Head Pose Estimation From a 2D Face Image Using 3D Face Morphing With Depth Parameters

Seong G. Kong, *Senior Member, IEEE*, and Ralph Oyini Mbouna, *Member, IEEE*

***Abstract*— This paper presents estimation of head pose angles from a single 2D face image using a 3D face model morphed from a reference face model. A reference model refers to a 3D face of a person of the same ethnicity and gender as the query subject. The proposed scheme minimizes the disparity between the two sets of prominent facial features on the query face image and the corresponding points on the 3D face model to estimate the head pose angles. The 3D face model used is morphed from a reference model to be more specific to the query face in terms of the depth error at the feature points. The morphing process produces a 3D face model more specific to the query image when multiple 2D face images of the query subject are available for training. The proposed morphing process is computationally efficient since the depth of a 3D face model is adjusted by a scalar depth parameter at feature points. Optimal depth parameters are found by minimizing the disparity between the 2D features of the query face image and the corresponding features on the morphed 3D model projected onto 2D space. The proposed head pose estimation technique was evaluated on two benchmarking databases: 1) the USF Human-ID database for depth estimation and 2) the Pointing'04 database for head pose estimation. Experiment results demonstrate that head pose estimation errors in nodding and shaking angles are as low as 7.93° and 4.65° on average for a single 2D input face image.**

***Index Terms*— Head pose estimation, 3D face model, morphing, feature disparity minimization, depth estimation.**

I. Introduction

HEAD pose is highly associated with the attention of a human operator to a specific object in the surroundings. Estimating head pose of a human subject has been an active research topic in computer vision because of its importance in many applications such as face recognition, eye gaze tracking, and human-computer interaction. In face recognition, for instance, variations in head pose as well as illumination changes can significantly decrease the performance of face recognition [1], [2]. Head pose estimation plays an important role in gaze estimation [3], where the gaze direction is determined by a combination of the deviation of the pupil center from the eye center and the head pose angles.

Estimating head pose angles from a single 2D image, rather than multiple 2D images or a video sequence, is challenging due to insufficient information. There have been two major approaches [4] to head pose estimation from a single 2D face image: appearance-based and feature-based. Appearance-based approaches [5], [6] attempt to match a portion of the image containing the face to similar face images in the database to estimate the head pose. In this approach, head pose estimation becomes a pattern classification problem. Stiefelhagen [7] compares a query image of the face to a set of face images with known pose angles stored in the database. The head pose of the query face image is determined by the pose of the best matched image in the database. Appearance-based techniques work with low-resolution images, but only a finite number of predefined pose angles can be estimated. Huang *et al.* [8] represent a face image in a low dimensional space and then use classification or regression to determine the head pose of the subject. They proposed a supervised local subspace learning scheme to build local linear models from non-uniform sampled training data. One of major challenges with this head pose estimation scheme is to gather a sufficient amount of training data uniformly distributed across various pose angles. Their performance may decrease considerably when the input face image contains variations such as background, lighting, facial expression, age, and identity. Feature-based approaches [9]–[11] estimate the head pose from correspondences between the features of a face image and those on a 3D face model. Feature-based approaches find reasonably accurate head pose estimates for all three head pose angles, i.e., nodding, shaking, and tilting. The methods proposed by Sun *et al.* [12] and Martins and Batista [13] extract a set of features on a face image to fit the features onto a 3D face model. The head pose is estimated by the amount of rotation during the fitting process. The performance depends on the accuracy of a 3D face model representing the subject. Feature-based methods [14], [15] often require a computationally expensive fitting process as more 3D model templates are added to the set of exemplar 3D models. A Generic Elastic Model (GEM) [22] reconstructs a 3D face model from a 2D face image using a generic 3D model of the same gender and ethnicity as the query subject.

This paper presents head pose estimation from a single 2D face image using a 3D face model morphed from

Manuscript received April 30, 2014; revised August 25, 2014 and November 29, 2014; accepted February 5, 2015. Date of publication February 19, 2015; date of current version March 27, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. David Frakes.

S. G. Kong was with Temple University, Philadelphia, PA 19122 USA. He is now with the Department of Computer Engineering, Sejong University, Seoul, Korea (e-mail: skong@sejong.edu).

R. O. Mbouna is with the Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA 19122 USA (e-mail: oyini@temple.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2405483

a reference 3D face model. A reference model refers to a 3D face of a person of the same ethnicity and gender as the query subject. The proposed method uses a small number of prominent facial features such as eye corners, nose tip, and lip corners extracted from the query face image. An automatic feature extraction method based on Active Appearance Model (AAM) [21] has been used as a primary means to detect the key points. For evaluation purposes, the feature points were selected manually on a 2D face image and mapped onto a reference 3D face model. Head pose angles are estimated by minimizing the disparity between the features on the query face image and the corresponding points on the 3D face model projected onto the 2D space. In case more face images with known pose of the query subject are available, the reference 3D model can be morphed for more accurate model fitting. The refined 3D face model becomes more specific to the query subject in terms of depth errors at the feature points. The morphing process involves multiplying a scalar depth parameter to the depth of a reference 3D face model at each feature point. Optimal depth parameters are found by minimizing the disparity between the features of the 2D query face image and the corresponding features on the morphed 3D model projected onto the 2D space. The proposed morphing process is computationally efficient since the depth of the 3D reference model is represented by the depth at each feature point multiplied by a depth parameter. The use of depth parameters simplifies the morphing process to allow local deformation at each facial feature point for accurate 3D model building. An optimal set of depth parameters are found by minimizing the disparity between the features of the query 2D face image and the corresponding features of the morphed 3D model projected onto the 2D space. The proposed head pose estimation technique was evaluated on two benchmarking databases: 1) the USF Human-ID database for depth estimation and 2) the Pointing'04 database for head pose estimation.

