IEICE TRANS. INF. & SYST., VOL.E98–D, NO.6 JUNE 2015

PAPER

Face Recognition Across Poses Using a Single 3D Reference Model

Gee-Sern HSU[†a)], ***Member***, **Hsiao-Chia PENG**[†], **Ding-Yu LIN**[†], ***and*** **Chyi-Yeu LIN**[†], ***Nonmembers***

SUMMARY Face recognition across pose is generally tackled by either 2D based or 3D based approaches. The 2D-based often require a training set from which the cross-pose multi-view relationship can be learned and applied for recognition. The 3D based are mostly composed of 3D surface reconstruction of each gallery face, synthesis of 2D images of novel views using the reconstructed model, and match of the synthesized images to the probes. The depth information provides crucial information for arbitrary poses but more methods are yet to be developed. Extended from a latest face reconstruction method using a single 3D reference model and a frontal registered face, this study focuses on using the reconstructed 3D face for recognition. The recognition performance varies with poses, the closer to the front, the better. Several ways to improve the performance are attempted, including different numbers of fiducial points for alignment, multiple reference models considered in the reconstruction phase, and both frontal and profile poses available in the gallery. These attempts make this approach competitive to the state-of-the-art methods.

key words: *face recognition, face reconstruction, sparse representation classification*

1. Introduction

Approaches for cross-pose recognition can be split into two categories, one is 2D image based [1]–[4] and the other is 3D model based [5]–[8]. More advances have been made on the former which appear to outnumber the latter considerably [9]. However, most 2D approaches can only work for poses available in the training set. Because 3D facial information is considered crucial for recognition across arbitrary poses [9], more 3D based methods are yet to be developed. Here we focus on the 3D based methods which match 2D probe images with synthesized 2D images from 3D models because many well proved image features for face recogniton can be applied in the scenario.

In 2D image based methods, the Eigen Light-Fields (ELF) [1] assumes that the pixel intensities correspond to the radiance of light emitted from the face along certain rays in space, and estimates the basis set of the radiance values at each pose using samples of the same pose in the training set. The eigen light-field is defined on this basis set, which allows the gallery and probe faces represented in ELF coefficients, and recognition can be performed by matching these coefficients. It is an effective method dealing with poses, but suffers from the requirements that the probe images must align with light-field vectors. In the Tied Factor Analysis (TFA) [2], a face is decomposed into a latent variable (or factor) in the identity space, a pose-dependent mapping from identity to observation, a pose-dependent mean and a noise. Since the pose-dependent mapping and mean are independent of the subject, they can be obtained from a training set. 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-specific alignment. The performance degrades, sometimes significantly, when local features fail to be accurately localized. A stereo matching approach with epipolar geometry is applied in [3] to evaluate the similarity between two faces of different poses. Given three or four corresponding feature points on both faces, two sets of scanlines with epipolar constraints can be determined, and a stereo matching cost can be computed and optimized to reveal how well the two faces match each other. The regression-based method in [4] estimates the coefficients of linear combinations of 2D faces in the training set for approximating the linear subspaces for 3D face. To reduce the high variances in the estimated coefficients, the method exploits the regressors with local Gabor features for bias-variance balancing. Although these 2D-based methods report performances better than many 3D-based ones, all of them and many other 2D methods suffer from the limitation that they only work for poses available in the training set, making them ineffective in some practical applications.

