

Estimating Pose and Illumination Direction for Frontal Face Synthesis

Xinyu Huang
University of Kentucky
xhuan4@cs.uky.edu

Xianwang Wang
University of Kentucky
xwangc@uky.edu

Jizhou Gao
University of Kentucky
jgao5@cs.uky.edu

Ruigang Yang
University of Kentucky
ryang@cs.uky.edu

Abstract

Face pose and illumination estimation is an important pre-processing step in many face analysis problems. In this paper, we present a new method to estimate the face pose and illumination direction from one single image. The basic idea is to compare the reconstruction residuals between the input image and a small set of reference images under different poses and illumination directions. Based on the estimated pose and illumination direction, we develop a face synthesis framework to rectify the input image to the frontal view under standard illumination. Experiments show that our estimation method is both fast (less than one second per frame) and accurate (even less than three degrees) and our face synthesis method can generate visually plausible results, in particular for challenging inputs with large pose changes and poor lighting conditions. The synthesized frontal face views increase the face recognition rate significantly from 1.5% to 62.1%.

1. Introduction

Pose and illumination are the two most important factors that affect the appearance of a face. It is known that the intra-class differences caused by changes of pose and illumination are much larger than the inter-class differences caused by the face appearances [14]. To deal with this problem in biometrics, most commercial face recognition systems typically use active illumination and record several face images under different poses for each subject in the gallery. However, there are applications in which such user cooperation and controlled environment are not available. For example, in video surveillance, usually only one face image is available. In these biometric-unfriendly situations, face recognition has to be designed to function across pose and illumination variations. One strategy to address this problem is to use computational approach to synthesize a canonical view, e.g., a frontal face under an ideal illumination before recognition.

In this paper, we present a novel method to estimate the

pose and the direction of illumination simultaneously from one input image. Our method computes the reconstruction residuals between the input image and the images in the data set, assuming the input can be better reconstructed with the group of prototype images that possess the similar pose and illumination direction as the input. This method can provide high accuracy results (around three degrees) from a sparse set of gallery images (e.g., about 20 degrees apart) in less than one second.

With the pose and illumination direction estimation, we form a face synthesis framework to synthesize the frontal faces with similar illumination in the gallery. Based on the Active Appearance Models (AAMs) [6], our approach can synthesize high-resolution frontal faces from one input image with large pose and illumination direction changes. We demonstrate the effectiveness of our complete estimation and synthesis framework on the subjects in the CMU-PIE dataset [16].

2. Related Works

There are many existing methods for pose estimation. In [8, 2], they are divided in five categories: shape-based geometric analysis [9], appearance-based methods [11], model-based methods [6], template-based methods [13], and dimensionality reduction based methods [2, 8]. In shape-based geometric analysis, the head poses are estimated by geometric parameters defined by the facial landmarks. The pose estimation problem in the appearance-based methods is viewed as a pattern classification problem. Many methods in this category only estimate poses in a limited range. The model-based methods fit the input image with a face model such as AAM and a neural network is trained to classify the poses. The template-based methods are based on nearest neighbor classification with texture templates.

Dimensionality reduction based methods are promising based on recent research. With the non-linear dimensionality reduction (NLDR) techniques, pose estimation is performed on a smooth and nonlinear manifold embedding of

the high-dimensional input space. While NLDR is typically considered state of the art, it has its limitations. The sparsely sampled training data, presence of noise in a local area, and curse of dimensionality make the manifold learning generally ineffective [3].

Our estimation method is quite different than these existing approaches. It can be categorized into template-based methods. However, our formulation is very different from [13]. In addition, we estimate illumination directions.