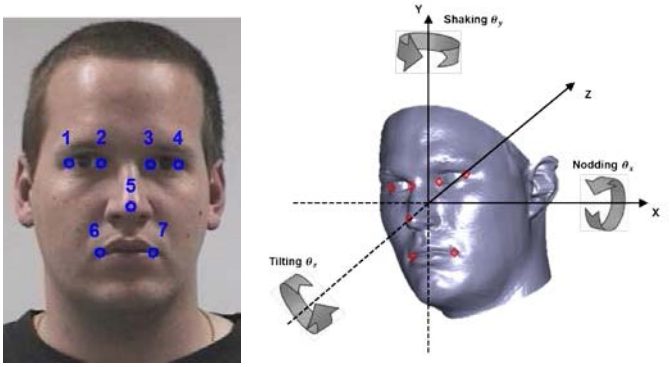
II. Projection of 3D Features Onto the 2D Space

According to the rotation-scale-translation camera model, a projection matrix relates the coordinates of a point in the 3D space to its projection onto the 2D image plane. A mapping of a point (X_0, Y_0, Z_0) in the 3D space onto a point (x_0, y_0) in the 2D image plane can be represented by the projection matrix $\mathbf{P}_0$:

$$\begin{bmatrix} x_0 \\ y_0 \\ 1 \end{bmatrix} = \mathbf{P}_0 \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{bmatrix} = s \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{bmatrix} \quad (1)$$

where s is a scale parameter, and t_1 and t_2 denote the amount of translation. We select N facial feature points on a 2D query face image and N corresponding facial features on the 3D model. The shape of a face can be described by a set of feature vectors $\mathbf{b}_0$ in the 2D space and the corresponding feature vectors $\mathbf{a}_0$ in the 3D space:

$$\mathbf{b}_0 = \begin{bmatrix} x_{0i} \\ y_{0i} \end{bmatrix}, \quad \mathbf{a}_0 = \begin{bmatrix} X_{0i} \\ Y_{0i} \\ Z_{0i} \end{bmatrix}, \quad i = 1, 2, \cdots, N \quad (2)$$



(a) (b)

Fig. 1. Facial feature extraction. (a) Seven feature points selected on a 2D face image, (b) Corresponding features on a 3D face model.

The proposed head pose estimation method works best with a 3D face model of the query subject itself. When a 3D face model of the subject is not available, as in many practical situations, a 3D face model of a subject of the same gender and ethnicity as the query subject is selected as a reference model. This process can be automated with a gender/ethnicity classification routine [23], [24]. Figure 1(a) shows the locations of seven ($N = 7$) selected feature points, two eye corners, nose tip, and two lip corners, used to represent the shape of a human face. Figure 1(b) shows corresponding feature points on a reference 3D face model along with the three rotation angles of head pose; nodding (θ_x), shaking (θ_y), and tilting (θ_z) with respect to the coordinate origin set to the centroid of the feature points. A reference 3D face model was arbitrarily chosen from a group of male Caucasian subjects, which is of the same ethnicity and gender as the query subject. We scale and align the 2D features of the query face image and the feature points on a reference 3D face model. The centroid $\overline{\mathbf{b}}_0$ of the 2D shape is an arithmetic average of the feature vectors $\mathbf{b}_0$.

We normalize the feature vectors by subtracting the centroid from each vector $\mathbf{b}_0$ to move the origin to the centroid and dividing by the norm of the difference. Then the normalized 2D feature vectors are given by:

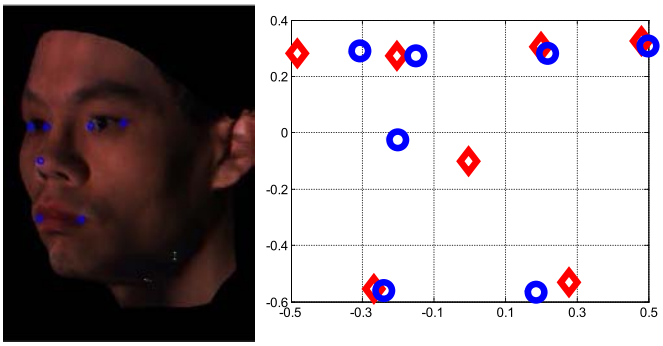
$$\mathbf{b} = \frac{\mathbf{b}_0 - \overline{\mathbf{b}}_0}{\|\mathbf{b}_0 - \overline{\mathbf{b}}_0\|} = \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \quad \overline{\mathbf{b}}_0 = \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} x_{0i} \\ y_{0i} \end{bmatrix}, \quad i = 1, 2, \cdots, N \quad (3)$$

The same normalization procedure is applied to the 3D feature vectors $\mathbf{a}_0$ to obtain normalized 3D feature vectors:

$$\mathbf{a} = \frac{\mathbf{a}_0 - \overline{\mathbf{a}}_0}{\|\mathbf{a}_0 - \overline{\mathbf{a}}_0\|} = \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix}, \quad \overline{\mathbf{a}}_0 = \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} X_{0i} \\ Y_{0i} \\ Z_{0i} \end{bmatrix}, \quad i = 1, 2, \cdots, N \quad (4)$$

Then the projection matrix $\mathbf{P}$ of a normalized feature point on a 3D face model onto the 2D space is given by

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \mathbf{P} \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix}, \quad i = 1, 2, \cdots, N \quad (5)$$



(a) (b)

Fig. 2. Normalization of the feature points in the 2D and 3D space. The 2D features (blue circles) from a face image with a certain pose and the 3D feature points (red diamonds) from a frontal 3D face model involve large disparity. (a) Feature points on a face image, (b) Normalized feature points in the 2D space.