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 recognition under pose and illumination variations, it is expensive in storage and computation because of the storage of the hundreds of 3D scans and the search for the correspondences to the reference model. A similar approach but modified with automatic feature localization is given in [6], which reports a satisfactory performance for poses less than 45°, but degrades significantly for large poses. Because the conventional PCA is used after synthesizing the views to match against the probe, we consider this a baseline for 3D methods in our performance evaluation. The Generic Elastic Model (GEM) [7] reconstructs the 3D face from a 2D face annotated with 79 fiducial points and a generic facial depth map. The reconstruction accuracy strongly depends on the correspondences between the fiducial points on the reference faces and those

Manuscript received October 21, 2014.
Manuscript publicized February 24, 2015.
[†]The authors are with National Taiwan University of Science and Technology, Taiwan R.O.C.
a) E-mail: jison@mail.ntust.edu.tw
DOI: 10.1587/transinf.2014EDP7352

on the gallery face, which can be difficult to define precisely. The Heterogeneous Specular and Diffuse (HSD) [8], one of the latest approaches, allows both specular and diffuse reflectance coefficients to vary spatially to better accommodate surface properties of real faces. A few face images under different lighting conditions are needed to estimate the 3-D shape and surface reflectivity using stochastic optimization. The resultant personalized 3-D face model is used to render novel gallery views of different poses for cross-pose recognition. A similar scheme to the proposed one [10] applied procrustes analysis which aligns synthetic image and the query image after 3D model is built. The block based MLBP feature is extracted for matching.

Our method extends the latest work on 3D face reconstruction proposed by Kemelmacher-Shlizerman and Basri [11] to tackle cross-pose recognition. It is 3D model based in nature, but different from [5]–[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 correspondences. It consists of the following steps: (1) 3D reconstruction based on the reference model, (2) model-based synthesis of novel poses, and (3) pose-oriented feature extraction and matching. The proposed method is a low-cost alternative to many 3D model-based approaches that require a large number of 3D scans. We evalutate the impacts made by multiple reference models, multiple sets of fiducial points for alignment, various settings on the features and algorithms, along with an extensive experimental study.

The rest of the paper is organized as follows: The preparation of the 3D reference model and the model-based reconstruction are presented in Sect. 2. Although the reconstruction part is mostly based on [11], our interpretation from a different viewpoint can be easier for implementation. The model-based synthesis of novel views and the pose-clustered recognition using the Sparse Representation-based Classification (SRC) are elaborated in Sect. 3. An extensive experimental study on the performance of SRC with different settings is presented in Sect. 4, along with a comparison with state-of-the-art approaches. A conclusion of this study is then given in Sect. 5.

2. Reconstruction Using Single Reference Model

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 in [11], making our reconstruction different from theirs, although the results are similar. Nevertheless, the investigations that we have added to the reconstruction phase include the rendering of a smooth surface given the noisy data of a 3D face scan for the reference model, model parameter estimation and the study on different numbers of fiducial points used for the alignment between the 2D image and 3D reference model. The former is presented in Sect. 2.1, and the latter in Sect. 4 with experimental results.

2.1 Reference Model Surface Rendering and Parameter Estimation

This step is not described explicitly in [11], but considered an essential part of the reconstruction when making a raw 3D face scan good as the reference model. Instead of using samples from the USF database as the reference models as in [11], we select samples from the FRGC database [12] because of its popularity. Each FRGC 3D face scan consists of a range image and a texture image that we can use to estimate the surface normal $\vec{n}_r(x, y)$ and albedo $\rho_r(x, y)$, which are required for the reconstruction of other faces.

We applied the Moving Least Squares (MLS) [13] to smooth $z_{r,0}$, the raw depth data of the reference model, so that the measurement noise in $z_{r,0}$ can be removed and the smoothed surface z_r can best approximate $z_{r,0}$. Given a subset of $z_{r,0}$ in the form of point clouds with N_k points in the subset, denoted as $\mathbf{P}_k = \{\vec{p}_i\}_{i=1,\cdots,N_k}$, the goal is to determine a novel set of points, $\mathbf{R}_k = \{\vec{r}_i\}_{i=1,\cdots,N_k}$, on a low-order polynomial that minimizes the distance between $\mathbf{P}_k$ and $\mathbf{R}_k$. The smoothed surface z_r can then be obtained from $\{\mathbf{R}_k\}_{\forall k}$. Modified from the MLS reported in [13] for better efficiency, our method is composed of the following step,

