MVA2013 IAPR International Conference on Machine Vision Applications, May 20-23, 2013, Kyoto, JAPAN

Pose-Invariant Face Recognition Using A Single 3D Reference Model

Gee-Sern Hsu*, Hsiao-Chia Peng
National Taiwan University of Science and Technology
No. 43, Sec.4, Keelung Rd., Taipei, 106, Taiwan
*jison@mail.ntust.edu.tw

Abstract

Approaches for cross-pose face recognition can be categorized into 2D image based and 3D model based. However, only a small number of 3D model based approaches are reported although 3D information is considered crucial for cross-pose analysis. Extended from a latest face reconstruction method using a single 3D reference model, this study focuses on using the reconstructed 3D face for recognition across poses. Given a 2D face image of frontal pose, one can reconstruct its 3D model from a 3D reference face using spherical harmonics to approximate the reflectance across the given face. Experiments on the PIE database show that addition of local correspondences for pose alignment, multiple reference models considered in the reconstruction phase, and the addition of a profile pose to the gallery set can make the performance competitive to the state-of-the-art.

1 Introduction

The approaches for face recognition across poses can be generally split into two categories, one is 2D image based [1, 2, 3, 4] and the other is 3D model based [5, 6, 7, 8]. An extensive review on both categories can be found in [9]. Although the 2D based approaches can be derived from 3D analysis, they only require multiple 2D images in the training phase, instead of the 3D facial models considered in the 3D model based methods. Because more advancement has been made on 2D based approaches than that on 3D based ones, the former appears to outnumber the latter significantly in the literature [9], and only a small number of research on the latter is available. Therefore, more 3D model based methods are yet to be developed as 3D facial information is considered crucial for cross-pose recognition.

In the 2D based methods, the Eigen Light-Feilds (ELF) [1] assumes that the pixel intensity corresponds to the radiance of light emitted from the face along certain rays in space, is defined on the set of all such radiance values over all possible rays. Tied Factor Analysis (TFA) [2] decomposes a face into a latent variable (or factor) in the identify space, a pose-dependent mapping from identity to observation, a pose-dependent mean and a noise. Given a non-frontal face with a known pose, its corresponding frontal pose can be estimated using the learned frontal pose mapping and mean, and then matched against those in the gallery. This method requires manual annotation of local features for pose alignment, and similar to ELF, it only works for the poses available in the training phase. A stereo matching approach with epipolar geometry [3] evaluates the similarity between two faces with different poses. Given a few matched points on both faces, the dense correspondences across the faces can be computed using an optimized stereo matching approach. The performance degrades when misalignment. A regression-based approach [4] estimates the coefficients of linear combinations of the 2D face images in the training set for approximating the face in 3D. Similar to most 2D-based methods, this approach also suffers from the limitation that it only works for poses available in the training set.

In 3D model based approaches, the morphable model [5] uses the prior knowledge, including the 3D face shapes and textures, collected from hundreds of 3D facial scans to build a 3D model for a given 2D image. Although considered as an effective solution for cross-pose recognition, it is expensive in storage and computation. A similar approach but modified with automatic feature localization is given in [6]. It is reported a satisfactory performance for poses less than 45° but degrades significantly for large poses. We consider this a baseline for 3D methods in our performance evaluation. The Generic Elastic Model (GEM) [7] reconstructs the 3D face from a single 2D face which has been annotated by as many as 79 fiducial points which influences the reconstruction accuracy significantly. Another latest work on cross-pose recognition, the Heterogeneous Specular and Diffuse (HSD) [8] allows both specular and diffuse reflectance coefficients to vary spatially to better accommodate the surface properties of faces. A few face images under different lighting conditions are needed to estimate the 3D shape and surface reflectivity using stochastic optimization. The resultant personalized 3D face model is used to render novel gallery views under different poses for recognition across pose.

Our method extends the work in [10], one of the latest research on 3D face reconstruction, to cross-pose face recognition. It is 3D model based in nature, but different from [5, 6, 7, 8] and others in that it exploits a single 3D reference model and recovers the 3D shape of a 2D face image in the gallery without the need of a dense set of correspondence points. This method consists of two phases: **I**. the 3D reconstruction using the reference model and spherical function approximation, as presented in Sec. 2, and **II**. the model-based synthesis of novel views and pose-oriented feature extraction and matching, as described in Sec. 3. It is observed in our experimental study, presented in Sec. 4, that additional correspondence points, an additional 3D reference model considered in the reconstruction phase, and the addition of profile pose to the gallery set can substantially improve the performance, followed by a conclusion given in Sec. 5.