Face synthesis and recognition algorithms across pose and illumination can be divided into three categories, model-based methods [4, 6], appearance-based methods [1], and style-content separation based methods [17, 12]. 3D Morphable Model (3DMM) and AAM are two well-known model-based methods. In [4], 3D shape reconstruction is a fitting problem by minimizing the difference between the rendered model image spanned by a training set and the input image. This method generates face models with good qualities. However, it is computationally expensive and need a good initialization. The major difference between 3DMM and AAM is that AAM uses 2D shapes. In [7], three separate AAMs are trained to locate a face and predict new views. In [1], an illumination cone is built with seven images per person under different illuminations for each pose. With the illumination cone, faces in novel pose and illumination conditions can be synthesized. In [17], Tenenbaum *et al.* first propose the bilinear model to separate content (intra-class) and style (inter-class). In [12], Li *et al.* extend the model to the nonlinear case in which the input space is transferred to a feature space using the Gaussian kernel.

3. Pose and Illumination Estimation

In this section, we present our approach for pose and illumination direction estimation. Our goal is to estimate unknown pose and illumination labels from an input face image that is not in the training database. Although face poses can be easily represented by horizontal and/or vertical angle changes, many parameters are needed to model illumination variations in real scenes. Hence, our illumination estimation is to find the nearest illumination condition in the database. This estimation is particularly useful for our face synthesis framework. For example, illumination could be modeled by the direction of one major light source.

Our approach is based on the assumption that *the new input face image is usually better reconstructed from faces in the database with similar pose and illumination labels*. This assumption is based on two reasons. (1) Face could be linearly represented by a finite set of faces in the training database (e.g., 3DMM and AAM). (2) Intra-class differences are much larger than the inter-class differences as mentioned before. In our approach, the residual is obtained by minimizing the residual function:

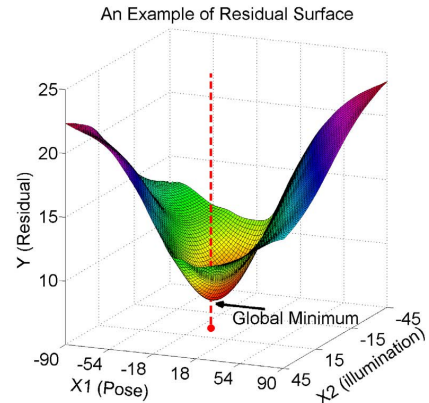


Figure 1. One example of the residual surface. The pose changes from -90° to $+90^\circ$ in 2° increments. The illumination direction changes from -45° to $+45^\circ$ in 2° increments. The new input image is located at pose -7° and illumination 6° .

$$R(W) = \|Z - \sum_{i=1}^n w_i Z_i\|^2, \text{ with } \sum_{i=1}^n w_i = 1, \quad (1)$$

where Z is the input image, Z_i is the image in the database, $W = (w_1, w_2, \dots, w_n)$ is the weight vector, n is the number of faces in the database with same pose and illumination, and R is the reconstruction residual. This constrained least squares problem can be computed in a closed form by first computing covariance matrix $C = (Z - Z_i)^T(Z - Z_j)$, and solving the linear system $CW = 1$ for the weight vector W .

3.1. Estimation by Brute-Force Search

The simplest way to estimate pose and illumination of an input image is to compute residuals for all the possible combinations of pose and illumination conditions and the location of the minimal residual is the expectation of pose and illumination. In Figure 1, a residual surface is plotted with horizontal pose and illumination direction changes. By examining residual surfaces from different input images, we observed that the surface should have only one well defined global minimum that location is the expected pose and illumination. Furthermore, the residual gradually increases (with some perturbations) when the pose and illumination are far away from the global minimum. Hence, the residual surface is approximately convex. This nice property of residual surface makes our estimation approach more efficient than the approaches based on NLDR techniques in which poses are estimated on the complex nonlinear manifold of the original input space.

3.2. Estimation by Polynomial Regression

When many variables are used to model pose and illumination variations, it is very time-consuming (e.g., over 10^7 combinations on a dense sampled database) to obtain an accurate estimation by brute-force search. Moreover, most

existing databases are not densely sampled. In this case, the brute-force search method can not generate accurate estimations. Therefore, our goal is to estimate pose and illumination accurately only based on a small sparsely-sampled database.