The alignment and normalization process removes the scale factor and translation parameters in the projection matrix expression. Figure 2 shows the feature points from a 2D face image and the corresponding features from a reference 3D face model after the normalization. The two sets of feature points involve large disparity since the query face image has a certain pose while the initial position of a reference 3D model is frontal.

III. 3D Morphing With Depth Parameters

In case more 2D face images with known pose of the query subject are available, the depth of the reference face model can be refined through a 3D morphing process to improve the accuracy of pose estimation. The reference 3D face model is rotated by the pose angles of the given training image to reduce the disparity between the 2D and the projected 3D feature vectors. After normalization of the 2D and 3D feature points, spatial deviations dX and dY are found to be substantially small compared to the depth variation dZ. Only depth morphing is performed since the feature disparity largely comes from the mismatch in depth between the reference model and the query subject. The percentage variations of dX and dY were below 30% of the variation of dZ. (See APPENDIX for details).

In our morphable 3D face model, the depth information at each feature point is represented by a multiplication of the z-coordinate value Z_i with a depth parameter k_i. A feature point in the morphed 3D face model can be represented by

$$\mathbf{a}_i(\mathbf{k}) = \begin{bmatrix} X_i \\ Y_i \\ k_i Z_i \end{bmatrix}, \quad i = 1, 2, \cdots, N \tag{6}$$

Now the morphing process is given by updating the depth parameters $\mathbf{k} = [k_1, k_2, \cdots, k_N]$ until the feature disparity is minimized. The disparity $\mathbf{d}_i(\mathbf{k})$, $i = 1, 2, \ldots, N$, between the features on a 2D face image and the projected 3D features of the reference model is given by

$$\mathbf{d}_i(\mathbf{k}) = \mathbf{b}_i - \mathbf{P}\mathbf{a}_i(\mathbf{k}) = \begin{bmatrix} u_i(\mathbf{k}) \\ v_i(\mathbf{k}) \end{bmatrix}, \quad i = 1, 2, \cdots, N \tag{7}$$

The mismatch between a reference 3D model and the 3D shape of the query subject gives a substantial error in terms of feature disparity. The optimum depth parameter can be found by minimizing the objective function of feature disparity:

$$\varepsilon_0 = \sum_{i=1}^{N} \|\mathbf{d}_i(\mathbf{k})\|^2 = \sum_{i=1}^{N} \left[u_i^2(\mathbf{k}) + v_i^2(\mathbf{k})\right] \tag{8}$$

From the fact that the shapes of human faces share a large amount of overall similarities, a soft constraint needs to be added to penalize the objective function not to produce any degenerate solutions, but geometrically meaningful faces, in the morphing process. Degenerate solutions may occur in the case where the minimum correspondence error between the pairs of points picked on 3D model and 2D face images lead to meaningless 3D shape reconstruction of the face. A penalty function P describes the depth disparity between the reference 3D model and the morphed 3D model at n-th iteration:

$$P(\mathbf{k}(n)) = \sum_{i=1}^{N} (k_i(0) - k_i(n))^2 Z_i^2 \tag{9}$$

This penalty term controls the amount of depth displacement before losing the distinctness of the 3D face structure. The revised objective function is now given by:

$$\varepsilon(n) = \varepsilon_0(n) + \alpha P(\mathbf{k}(n)) \tag{10}$$

with a proper weight α so the optimization converges to a feasible solution. The optimal $\mathbf{k}$ is determined by the depth parameter that minimizes the objective function:

$$\mathbf{k}^* = \arg\min_{\mathbf{k}} \varepsilon(n) \tag{11}$$

We use the gradient descent method to minimize $\varepsilon(n)$ with respect to the parameter vector $\mathbf{k}$. The depth parameters $\mathbf{k}$ is updated in each iteration starting from an initial condition $\mathbf{k}(0) = [1, 1, \cdots, 1]$

$$\mathbf{k}(n+1) = \mathbf{k}(n) - c\nabla_{\mathbf{k}}\varepsilon(n) \tag{12}$$

where c denotes the step size of the gradient decent method. The gradient vector $\nabla_{\mathbf{k}}\varepsilon = \left[\frac{\partial\varepsilon}{\partial k_1}, \frac{\partial\varepsilon}{\partial k_2}, \cdots, \frac{\partial\varepsilon}{\partial k_N}\right]^T$ can be approximated using the difference

$$\frac{\partial\varepsilon}{\partial k_i} = \frac{\varepsilon(k_1, \cdots, k_i + \Delta, \cdots, k_N) - \varepsilon(k_1, \cdots, k_i - \Delta, \cdots, k_N)}{2\Delta} \tag{13}$$

The optimization process is carried out with $\Delta = 0.01$ and $c = 0.001$. And $\mathbf{k}$ is updated until the condition is satisfied: $|\varepsilon(n+1) - \varepsilon(n)| < (0.0001)\,\|\nabla_{\mathbf{k}}\varepsilon\|$.