2 3D Reconstruction from a 2D Face

We reformulate the problem as a constrained minimization so that the well-known scheme with Lagrange multipliers can be applied. We also make some minor modifications to the original algorithm, making our reconstruction somewhat different from that in [10], although the overall workflow and results are similar. Nevertheless, the investigations that we have added to the reconstruction phase include the study on different numbers of fiducial points used for the alignment between the 2D image and 3D reference model, and the model parameter estimation when considering a 3D face scan from a different database as the reference model.

Assuming that the face surface is Lambertian, a 2D face image $I(x,y)$ can be written as

$$I(x,y) = \rho(x,y)\vec{h}(x,y) \cdot \vec{n}(x,y) = \rho(x,y)R(x,y) \quad (1)$$

where $\rho(x,y)$ is the surface albedo at the point (x,y), $\vec{h}(x,y) \in R^3$ is the lighting cast on (x,y) with intensity on each of the three directions, $\vec{n}(x,y)$ is the face surface normal at (x,y), and the reflectance $R(x,y) = \vec{h}(x,y) \cdot \vec{n}(x,y)$. For simplicity of notation, the coordinates (x,y) is dropped in the rest of the paper, and $\vec{n}(x,y)$, for example, is written as $\vec{n}$. With Lambertian surface and a few assumptions [10], the reflectance can be approximated using spherical harmonics,

$$R(x,y) \approx \vec{l} \cdot \vec{Y}(\vec{n}) \quad (2)$$

where $\vec{l}$ is the lighting coefficient vector and $\vec{Y}(\vec{n})$ is the spherical harmonic vector, which, in the second order approximation, takes the following form:

$$\begin{aligned}\vec{Y}(\vec{n}) \quad = \quad & [c_0, c_1 n_x, c_1 n_y, c_1 n_z, c_2 n_x n_y, c_2 n_x n_z, c_2 n_y n_z, \\ & c_2(n_x^2 - n_y^2)/2, c_2(3n_z^2 - 1)/2\sqrt{3}\Big]^T \quad (3)\end{aligned}$$

where $c_0 = 1/\sqrt{4\pi}$, $c_1 = \sqrt{3}/\sqrt{4\pi}$, $c_2 = 3\sqrt{5}/\sqrt{12\pi}$.

The difference between (1) and (3) is that the lighting intensity and direction are all merged into $\vec{h}$ in (1), separated from $\vec{n}$, but in (3) they are split into the lighting vector $\vec{l}$ and the spherical harmonics $\vec{Y}(\vec{n})$, which is solely dependent on the components of $\vec{n}$, namely n_x, n_y and n_z.

The core problem can now be formulated as the minimization of $||I - \rho\vec{l} \cdot \vec{Y}(\vec{n})||$ over ρ, $\vec{l}$ and $\vec{n}$. The solution in [10] uses a reference model M_r, which offers the depth $z_r(x,y)$, the surface normal $\vec{n}_r(x,y)$ and the albedo $\rho_r(x,y)$ as reference for initialization, making the problem solvable by regularization. Because of a better computational efficiency, we choose DoG (Difference of Gaussian) instead of LoG (Laplacian of Gaussian) adopted in [10] in the minimization.

$$\min_{\vec{l},\vec{z},\rho} \int (I - \rho\vec{l}\cdot\vec{Y}(\vec{n}))^2 + \lambda_1(D_g * d_z)^2 + \lambda_2(D_g * d_\rho)^2 dxdy \quad (4)$$

where $d_z = z(x,y) - z_r(x,y)$, $d_\rho = \rho(x,y) - \rho_r(x,y)$, and D_g* denotes the convolution with the DoG; λ_1 and λ_2 are constants. Although this is not described explicitly in [10], the formulation in (4) can be better interpreted as the minimization of $||I - \rho\vec{l} \cdot \vec{Y}(\vec{n})||$ subject to the constraints $D_g * d_z \approx 0$ and $D_g * d_\rho \approx 0$. Such a formulation allows the interpretation of λ_1 and λ_2 as the Lagrange multipliers. Assuming that I is aligned to the reference model, the reconstruction tackles the minimization in (4) by first solving for the spherical harmonic coefficients $\vec{l}$ using the references z_r and ρ_r , then the depth $z(x,y)$, and then the albedo $\rho(x,y)$.