With a sparse set of database images, one possible method is to estimate the residual surface first by multiple regression and then find the global minimum of the surface, which location is the expected pose and illumination. In general, the relationship between the residual and pose and illumination is

$$\mathbf{Y} = f(x_1, x_2, \dots, x_m) + \varepsilon, \quad (2)$$

where \mathbf{Y} is the univariate response vector that represents the computed residuals, (x_1, x_2, \dots, x_m) are m variables used to model pose and illumination variations, and ε is a random error term. The form of the response function f is unknown, we could approximate it with a 2- or high-order polynomial model.

However, this straightforward method is not suitable in practice. First, we still need to compute quite a large number of residuals to estimate regression coefficients accurately. Second, it is not easy to choose a regression model. In order to obtain the best model, we may need to evaluate many candidate regressors including interaction variables by using the F -statistic to test the significance of regression.

In our approach, we simplify the multiple regression problem by reducing the original high-dimensional space to m 2-dimensional spaces (i.e., $(\mathbf{Y}, x_1), (\mathbf{Y}, x_2), \dots, (\mathbf{Y}, x_m)$). That is, we project all the residuals on the (\mathbf{Y}, x_i) plane, where $i \in [1..m]$, and apply a simple curve fitting to estimate the “sub-interval” minimum. The reduction can be done because, first, convexity is preserved by the projections, and second, we only apply the curve fitting on the lower boundary (i.e., an approximate convex hull) of each projection. The scatterplots with fitted curves in Figure 2 show the response \mathbf{Y} against each variable x_i in turn.

One way to compute this convex hull is to simply sort all the residuals for each value of x_i and select the minimum residuals, which forms, for example, the red lower boundary in the first panel in Figure 2. However, even in a sparsely sampled database, it could be quite expensive to compute the residuals for all parameter combinations. For example, the residuals need to be computed over 6,000 images (for 20° intervals over pose and illumination directions). In practice, we use an approximation to achieve a faster speed. We observed that the residuals on the lower boundary correspond to the database images that closely resemble the input image. So instead of sorting the reconstruction residual, we sort the appearance residual, which simply involves the computation of the sum of squared per-pixel differences. The fitted curve based on this sampling method is the dark-colored curve in each panel in Figure 2. The location of

the global minimum of the fitted curve is the estimation of pose and illumination. The detailed algorithm is shown in Algorithm 1.

Algorithm 1 Pose and Illumination Estimation

Input: A new face image z .

Output: Estimated pose and illumination.

1: **Compute reconstruction residuals.**

(a) Compute Euclidean distances between z and the images in the sparsely sampled database.

(b) Find the k nearest neighbors (knn) of z along each variable direction x_i , i.e., $knn(z) = \{knn(x_1, \dots, x_{i-1}, a, x_{i+1}, \dots, x_m)\}$, for a is in the discrete value set of x_i .

(c) Compute residuals for the sampled data points by solving Eq. (1).

2: **Fit curves between the residual and each variable.**

Curve fitting between \mathbf{Y} and \mathbf{X}_i ($i = 1, 2, \dots, m$) by Weighted Polynomial Regression (3-Order),

$$\hat{\beta} = (\mathbf{X}_i^T \mathbf{W} \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{W} \mathbf{Y}, \quad (3)$$

where $\mathbf{X}_i = (\mathbf{x}_i^1, \dots, \mathbf{x}_i^n)^T$ is the input matrix with $\mathbf{x}_i^j = (1, x_i, x_i^2, x_i^3)^T$, $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_3)^T$ is the vector of coefficients, $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$ is the weight matrix, and w is expressed as Gaussian kernel function,

$$w(x) = \exp(-(x - x_{min})^2/\lambda), \quad (4)$$

where x_{min} is the position of current smallest residual, and λ is the parameter to control the width of the kernel.

3: **Find the global minimums of the fitted curves.**

Solve the following equation by quasi-Newton (BFGS) algorithm for $i = 1, 2, \dots, m$.

$$\min_{x_i} g(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \hat{\beta}_3 x_i^3, \quad (5)$$

such that $lb \leq x_i \leq ub$, where lb and ub are lower and upper bounds for x_i . The position of the current smallest residual is the initial guess. The estimated minimums $\hat{x}_1, \hat{x}_2, \dots, \hat{x}_m$ are the expected pose and illumination.