1. Use $\mathbf{P}_k$ to determine a local plane H_0 with origin $\vec{q}_0$ and normal $\vec{n}_0$ so that the following weighted sum can be computed,

$$\sum_{i=1}^{N_k} (u_0(x_i, y_i) - \mu_{i,0})^2 \, \phi\left(\left\|\vec{p}_i - \vec{q}_0\right\|\right) \tag{1}$$

   where $u_0(x_i, y_i)$ is the distance of $\vec{r}_i$ to H_0 with the location of its projection onto H_0 given by (x_i, y_i); $\mu_{i,0}$ is the distance of $\vec{p}_i$ to H_0, i.e., $\mu_{i,0} = \vec{n}_0 \cdot (\vec{p}_i - \vec{q}_0)$; and $\phi(\cdot)$ is a Gaussian function so that the points closer to $\vec{q}_0$ are weighted more. Assuming that $\mathbf{R}_k$ are described by a low-order polynomial in terms of the coordinates (x_i, y_i) on H_0, i.e., $\vec{r}_i = f(x_i, y_i|\Lambda_0)$ and $u_0(x_i, y_i) = \vec{n}_0 \cdot (f(x_i, y_i|\Lambda_0) - \vec{q}_0)$, where $f(x_i, y_i|\Lambda_0)$ is a polynomial surface with parameter Λ_0 that defines the local geometry of $\mathbf{R}_k$.
2. Because H_0 can be uniquely defined given $\vec{q}_0$ and $\vec{n}_0$, one can change them to $\vec{q}_1$ and $\vec{n}_1$ and obtain a novel plane H_1. Given that the order of the polynomial $f(x_i, y_i|\Lambda_0)$ is fixed (so that the number of parameters of $f(x_i, y_i|\Lambda_0)$ is fixed), a parameter estimation problem can be defined as the minimization of the weighted sum as:

$$\Lambda_k^*, \vec{n}_k^*, \vec{q}_k^* = \underset{\Lambda,\vec{n},\vec{q}}{\operatorname{argmin}} \sum_{i=1}^{N_k} (u(x_i, y_i) - \mu_i)^2 \, \phi\left(\left\|\vec{p}_i - \vec{q}\right\|\right) \tag{2}$$

The above can be repeated on other subsets $\{\mathbf{P}_k\}_{\forall k}$ for estimating $\{\Lambda_k, \vec{n}_k, \vec{q}_k\}_{\forall k}$ and $\{\mathbf{R}_k\}_{\forall k}$. A key issue in this scheme is the initial estimates of $\vec{n}_0$ and $\vec{q}_0$. A few possible ways are given in [13]; however, from our experiments we found that

the minimum principal component extracted from $\mathbf{P}_k$ offers a good estimate of $\vec{n}_0$ and the centroid of $\mathbf{P}_k$ can be appropriate as $\vec{q}_0$. To extract the principal components, one needs to solve the eigenvectors of the covariance C_k,

$$C_k = \frac{1}{k}\sum_{i=1}^{k}(\vec{p}_i - \vec{q}_0)\cdot(\vec{p}_i - \vec{q}_0)^T \tag{3}$$

where $\vec{q}_0$ is the centroid of $\mathbf{P}_k$, and considered as the origin of the initial plane H_0. $\vec{n}_0$, the normal vector of H_0, is given by the eigenvector of C_k associated with the lowest eigenvalue. Following the above approach, the surface normal $\vec{n}_r$ can be obtained from the estimated polynomials $f(x_i, y_i|\Lambda_k)$. Given $\vec{n}_r$ and the associated 2D image I_r, ρ_r can be estimated using the method presented in the next section with some simplification, as described at the end of Sect. 2.2.