The alignment between I and the reference model needs corresponding fiducial points on both I and the reference model. We applied the method in [11] for automatic detection of facial features, and adjusted the results manually in case the method failed to perform ideally. Given a set of fiducial points that split I and the reference face into corresponding local regions, perspective and affine transforms are applied to fit each local region of the reference model to the corresponding region in I. Our experiments show that it is not always true that more corresponding points lead to a better performance.

Instead of using the samples from the USF database as the reference models as in [10], we select the 3D images from the FRGC database [12] for its popularity. Each FRGC 3D image consists of a range image and a texture image as shown in Fig. 1(a), on which one needs to estimate $\vec{n}_r(x,y)$ and $\rho_r(x,y)$. This step is excluded in [10], but considered an essential part of the method when one is considering a different reference model. Rather than computing the surface mesh to obtain the surface normal vector $\vec{n}_{ref}$ from each mesh, we apply the approximation on the point cloud dataset of each point p_i directly by Total Least Squares (TLS) proposed in [13].

$$C = \frac{1}{k}\sum_{i=1}^{k}(p_i - \overline{p}) \cdot (p_i - \overline{p})^T \quad (5)$$

$$C \cdot \vec{v_j} = \lambda_j \cdot \vec{v_j}, \; j \in \{0, 1, 2\} \quad (6)$$

TLS decomposes the covariance matrix C of the k-nearest neighboring points of p_i and estimates $\vec{n}_{ref}(x,y)$ using the regional eigen-structure. $\overline{p}$ represents the 3D centroid of the nearest neighbors. The j-th eigenvalue and eigenvector of the covariance matrix is represented in λ_j and $\vec{v_j}$.

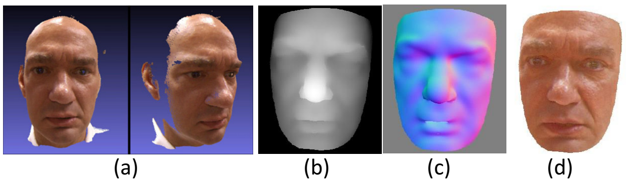

Figure 1: (a) 3D image from FRGC (b) depth map (c) normal map (d) albedo after model parameter estimation.

3 Face Recognition Across Poses

We assume a common scenario that the gallery has one frontal face image per subject for enrollment, and the probe set contains face images of other poses for recognition. A couple issues must be solved for this scenario: the preparation of the images good for training from the reconstructed 3D face, and the estimate of the pose of a given probe so that its matching to the gallery can be fast. These are discussed below. To refrain the scope of this paper from covering facial feature localization, which can be solved by many algorithms, e.g., [11], we assume that the fiducial points on a probe can be available using these algorithms or by manual annotation.

3.1 Reconstructed Model Based Training Images

Each face image in the galley set is taken as the $I(x, y)$ in (4) for making its corresponding 3D face from the reference model. The alignment between $I(x, y)$ and the reference model is performed using a number of fiducial points manually selected. It is observed in our experiments that the precision of the alignment makes a strong impact on the recognition performance. This is discussed along with experimental results in Sec. 4.

Following the approach presented in Sec. 2, one can obtain a 3D reconstructed face for each gallery image. However, one cannot directly use this 3D face to generate 2D images of other poses good for recognition because the discontinuities among the depth points of the reconstructed face induce many null spots when projecting onto the 2D image planes of other poses. To fill in the null spots, we compute the triangle mesh on each depth point with its two nearest neighbors. The gallery image is then used as the texture on the meshed surface and projected onto the image plane of a desired pose. The work flow of the recognition is illustrated in Fig. 2.

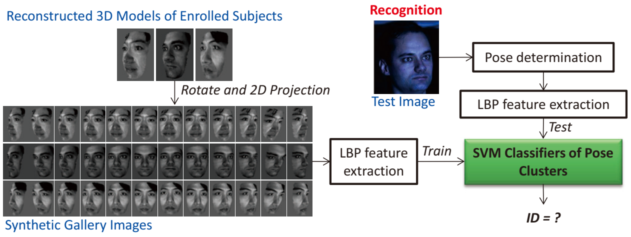

Figure 2: Work flow of face recognition across poses base on single reference model.