4. Frontal Face Synthesis

In this section, we present a method for frontal face synthesis by using our estimation result. The goal of our synthesis framework is to improve the recognition accuracy under varying poses and illuminations, especially for the large pose and illumination changes. Our method can be formalized as learning and applying a transformation between two face classes under different pose and illumination conditions. This is similar to the methods in these papers [18, 7]. Most of them focus on the pose transformations, we included illumination transformations too.

In our method, we use multiple AAMs as the face models to extract facial features and compute shape and texture parameters. The shape parameters are computed by applying PCA to the aligned landmark points that outline the main facial features.

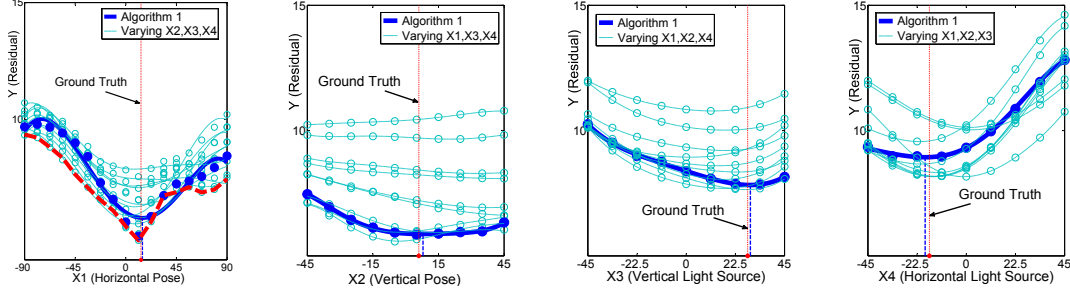


Figure 2. Scatterplots with fitted curves for the reconstruction residuals. The light-colored curves in each panel fit to the residuals that are computed by different values of other three variables. The dark-colored curve in each panel fits to the residuals sampled by Algorithm 1. The red curve in the first panel gives the lower boundary of all the residuals.

$$s = \bar{s} + \sum_{i=1}^n s_i \alpha_i \quad (6)$$

where α_i is a set of shape parameters, \bar{s} is the mean shape, and s_i are n shape template vectors. The texture parameters are obtained by applying PCA on the warped shape-free textures.

$$t = \bar{t} + \sum_{i=1}^n t_i \beta_i \quad (7)$$

where β_i is a set of texture parameters, \bar{t} is the mean texture, and t_i are n texture template vectors.

Figure 3 shows this synthesis framework. During the online synthesis stage, pose and illumination direction are first estimated by Algorithm 1. The estimation tells us the location of the pose and illumination direction that can best reconstruct the input image. Shape and texture parameters are computed by Equation (6) and (7) based on the template images with similar pose and illumination conditions. Since a class of images with small pose and illumination variations can be considered as a linear object class [18], the shape and texture parameters almost remain same. However, when there are large changes between the input image and the target image (e.g., profile view and frontal view), the linear class assumption is not valid and a transformation of parameters is needed before the synthesis. In our framework, the transformation of parameters is learned during the offline training stage using Multivariate Linear Regression (MLR). An important step in the original AAM that combines the shape and texture parameters together into one set of appearance parameters is not suitable in our learning stage. In order to learn an accurate transformation, it is necessary to learn the transformation of shape parameters and shape-free texture parameters separately. Furthermore, textures should not be normalized such that illumination variation remains.