IV. Head Pose Angle Estimation

Head pose angles are estimated by minimizing the disparity between the features on the query face image and the corresponding points on the 3D face model projected onto the 2D space. Let $\mathbf{B}$ be a $2 \times N$ matrix of a set of normalized feature vectors from the 2D query face image and $\mathbf{A}(\mathbf{k}^*)$ be a $3 \times N$ matrix of corresponding feature vectors normalized

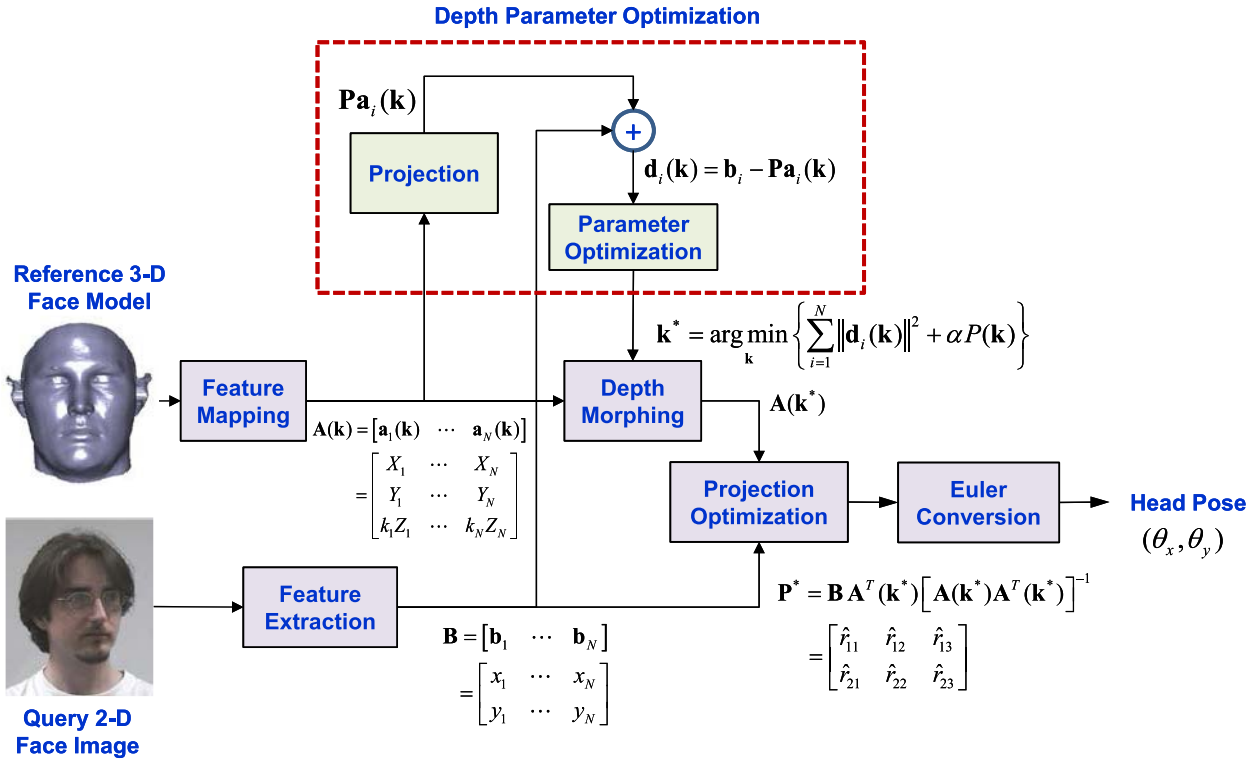


Fig. 3. Schematic diagram of the proposed head pose estimation technique using a single 2D face image.

from a 3D face model morphed from the reference model with the optimal depth parameters:

$$\mathbf{B} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \\ y_1 & y_2 & \cdots & y_N \end{bmatrix} \tag{14}$$

$$\mathbf{A}(\mathbf{k}^*) = \begin{bmatrix} X_1 & X_2 & \cdots & X_N \\ Y_1 & Y_2 & \cdots & Y_N \\ k_1^* Z_1 & k_2^* Z_2 & \cdots & k_N^* Z_N \end{bmatrix} \tag{15}$$

We determine the head pose of the query subject using the rotation matrix of the 3D face model that minimizes the error between the feature matrix $\mathbf{B}$ of a 2D query image and the 3D feature matrix $\mathbf{A}(\mathbf{k}^*)$ projected onto the 2D space using the least-squares minimization. The optimal projection matrix $\mathbf{P}^*$ that minimizes the pose error $\|\mathbf{B} - \mathbf{P}^*\mathbf{A}(\mathbf{k}^*)\|$ is given by

$$\begin{aligned} \mathbf{P}^* &= \mathbf{B}\,\mathbf{A}^T(\mathbf{k}^*)\left[\mathbf{A}(\mathbf{k}^*)\mathbf{A}^T(\mathbf{k}^*)\right]^{-1} \\ &= \begin{bmatrix} \hat{r}_{11} & \hat{r}_{12} & \hat{r}_{13} \\ \hat{r}_{21} & \hat{r}_{22} & \hat{r}_{23} \end{bmatrix} \end{aligned} \tag{16}$$

The least-squares optimization solves for the first two rows of the rotation matrix $\mathbf{R}$ because of the orthographic projection assumption. The last row is given by the cross product of the first two rows $\begin{bmatrix} \hat{r}_{31} & \hat{r}_{32} & \hat{r}_{33} \end{bmatrix} = \begin{bmatrix} \hat{r}_{11} & \hat{r}_{12} & \hat{r}_{13} \end{bmatrix} \times \begin{bmatrix} \hat{r}_{21} & \hat{r}_{22} & \hat{r}_{23} \end{bmatrix}$. For three Euler angles of head pose $(\theta_x, \theta_y, \theta_z)$, the rotation matrix $\mathbf{R}$ can be represented as the product of three matrices representing individual rotation along each axis (17), as shown at the bottom of this page.

Then each head pose angle can be determined by:

$$\theta_x = \tan^{-1}\frac{\hat{r}_{32}}{\hat{r}_{33}} \tag{18}$$

$$\theta_y = -\tan^{-1}\frac{\hat{r}_{31}}{\sqrt{\hat{r}_{32}^2 + \hat{r}_{33}^2}} \tag{19}$$

$$\theta_z = \tan^{-1}\frac{\hat{r}_{21}}{\hat{r}_{11}} \tag{20}$$

In case more than one 2D face image of the query subject are available, the head pose estimation accuracy can be improved. For the first training image, parameter optimization is carried out with an initial value of $\mathbf{k}$, e.g. $\mathbf{k}(0) = [1, 1, \ldots, 1]$. For the next available training image, the optimization process is repeated with the optimum $\mathbf{k}^*$ of the previous round as an initial value.