3.2 Pose-Oriented Recognition

Although one can generate training images of arbitrary poses using the above approach, we consider it a better option to generate pose-oriented clusters of training images. Take the pose subset with 13 variations in the CMU PIE database [14] as an example, which is used in our experiments for performance evaluation. When generating the training set, each of these poses is considered as the center of a pose-oriented cluster, and four neighboring poses are synthesized and added to the cluster, including ±10° in yaw and pitch angles. Instead of using the PIE original pose tags, such as c02, c37, ..., we use the approximated pose angle with an alphabet in the front to denote its direction. For example, R67.5° refers to 67.5° to the right, U22.5° is 22.5° upward and D22.5° is 22.5° downward. All synthesized face images are normalized in size to either the distance between the eyes and mouth when the poses are primarily caused by horizontal rotations, or to the distance between both eyes when the poses are caused by vertical rotations.

Given a probe with fiducial points available[1] for size normalization and pose matching to the pose cluster in the gallery set, the size of the pose is first normalized so that the distance between the eyes and mouth is the same as those used in the images. The texture feature LBP is extracted. We applied the uniform pattern (ULBP)[2] with the feature dimension reduced to 59. The ULBP is extracted for each pixel, and the associated histogram is obtained as the feature vector for a block of 32×32 in size. Each 128×128 input image is split into 4×4 blocks, and the feature vector of each 32×32 block is cascaded into a $4 \times 4 \times 59 = 944$ dimensional feature vector. Following the above procedure, each person in the gallery can have 13 pose clusters, and each cluster has five synthesized images with ULBP feature extracted.

4 Experiments

All experiments were carried out on a Linux platform with Intel Core i3-2120 processor with 3.30 GHz. We used OpenCV (`http://opencv.org`) for image processing, CLAPACK (`http://www.netlib.org/clapack`) for solving optimization and Freeglut (`http://freeglut.sourceforge.net`) for handling 3D models of different poses. The Point Cloud Library (`http://pointclouds.org`) was used for preprocessing on both the reference and reconstructed models. All reference models were taken from the FRGC database [12] and resized into 250×300. The recognition was evaluated on the PIE pose subset, with frontal pose as the gallery and the remaining poses as the probe. The processing time for reconstruction was found to increase exponentially with the scale factor imposed on the gallery image considered for reconstruction. 0.3x was selected in all experiments for better efficiency with 133 seconds processing time.

Three issues were studied: additional fiducial points for local correspondences and pose alignment, an additional 3D reference model considered in the reconstruction phase, and the addition of profile pose to the gallery set.

Additional Fiducial Points for Pose Alignment

The more fiducial points available for the local correspondences between the gallery image and reference model, the more accurate the pose alignment can be. We compared a case with 3 and 12 fiducial points as shown in Fig. 3(b). The latter case with 12 points split the face into 19 local regions as shown in Fig. 3(a). Perspective transform and affine transform were applied on these regions to fit each gallery image to the reference model.

Additional Reference Model

The default reference model was arbitrarily selected, as the one shown in the previous figures. We selected a different gender and age as the additional one. Following the same approach, each galley image had two

[1] To better confine the scope of this paper, the detection of fiducial points on a given face is considered solved by existing methods, for example [11].

[2] Compared with other forms of LBP, the ULBP shows the most consistent result in our experiments.

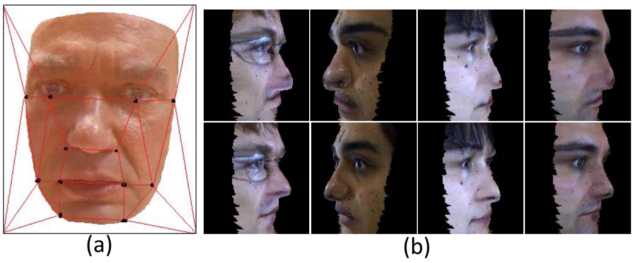

Figure 3: (a) 12 fiducial points. (b) Comparison of a case with 3 (top row) and 12 (bottom row) fiducial points.

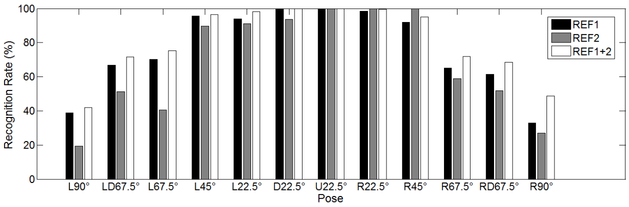

Figure 4: Performance comparison of single and double reference models.

reconstructed models, generating an additional set of pose clusters for training. The comparison between the two cases is shown in Fig. 4, where the case with both REF1 and REF2 models outperforms the case with either one alone.