5. Experimental Results

5.1. Data sets

Two data sets are used in our experiments. **(1)** Face images with pose and illumination changes generated from

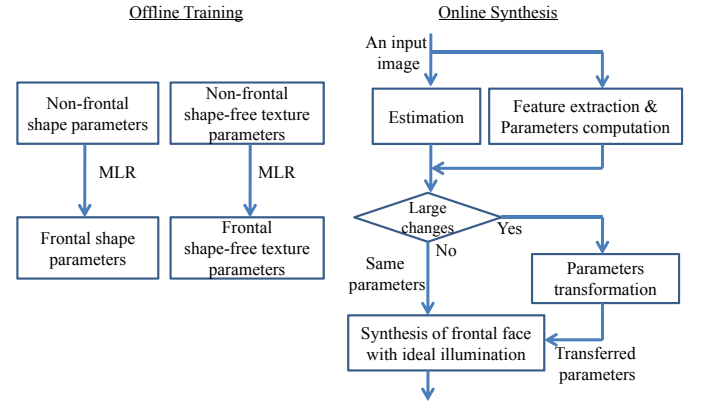


Figure 3. The frontal face synthesis framework.

3D face scans [5]. The pose changes from -90° to $+90^\circ$ horizontally and from -45° to $+45^\circ$ vertically. The illumination direction is changed from -45° to $+45^\circ$ horizontally and from -45° to $+45^\circ$ vertically. **(2)** CMU-PIE database [16]. We choose a portion of the database, which contains face images with neutral expressions captured from 4 cameras (camera index: 34, 27, 05, and 37) under 3 illumination directions (flash index: 09, 11, and 21).

5.2. Estimation Results

In the experiments of estimation of pose and illumination, 5 different individuals from data set (1) are used as the training data set. The speed for estimating one cropped input image is between less than 1 second and 2 seconds based on the Matlab implementation.

Table 1 shows the estimation accuracy when only horizontal pose changes under the similar illuminations are considered. Five different individuals from data set (1) are used for the testing. To our knowledge, the best estimation results are from [2, 8], which only estimate horizontal poses. The accuracy of our method is similar to the results from [2] and almost remains the same when the pose increments are as large as 20° . However, the training data for [2, 8] must be densely sampled (e.g., 1° and 2° increments in [8] and [2]

Pose Range	Illumination Range	Average Error (Pose)	Standard Deviation (Pose)	Average Error (Illumination)	Standard Deviation (Illumination)
$\pm 90^\circ$, I-10-H	$\pm 45^\circ$, I-4-H	2.91°	2.76°	7.32°	6.36°
$\pm 90^\circ$, I-20-H	$\pm 45^\circ$, I-10-H	3.35°	3.63°	8.27°	7.19°
$\pm 90^\circ$, I-20-H	$\pm 45^\circ$, I-10-H	3.26°	3.01°	5.25°	4.46°
$\pm 45^\circ$, I-10-V	$\pm 45^\circ$, I-10-V	6.16°	7.01°	9.12°	10.40°

Table 2. Average estimation error with varying pose and illumination. $\pm 90^\circ$, I-2-H, indicates horizontal (i.e., H) changes from -90° to $+90^\circ$ with 2° increments.

Method	Pose Increments (horizontal)	Average Error
(a)	2°	3.34°
(b)	2°	3.15°
(b)	10°	3.25°
(b)	20°	3.63°
(c1)	2°	5.66°, 6.59°, 8.21°
(c2)	2°	5.02°, 2.11°, 1.44°
(d)	1°	1.65°

Table 1. Performance comparison of different methods with varying pose increments. (a) Brute-force method, (b) Algorithm 1, (c1) BME using Isomap, LLE, and Laplacian Eigenmap with 3 dimensions of embedding in the grayscale feature space [2], (c2) BME using Isomap, LLE, and Laplacian Eigenmap with 100 dimensions of embedding in the grayscale feature space [2]. (d) Locally Embedded Analysis (LEA) with 20 dimensions of embedding [8].

respectively) for manifold learning. This could be one limitation for the manifold learning when we need to estimate vertical poses and illumination directions. Table 2 shows the estimation accuracy from 5 different individuals from data set (1) when pose and lighting directions are both considered. It turns out vertical changes are difficult to estimate and the lighting directions are harder to estimate than the poses. In Table 3, we show the estimation accuracy by testing face images from CMU-PIE database on data set (1). The original face images are cropped and transformed to a template (64×64) by three points (the centers of two eyes and the center of the mouth). The three points can be obtained by manual clicks or computed from the AAM model. Here we compute the relative angles that are nearly constant. The results show that the pose estimation is robust over different databases. Although estimations of illumination directions are not as accurate as our pose estimations, it turns out that the illumination estimations still are good enough for our face synthesis framework.