2.2 Irradiance Evaluation Using Constrained Minimization

The goal in this section is to estimate the 3D shape model of any given 2D face image $I(x, y)$ using the depth $z_r(x, y)$, surface normal $\vec{n}_r(x, y)$ and albedo $\rho_r(x, y)$ of the reference model. Assuming that the face surface is Lambertian, $I(x, y)$ can be decomposed as

$$I(x, y) = \rho(x, y)\vec{h}(x, y)\cdot\vec{n}(x, y) = \rho(x, y)R(x, y) \tag{4}$$

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 a few assumptions [11], the reflectance can be approximated using spherical harmonics,

$$R(x, y) \approx \vec{l}\cdot\vec{Y}(\vec{n}) \tag{5}$$

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:

$$\vec{Y}(\vec{n}) = \Big[c_0, c_1 n_x, c_1 n_y, c_1 n_z, c_2 n_x n_y, c_2 n_x n_z, c_2 n_y n_z, c_2(n_x^2 - n_y^2)/2, c_2(3n_z^2 - 1)/2\sqrt{3}\Big]^T \tag{6}$$

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 (4) and (6) is that the lighting intensity and direction are all merged into $\vec{h}$ in (4), separated from $\vec{n}$, but in (6) 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 [11] uses the depth z_r, the surface normal $\vec{n}_r$ and the albedo ρ_r of the reference model for initialization, making the problem solvable by regularization. We choose DoG (Difference of Gaussian) instead of LoG (Laplacian of Gaussian) adopted in [11] in the minimization because of a better computational efficiency without loss of accuracy:

$$\min_{\vec{l},z,\rho}\int (I - \rho\vec{l}\cdot\vec{Y}(\vec{n}))^2 + \lambda_1(D_g * d_z)^2 + \lambda_2(D_g * d_\rho)^2 dxdy \tag{7}$$

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 [11], the formulation in (7) 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 (7) 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 [14] 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 splits I and the reference face into corresponding local regions, perspective and affine transforms are then applied to fit each local region of the reference model to the corresponding region in I.

The minimization (7) is also used for computing ρ_r given I_r and $\vec{n}_r$. In such a case, there are no constraint terms in (7), and one can use the average of 2D faces in the gallery as the initial guess of the albedo, $\rho_r^{(0)}$, to solve the lighting coefficients $\vec{l}^{(0)}$ and search for the desired ρ_r iteratively.

3. Cross-Pose Recognition Using SRC

3.1 Generation of Model-Based Training Images

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 generation of 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. To constrain the scope of this paper from covering facial feature localization, which can be solved by many algorithms, e.g., [14], we assume that the fiducial points on a probe can be available using these algorithms or manual annotation.

The overall recognition workflow is given in Fig. 1. Each frontal face image in the gallery is taken as the $I(x, y)$ in (7) for making its corresponding 3D face. The alignment between $I(x, y)$ and the reference model is performed using a set of fiducial points. Our experiments reveal that the fiducial-points-based alignment makes a strong impact on the reconstruction and recognition performance. Figures 2 (a) and (b) show the reconstruction using 3 and 12 fiducial points. This, however, does not imply that more fiducial points always lead to better reconstruction. This issue is discussed along with experimental results in Sect. 4.

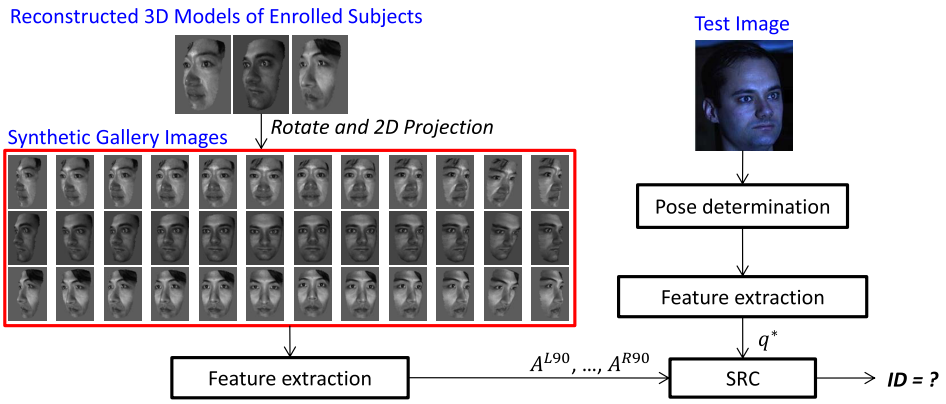

Fig. 1 Workflow of recognition across pose given the reconstructed 3D face in the gallery.