Addition of Profile Pose in the Gallery

A common scenario in forensic and law enforcement applications considers both frontal and profile poses available in the gallery. This scenario was considered in our experiments, which compared the performance of the proposed methods with different settings to several approaches reported in Zhang's review [9] on the PIE pose subset. Two reference models with 12 and 9 fiducial points were considered in our settings. Both the case with frontal pose only and the case with frontal and 90° profile poses in the gallery were tested. The results are shown in Fig. 5. The best three are the stereo matching [3], the HSD [8] and the proposed with both frontal and profile poses in the gallery. In addition, the proposed method with frontal pose only in the gallery performs similarly well as the best ones do when recognizing faces with poses less than 67.5°.

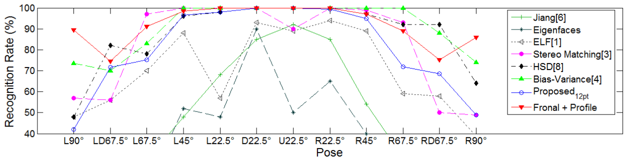

Figure 5: Recognition comparison on PIE database with ambient light on. Recognition rates less than 40% are ignored.

5 Conclusion

3D models and features are mostly considered important for pose invariant face recognition; however, only a small number of research has been carried out in this regard. This study focuses on using a reconstructed 3D face for recognition across pose. Given a 2D face image, one can reconstruct its 3D surface from the reference model using spherical harmonic functions to approximate its irradiance. The reconstructed 3D face allows the synthesis of novel views with arbitrary poses. Experiments on the PIE database show that the method can be competitive to the state-of-the-art with multiple reference models considered in the reconstruction phase, additional fiducial points for pose alignment, and the addition of profile pose to the gallery.

References

[1] Ralph Gross, Iain Matthews, and Simon Baker: "Appearance-based face recognition and light fields," *TPAMI*, vol. 26, pp. 449–465, 2004.

[2] Simon J.D. Prince, James H. Elder, Jonathan Warrell, and Fatima M. Felisberti': "Tied factor analysis for face recognition across large pose differences," *TPAMI*, vol. 30, pp. 970–984, June 2008.

[3] Carlos D. Castillo and David W. Jacobs: "Using stereo matching for 2-D face recognition across pose," *CVPR*, 2007, pp. 1–8.

[4] Annan Li, Shiguang Shan, and Wen Gao: "Coupled bias-variance tradeoff for cross-pose face recognition," *TIP*, vol. 21, no. 1, pp. 305–315, 2012.

[5] Volker Blanz and Thomas Vetter: "Face recognition based on fitting a 3D morphable model," *TPAMI*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[6] Dalong Jiang, Yuxiao Hu, Shuicheng Yan, Lei Zhang, Hongjiang Zhang, and Wen Gao: "Efficient 3D reconstruction for face recognition," *PR*, vol. 38, pp. 787–798, June 2005.

[7] Utsav Prabhu, Jingu Heo, and Marios Savvides: "Unconstrained pose-invariant face recognition using 3d generic elastic models," *TPAMI*, vol. 33, pp. 1952–1961, 2011.

[8] Xiaozheng Zhang and Yongsheng Gao: "Heterogeneous specular and diffuse 3-D surface approximation for face recognition across pose," *IEEE Trans. Inf. Forensics and Security*, vol. 7, no. 2, pp. 1952–1961, 2012.

[9] Xiaozheng Zhang and Yongsheng Gao: "Face recognition across pose: A review," *PR*, vol. 42, pp. 2876–2896, Nov. 2009.

[10] Ira Kemelmacher-Shlizerman and Ronen Basri: "3D face reconstruction from a single image using a single reference face shape," *TPAMI*, vol. 33, no. 2, pp. 394–405, Feb. 2011.

[11] Liya Ding and Aleix M. Martínez: "Features versus context: An approach for precise and detailed detection and delineation of faces and facial features," *TPAMI*, vol. 32, no. 11, pp. 2022–2038, 2010.

[12] P.Jonathon Phillips, Patrick J.Flynn, Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek: "Overview of the face recognition grand challenge," *CVPR*, 2005, vol. 1, pp. 947–954.

[13] Niloy J. Mitra, An Nguyen, and Leonidas J. Guibas: "Estimating surface normals in noisy point cloud data," *IJCGA*, vol. 14, no. 4-5, pp. 261–276, 2004.

[14] Terence Sim, Simon Baker, and Maan Bsat: "The CMU pose, illumination, and expression (PIE) database," *FG*, 2002, pp. 46–51.