5.3. Synthesis and Recognition

We test our synthesis framework on three different cases that represent small, moderate, and large pose and illumination variations. The face synthesis is evaluated in a leave-one-out manner, i.e., one of 68 individuals is taken out as an unseen face and 67 individuals at the pose and illumination direction that are the nearest to the estimated ones are used

	Average Estimation	Standard Deviation	Computed Values
Pose: 27 \leftrightarrow 34	66.92°	7.27°	66.19°
Pose: 5 \leftrightarrow 27	11.38°	5.52°	16.50°
Illum.: 11 \leftrightarrow 9	10.28°	20.04°	16.53°
Illum.: 9 \leftrightarrow 21	1.89°	5.63°	4.30°

Table 3. Estimation accuracy by computing the relative horizontal angles for different cameras and flashes. Camera numbers are 5, 27, and 34. Illumination numbers are 9, 11, and 21. The last column (Computed Values) is calculated from geometric information provided in the database.

as template images. Results of frontal face synthesis are shown in Figure 4. The results show that our method is able to handle large pose and illumination changes. The time of online synthesis is only several seconds without considering the time for feature extraction using AAM. Figure 5 shows the identification rate based on the traditional PCA method. The identification rates by applying synthesis are significantly increased compared with the rate without synthesis. Our method is better performed than the method in [10], in which the recognition rates using PCA for pose 34, 37, and 5 are about 10%, 46%, and 71% respectively. Our method obtains similar identification rates as shown in [4], which is 70% for 45° left view and 80% for 45° right view using one algorithm from the FRVT 2002 [14] with 87 individuals from the FERET database [15].

6. Conclusions

In this paper, we first develop a new estimation method for face pose and illumination direction. Our method is efficient and robust to estimate pose and illumination direction from one single image based on a small set of reconstruction residuals. We also present a framework to synthesize the frontal face image under an ideal illumination. This framework is based on estimation of pose and illumination direction and learning a transformation between two different face classes before synthesis. The dramatically improved recognition rates with the rectified (frontal-view) face images show the effectiveness of our methods.

References

- [1] A.S.Georghiadis, P.N.Belhumeur, and D.J.Kriegman. From Few to Many: Illumination Cone Models for Face Recogni-

