels. At this resolution, the neural network was able to determine head pan angle class 88% of the time for 9 poses with a step size of 22.5°. For 5 poses with a step size of 45° recognition was 96%. The neural network was also considerably faster running at over 300Hz on a standard PC. Both methods were very sensitive to head tilt.

Since the PIE database contains data for varying lighting conditions and facial expressions including instances of subjects wearing glasses, we plan to perform further tests to evaluate system sensitivity to these conditions.

In our initial tests, the PM approach appeared to be more extensible to data acquired under different conditions. We conjecture that there is a tradeoff between model complexity, extensibility and accuracy. In general for head pose estimation and tracking (fine or coarse), there is a consistent tradeoff between complex models i.e., 3D geometric models with elaborate initialization or specialized training sets, accuracy, and lack of extensibility – i.e. to people who do not fit the model, for which initialization is not as good or for individuals or lighting conditions which differ from the training set.

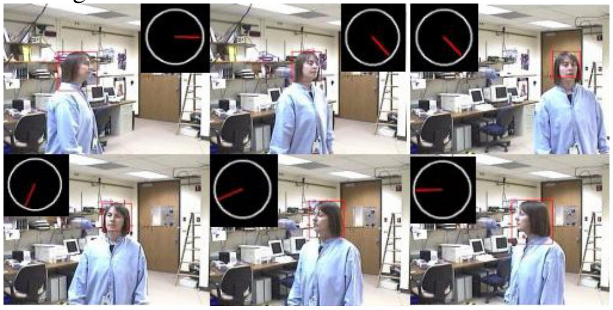

Figure 5. Results from lab

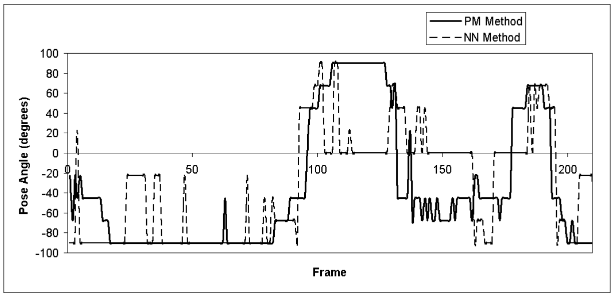

Figure 6. Pose estimation for lab sequence

Acknowledgements

We would like to thank T. Sim, S. Baker, and M. Bsat at CMU for the creation and public access of the CMU Pose, Illumination, and Expression (PIE) Database of Human Faces. We would also like to thank Sharatchandra Pankanti, our colleague at IBM Research for the development of the head finding algorithm and for help with data acquisition. Arun Hampapur and Andrew Senior, the other members of the IBM PeopleVision Project developed the people tracking system that was used extensively.

References

[1] S. Basu, I. Essa, and A. Pentland, "Motion Regularization for Model-Based Head Tracking," in *13th Int'l Conf on Pattern Recognition*. Austria, Vienna, August 25-30, 1996.

[2] J. Bruske, E. Abraham-Mumm, and J. Pauli, "Head-Pose Estimation from Facial Images," in *Proc. of Int'l Neural Network and Brain*, 1998, pp. 528-531.

[3] J. Heinzmann and A. Zelinsky, "3-D Facial Pose and Gaze Point Estimation Using a Robust Real-Time Tracking Paradigm," in *Proc. of the 3rd Int'l Conf. on Automatic Face and Gesture Recognition*. Los Alamitos, CA, April 14-16, 1998, pp. 142-7.

[4] T. Jebara and A. Pentland, "Parameterized Structure from Motion for 3D Adaptive Feedback Tracking of Faces," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[5] B. Kruger, S. Bruns, and G. Sommer, "Efficient Head Pose Estimation with Gabor Wavelets," in *Proc. of the 11th British Machine Vision Conference*, Sept. 11-14, 2000, Vol. 1, pp. 72-81.

[6] M. La Cascia, S. Sclaroff, and V. Athitsos, "Fast, Reliable Head Tracking Under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, April, 2000.

[7] M. Malciu and F. Preteux, "A Robust Model-Based Approach for 3D Head Tracking in Video Sequences," in *Proc. of the Fourth Int'l Conf. on Automatic Face and Gesture Recognition*. Grenoble, France, March 28-30, 2000, pp. 169-74.

[8] S. Niyogi and W. Freeman, "Example-Based Head Tracking," in *Proc. 2nd Int'l Conf. on Automatic Face and Gesture Recognition*. Killington, VT, October 14-16, 1996, pp. 374-378.

[9] R. Rae and H. Ritter, "Recognition of Human Head Orientation Based on Artificial Neural Networks," *IEEE Trans. on Neural Networks*, vol. 9, no. 2, pp. 257-265, March 1998.

[10] H. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, January, 1998.

[11] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces," in *Proc.Int'l Conf. on Automatic Face and Gesture Recognition*. Washington, DC, May 20-21, 2002.

[12] Y. Tian, T. Kanade, and J. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, Feb. 2001.

[13] Y. Wu and K. Toyama, "Wide-Range Person- and Illumination-Insensitve Head Orientation Estimation," in *Proc. of the Fourth Int'l Conf. on Automatic Face and Gesture Recognition*. Grenoble, France, March 28-30, 2000, pp. 183-8.

[14] L. Zhao, G. Pingali, and I. Carlbom, "Real-Time Head Orientation Estimation Using Neural Networks," in *Proc. Int'l Conf. on Image Processing*, Sept. 2002.

[15] Z. Zivkovic and F. van der Heijden, "A Stabilized Adaptive Appearance Changes Model for 3D Head Tracking," in *Proc of the 2nd Int'l Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Realtime Systems*. Vancouver, Canada: IEEE, July 13, 2001, pp. 175-181.