Figure 3 shows a schematic diagram of the proposed head pose estimation technique from a single 2D query

$$\begin{aligned} \mathbf{R} &= \mathbf{R}_z(\theta_z)\,\mathbf{R}_y(\theta_y)\,\mathbf{R}_x(\theta_x) \\ &= \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta_z\cos\theta_y & \cos\theta_z\sin\theta_y\sin\theta_x - \sin\theta_z\cos\theta_x & \cos\theta_z\sin\theta_y\cos\theta_x + \sin\theta_z\sin\theta_x \\ \sin\theta_z\cos\theta_y & \sin\theta_z\sin\theta_y\sin\theta_x + \cos\theta_z\cos\theta_x & \sin\theta_z\sin\theta_y\cos\theta_x - \cos\theta_z\sin\theta_x \\ -\sin\theta_y & \cos\theta_y\sin\theta_x & \cos\theta_y\cos\theta_x \end{bmatrix} \end{aligned} \tag{17}$$

(a) (b) (c)

Fig. 4. 2D images of three query subjects from USF Human-ID database. (a) A male Caucasian subject (S1). (b) A male Asian subject (S2). (c) A female Caucasian subject (S3).

face image. A set of 2D features are extracted from the query 2D face image. The corresponding 3D features are obtained from the depth information of the reference 3D face model at the same x- and y-coordinates as the 2D features. The 2D and 3D features are both scaled and aligned in the normalization step. The depth information of the morphable 3D face model is represented by z-coordinate of the reference model multiplied by a depth parameter. The depth parameter **k** is optimized to minimize the disparity between the 2D features and the projected 3D features using a projection matrix **P**. The reference 3D face model is morphed to a 3D face model more specific to the query subject using the optimal depth parameter $\mathbf{k}^*$. Then we use the morphed 3D face model and the query face image to find the optimum projection matrix $\mathbf{P}^*$. Finally, we compute the head pose angles from Euler conversion of the optimum projection matrix.

V. Experiment Results

A. Depth Estimation Results

The USF Human-ID database [16] consists of 100 laser-scanned 3D faces of 100 different subjects. Each face model in the database has 75,972 vertices. The database contains 25 female and 75 male subjects of various ethnicities including Caucasian, Asian, Indian, Latino, and African. The scanning process used Cyberware Head and Face Color 3D Scanner that captures an array of digitized points, with each point represented by X, Y and Z coordinates for the surface structure of a face and 24-bit RGB values for color texture.

We observe the patterns of feature disparity and the depth error as an optimization iterates and with more face images for training. A male Caucasian model shown in Figure 1(b) was arbitrarily selected as the reference face model (S0). 2D images of three query subjects are used for testing, a male Caucasian (S1), a male Asian (S2), and a female Caucasian (S3), as shown in Figure 4. Figure 5 shows the trends of feature disparity $\varepsilon(n)$ and depth error $E(n)$ for successive iterations for the three testing subjects. The depth error $E(n)$ at n-th iteration is given by the average error between ground truth depth $\tilde{Z}_i$ and the estimated depth $k_i(n)Z_i$:

$$E(n) = \frac{1}{N}\sum_{i=1}^{N}\left|\tilde{Z}_i - k_i(n)Z_i\right| \tag{21}$$

For all subjects, the feature disparity as well as the depth error decreases as the iteration progresses. This reveals that

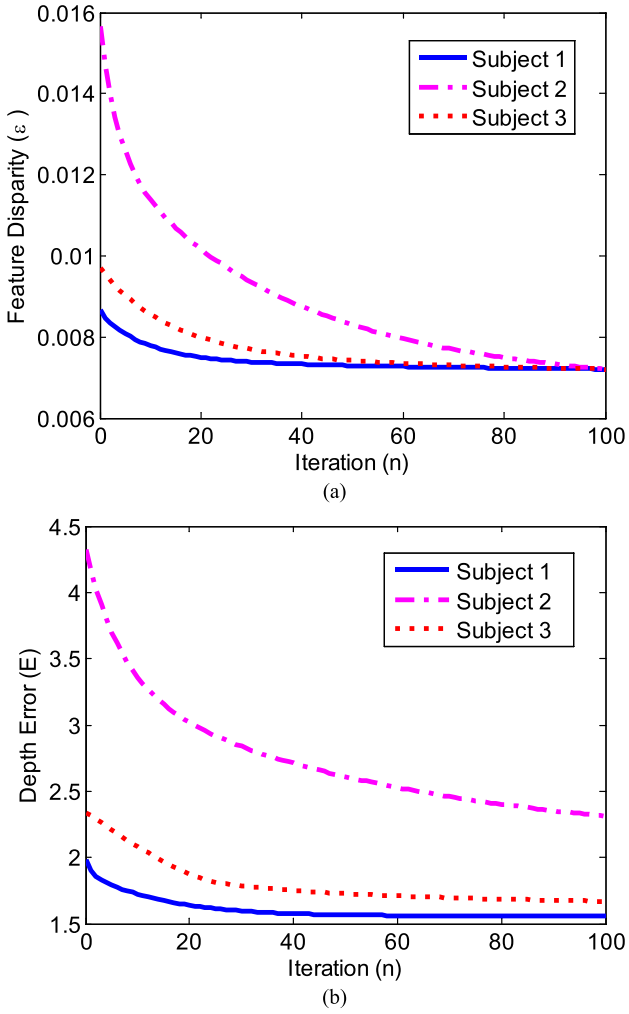


Fig. 5. Convergence of the feature disparity and the depth error. (a) Feature disparity in the normalized 2D face image space and (b) Depth error in the 3D face model space.