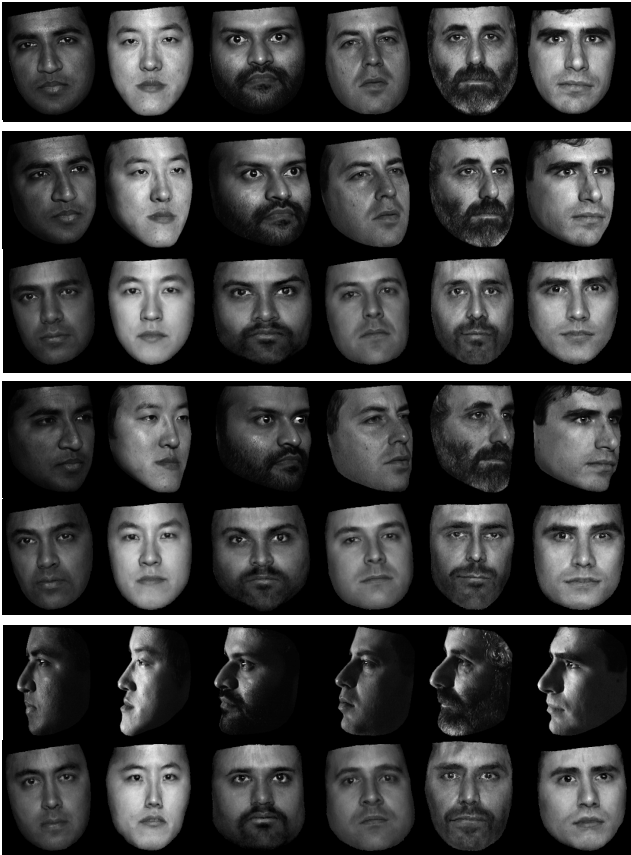


Figure 4. The frontal face synthesis from the CMU-PIE images. 1st row: ground truth of frontal faces (pose:27, illumination:11), 2nd row: input images (pose:5, illumination:11), 3rd row: Reconstruction from 2nd row, 4th row: input images (pose:37, illumination:21), 5th row: Reconstruction from 4th row, 6th row: input images (pose:34, illumination:9), 7th row: Reconstruction from 6th row.

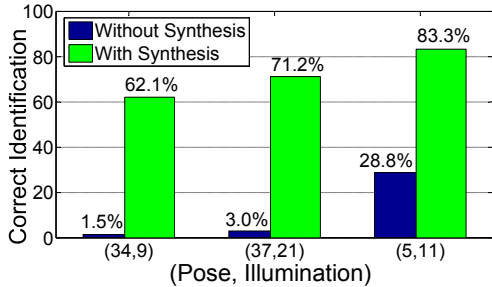


Figure 5. Identification accuracy comparison on the CMU-PIE data set using PCA. Three sets of pose and illumination conditions for 68 individuals are used, they are (34,9), (37,21), and (5,11).

tion under Variable Lighting and Pose. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

- [2] V. N. Balasubramanian, J. Ye, and S. Panchanathan. Biased Manifold Embedding: A Framework for Person-Independent Head Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007.
- [3] Y. Bengio, J.-F. Paiement, and P. Vincent. Out-of-Sample

Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. *Neural Computation*, 16(10):2197–2219, 2004.

- [4] V. Blanz, P. Grother, J. Phillips, and T. Vetter. Face Recognition Based on Frontal Views Generated from Non-Frontal Images. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 454–461, 2005.
- [5] V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. In *SIGGRAPH*, pages 187–194, 1999.
- [6] T. F. Cootes, G. Edwards, and C. J. Taylor. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.
- [7] T. F. Cootes, K. Walker, and C. J. Taylor. View-Based Active Appearance Models. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 227–232, 2000.
- [8] Y. Fu and T. S. Huang. Graph Embedded Analysis for Head Pose Estimation. In *International Conference on Automatic Face and Gesture Recognition*, pages 3–8, 2006.
- [9] Y. Hu, L. Chen, Y. Zhou, and H. Zhang. Estimating Face Pose by Facial Asymmetry and Geometry. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 651–656, 2004.
- [10] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang. Automatic 3D reconstruction for face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 843–848, 2004.
- [11] S. Z. Li, X. Lu, X. Hou, X. Peng, and Q. Cheng. Learning Multiview Face Subspaces and Facial Pose Estimation using Independent Component Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(6):705–712, 2005.
- [12] Y. Li, Y. Du, and X. Lin. Kernel-Based Multifactor Analysis for Image Synthesis and Recognition. In *IEEE International Conference on Computer Vision*, pages 114–119, 2005.
- [13] S. McKenna and S. Gong. Real-Time Face Pose Estimation. *Real-Time Imaging, Special Issue on Visual Monitoring and Inspection*, 4(5):333–347, 1998.
- [14] P. J. Phillips, P. Grother, R. Michaels, D. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002: Evaluation report. Technical Report NISTIR 6965, Nat. Inst. of Standards and Tech., 2003.
- [15] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Raussa. The FERET Database and Evaluation Procedure for Face-recognition Algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.
- [16] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database of human faces. Technical Report CMU-RI-TR-01-02, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, January 2001.
- [17] J. B. Tenenbaum and W. T. Freeman. Separating Style and Content with Bilinear Models. *Neural Computation*, 12(6):1247–1283, 2000.
- [18] T. Vetter and T. Poggio. Linear Object Classes and Image Synthesis From a Single Example Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):733–742, 1997.