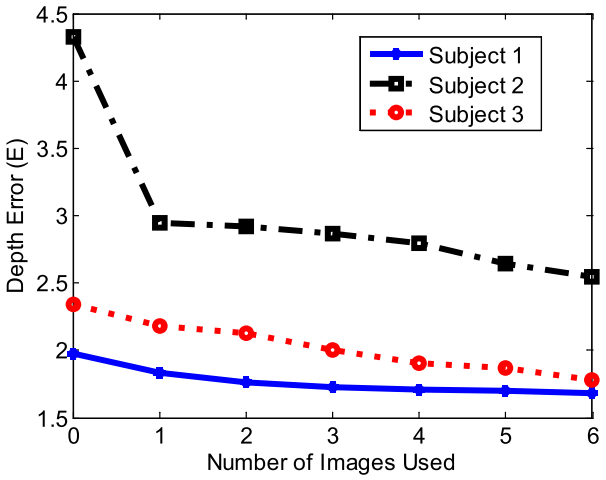


Fig. 6. Depth error as a function of the number of training images of the same person used to morph the reference model.

the morphing process reduces the feature disparity as well as the 3D depth error between the reference model and the 3D model of the query subject. Figure 6 shows the depth error for the three query subjects when multiple 2D face images

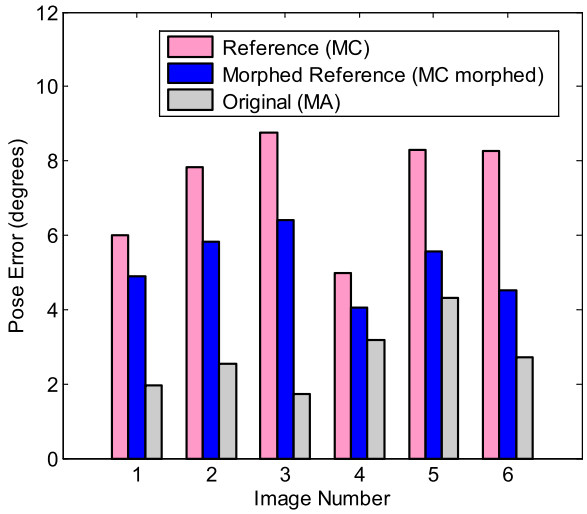


Fig. 7. Pose estimation errors for 6 different images of a male Asian (MA) subject using a male Caucasian (MC) reference model.

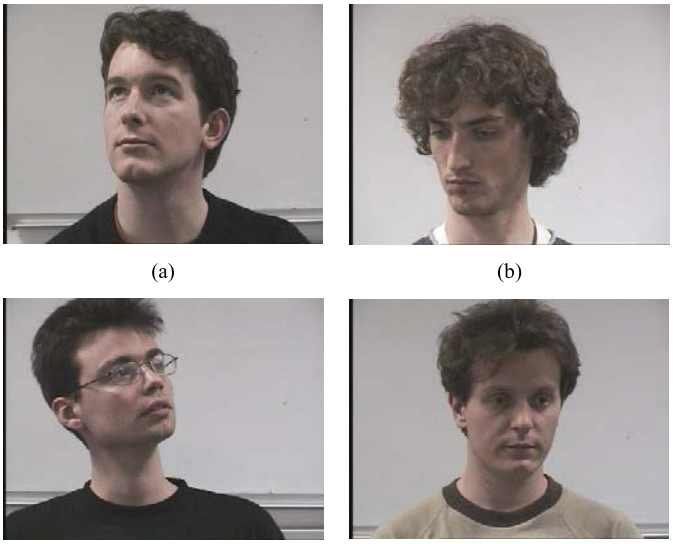


Fig. 8. Four sample images along with ground-truth pose angles of nodding and shaking from the Pointing'04 database. (a) (30°, 15°). (b) (−30°, 30°). (c) (15°, −45°). (d) (−15°, −15°).

of the query subjects were used to morph the reference 3D face model (S0). The depth error decreases as more 2D training images are used to morph the reference 3D model.

B. Head Pose Estimation Results

To evaluate the performance of the proposed head pose estimation technique, we use the mean absolute error (MAE) of nodding and shaking angles of head pose as the evaluation metric. The pose estimation error is computed by averaging the difference between the ground-truth and the estimated pose angles for all images. Figure 7 shows pose estimation errors for 6 different images of a male Asian (MA) subject with different pose using a male Caucasian (MC) reference model from the USF Human-ID database with and without morphing. When a 3D reference model of the original subject was used, the pose estimation error was the smallest. Pose estimation with an MC reference model, being selected from a different ethnic group, produces the biggest estimation errors. The use of morphed MC model using a single 2D non-frontal image of an MA subject produces better results than using the reference model without morphing.

The Pointing'04 database [17] consists of 15 sets of human face images of size 384×288 of 20 to 40 images with various poses. Out of 15 subjects, there are 12 male Caucasians, a female Caucasian, a male Indian, and a female Asian. Five subjects have facial hair and seven are wearing glasses. Each set contains two sessions of 93 images of the same person at different poses. During data acquisition people were asked to look successively at 93 markers. Each marker corresponds to a particular pose. Figure 8 shows four sample images with ground-truth pose angles from the Pointing'04 dataset.

We selected 70 images per subject of 12 MC subjects corresponding to the pose between -45 and $+45$ degrees. Table I summarizes head pose estimation errors for nodding and shaking angles for 12 MC subjects using three different face models: a graphical 3D model designed using the Blender software [18], a MA reference model, and a MC reference model. The use of MC reference model showed best results in pose estimation over the other two models since the query face was also a male Caucasian. Each reference model was morphed using an image of the query subject. The morphed 3D models improved head pose estimation compared with their original models without morphing. The ground-truth pose in the Pointing'04 database was obtained from the orientation of a marker at 15 degree steps for nodding and shaking.

Table II lists the overall performance of the proposed method compared to three state-of-the-art methods using the same database and evaluation metric. Gourier *et al.* [17] computed the head pose using auto-associative memories trained with the face images using the Widrow-Hoff correction rule. They classified the head pose by comparing input face image with those reconstructed using the auto-associative memory. The head pose of the highest similarity score was then selected. Tu *et al.* [19] used appearance variation caused by head pose changes to define a tensor model, where pose estimation was considered as a classification problem. The tensor model parameters were estimated during training by utilizing the classification performance. During testing, the tensor model was utilized to automatically localize the nose-tip position in the testing image and simultaneously estimate the head pose. They cropped face images manually in an experiment and automatically in another experiment according to the nose-tip location in both training and testing. Fenzi *et al.* [20] proposed a way to learn a class representation based on feature generative models derived from training class instances. Each model was based on a regression function learned from the descriptors of the same patch during training. During testing, the pose of the query image was estimated in a maximum *a posteriori* fashion by combining the regression functions that belong to the matching classes. Compared to the other

TABLE I
HEAD POSE ESTIMATION ERRORS FOR 12 MALE CAUCASIAN SUBJECTS FROM POINTING'04 DATABASE (DEGREES)

Test Subject	Nodding (θ_x)					Shaking (θ_y)				
	Generic	MA	MA (morphed)	MC	MC (morphed)	Generic	MA	MA (morphed)	MC	MC (morphed)
1	8.31	10.08	8.63	6.29	4.95	13.12	13.89	9.77	9.57	7.93
2	7.21	8.77	7.85	5.03	3.84	9.74	11.25	6.10	3.66	3.47
3	12.68	11.85	10.09	14.01	11.99	10.52	11.96	7.47	5.68	4.60
4	8.36	7.00	5.49	9.71	8.23	8.70	9.90	5.94	4.61	3.51
5	18.05	25.56	20.36	14.07	12.34	13.68	14.74	10.28	10.48	8.90
6	8.51	10.67	9.17	5.64	4.39	7.90	9.69	4.56	3.47	3.36
7	8.58	8.56	8.73	7.96	6.90	11.34	12.97	7.63	4.42	3.81
8	12.14	12.65	9.94	14.09	12.08	12.60	14.49	8.58	5.02	3.93
9	8.11	3.29	5.00	8.26	7.72	13.22	15.07	9.43	5.71	3.57
10	5.71	5.06	4.99	8.34	7.33	7.17	8.82	4.59	3.46	2.45
11	7.64	8.75	9.14	8.29	7.04	9.33	11.14	5.83	3.26	2.67
12	11.25	19.01	15.48	9.31	8.32	14.48	15.70	10.64	8.84	7.60
Average	9.71	10.94	9.57	9.25	7.93	10.98	12.47	7.56	5.68	4.65

TABLE II
COMPARISONS OF HEAD POSE ESTIMATION ERRORS OF DIFFERENT APPROACHES (DEGREES)

Authors	Methods	Nodding	Shaking
Gourrier (2007) [17]	Associative Memories	15.9	10.1
Tu (2009) [19]	Tensor Pose Models	11.37	12.28
	Tensor Pose Models (Manual)	4.37	5.01
Fenzi (2013) [20]	Class Generative Models	6.73	5.94
Proposed Method	Reference 3-D Model (Manual/Automatic)	9.25/9.43	5.68/5.41
	Morphed 3-D Reference Model (Manual/Automatic)	7.93/8.44	4.65/4.39

algorithms that use multiple training images of the query subject, the proposed method achieves similar performance (9.25°, 5.68°) using no training images of the query subject.

When one or more images of the query subject are available for training, the reference model can be morphed to improve the performance to (7.93°, 4.65°). Unlike the approaches that treat pose estimation as a classification problem with a finite number of pose classes, the proposed scheme provides a continuous value for the pose in all three directions from a single 2D face image. For each subject, the first 35 images were used to build the AAM and the remaining 35 images for testing. We selected seven facial features on the training face images used to build the AAM. The AAM fitted a shape model represented by the features to the query face image by iteratively adjusting the model parameters to minimize a distance measure between the 2D shape model and the face image. Once the fitting process is complete, the facial features positions are extracted from the model parameters. The average mean-absolute pose error with automatic feature extraction was (8.44, 4.39) degrees.

VI. CONCLUSION

This paper presents head pose estimation from a single 2D face image using a 3D face model morphed from a reference 3D face model of a person of the same ethnicity and gender as the query subject. Head pose angles are estimated by minimizing the disparity between the features on the query face image and the corresponding points on the 3D face model. When multiple images of the query subject are available for training, the reference model can be morphed for more accurate model fitting. The refined 3D face model becomes more specific to the query face image in terms of depth errors at the feature points. The proposed morphing process is computationally efficient since the 3D depth is adjusted by multiplying the depth at each feature point by a scalar depth parameter. Experiments with the USF Human-ID database suggest that morphed 3D face model decreases depth errors as well as feature disparity. Gender and ethnicity variations between test subject and 3D model affect the performance of head pose estimation. Experiments with the Pointing'04 database confirm that the proposed method successfully estimates head pose angles from a single 2D face test image with average errors in nodding and shaking angles of 9.25 and 5.68 degrees. In case a 2D face image of the query subject is available, the morphed 3D model estimates the head pose with the errors as low as 7.93 and 4.65 degrees.

APPENDIX

The variations of the features in each of three coordinates were found using the USF Human-ID database [16]. For each of M ($M = 100$) 3D face models in the database, a feature vector of N ($N = 7$) vertices can be represented by

$$\mathbf{v}_s = [v_{s1}, v_{s2}, \cdots, v_{sN}]^T, \quad s = x, y, z \tag{A1}$$

For the average feature vector $\bar{\mathbf{v}}_s$, the range of the feature points in each direction are given by

$$R_s = \|\max(\bar{\mathbf{v}}_s) - \min(\bar{\mathbf{v}}_s)\|, \quad s = x, y, z \tag{A2}$$

Then the average difference dX

$$dX = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left| v_{x\,j}^{i} - \bar{v}_{x\,j} \right| \tag{A3}$$

and similarly, dY and dZ in each of three coordinates represent the variation of the feature points. The percentage variations dX/R_x, dY/R_y, and dZ/R_z were 1.4%, 1.7%, and 3.5%, respectively. The feature variations in depth were greater than those of x- and y-coordinates.

REFERENCES

[1] H.-S. Koo and K.-M. Lam, "Recovering the 3D shape and poses of face images based on the similarity transform," *Pattern Recognit. Lett.*, vol. 29, no. 6, pp. 712–723, Apr. 2008.

[2] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 802–815, Feb. 2012.

[3] R. O. Mbouna, S. G. Kong, and M.-G. Chun, "Visual analysis of eye state and head pose for driver alertness monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1462–1468, Sep. 2013.

[4] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[5] S. Yan, H. Wang, J. Tu, X. Tang, and T. S. Huang, "Mode-kn factor analysis for image ensembles," *IEEE Trans. Image Process.*, vol. 18, no. 3, pp. 670–676, Mar. 2009.

[6] V. N. Balasubramanian, J. Ye, and S. Panchanathan, "Biased manifold embedding: A framework for person-independent head pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.

[7] R. Stiefelhagen, "Estimating head pose with neural networks—Results on the Pointing04 ICPR workshop evaluation data," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Aug. 2004, pp. 8–11.

[8] D. Huang, M. Storer, F. De la Torre, and H. Bischof, "Supervised local subspace learning for continuous head pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2921–2928.

[9] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int. J. Comput. Vis.*, vol. 9, no. 2, pp. 137–154, 1992.

[10] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2000, pp. 690–696.

[11] D. Jelinek and C. J. Taylor, "Reconstruction of linearly parameterized models from single images with a camera of unknown focal length," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 767–773, Jul. 2001.

[12] Z.-L. Sun, K.-M. Lam, and Q.-W. Gao, "Depth estimation of face images using the nonlinear least-squares model," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 17–30, Jan. 2013.

[13] P. Martins and J. Batista, "Single view head pose estimation," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1652–1655.

[14] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2D+3D active appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2004, pp. 535–542.

[15] J. Gonzalez-Mora, F. De la Torre, N. Guil, and E. L. Zapata, "Learning a generic 3D face model from 2D image databases using incremental structure-from-motion," *Image Vis. Comput.*, vol. 28, no. 7, pp. 1117–1129, Jul. 2010.

[16] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[17] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, "Head pose estimation on low resolution images," in *Multimodal Technologies for Perception of Humans* (Lecture Notes in Computer Science), vol. 4122. Berlin, Germany: Springer-Verlag, 2007, pp. 270–280.

[18] J. M. Blain, *The Complete Guide to Blender Graphics: Computer Modeling and Animation*. Boca Raton, FL, USA: CRC Press, 2012.

[19] J. Tu, Y. Fu, and T. S. Huang, "Locating nose-tips and estimating head poses in images by tensorposes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 1, pp. 90–102, Jan. 2009.

[20] M. Fenzi, L. Leal-Taixe, B. Rosenhahn, and J. Ostermann, "Class generative models based on feature regression for pose estimation of object categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 755–762.

[21] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, 2004.

[22] J. Heo and M. Savvides, "Gender and ethnicity specific generic elastic models from a single 2D image for novel 2D pose face synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2341–2350, Dec. 2012.

[23] G. Guo and G. Mu, "A study of large-scale ethnicity estimation with gender and age variations," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 79–86.

[24] H. Han and A. K. Jain, "Age, gender and race estimation from unconstrained face images," Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep. (MSU-CSE-14-5), 2014.

Seong G. Kong received the B.S. and M.S. degrees from Seoul National University, Seoul, Korea, in 1982 and 1987, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, CA, USA, in 1991, all in electrical engineering. He was an Associate Professor with the Department of Electrical and Computer Engineering, University of Tennessee, Knoxville, and Temple University. He is currently a Professor of Computer Engineering with Sejong University, Seoul. His research interests include image processing, pattern recognition, and intelligent systems. He was a recipient of the best paper award from the International Conference on Pattern Recognition in 2004, the Honorable Mention Paper Award from the American Society of Agricultural and Biological Engineers, the Professional Development Award from the University of Tennessee in 2005, and the Most Cited Paper Award from the COMPUTER VISION AND IMAGE UNDERSTANDING journal in 2007 and 2008. His professional services include as an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS, a Guest Editor of a special issue of the INTERNATIONAL JOURNAL OF CONTROL, AUTOMATION, AND SYSTEMS, a Guest Editor of a special issue of the JOURNAL OF SENSORS, and a Program Committee Member of various international conferences.

Ralph Oyini Mbouna received the B.S. and M.S. degrees in electrical and computer engineering from Temple University, Philadelphia, PA, USA, in 2012, and the Ph.D. degree in electrical engineering from Temple University, in 2014. He is currently an Assistant Professor of Electrical and Computer Engineering with Temple University. His research interests are pattern recognition and computer vision, including gaze estimation and tracking, driver alertness monitoring, 3D face reconstruction, and biometric identification.