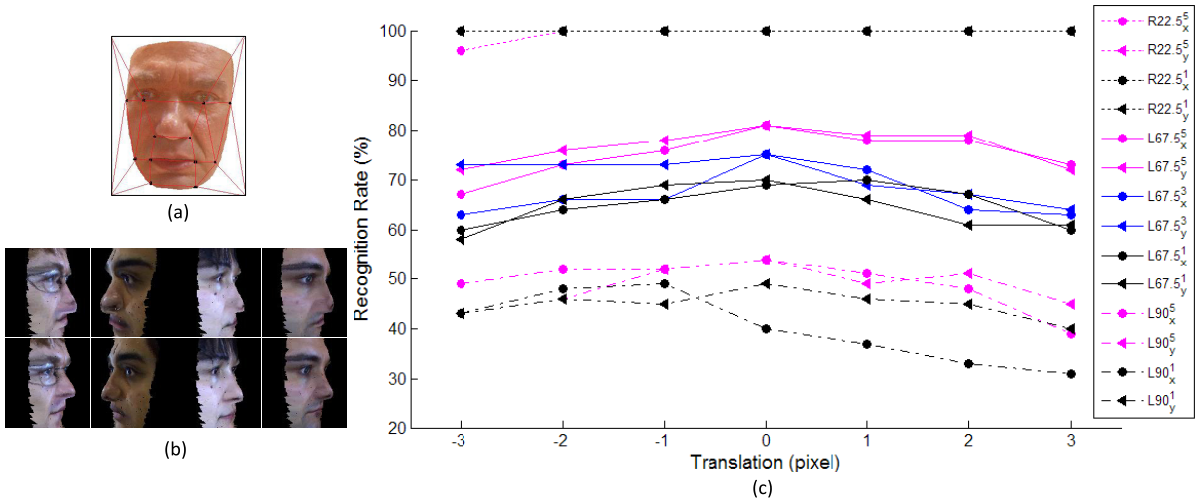

Fig. 2 (a) 12 fiducial points for piece-wise warping. (b) Reconstruction with alignment using 3 fiducial points in the upper row, and 12 fiducial points in the bottom row. (c) Translation error in x and y direction with different number of neighboring samples at pose R22.5°, L67.5° and L90°.

Following the approach in Sect. 2, one can obtain a 3D reconstructed face for each gallery image. The surface smoothing and rendering in Sect. 2.1 is performed on each reconstructed face to obtain the finalized face surface. A weak perspective transformation with a rotation matrix R_s and a translation vector t_s specified for pose $\mathbf{p}_s$ is then applied on the 3D facial surface to render its 2D projection on the image plane as the training image with pose $\mathbf{p}_s$.

To better tolerate possible pose deviation in a given probe, the training set is generated in pose-oriented clusters. Take the pose subset in the CMU PIE database [15] as an example which is used in our experiments. The pose subset offers 13 poses in total, 9 taken from horizontal views with yaw angle roughly 22.5° apart (so the central one corresponding to the frontal), 2 taken from the vertical views with pitch angle 22.5° up and down, and the rest 2 taken from *surveillance* views with yaw angle 67.5° to both side and pitch angle 22.5° down.

When generating the training set, each of these poses is considered as the center of a pose-oriented cluster, several neighboring poses are synthesized and added to the cluster. We compare the number of 5, 3, 1 with 5° interval at different poses. 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. Translation errors are calculated by misalignment of the probe face in the x and y directions around the highest performance. As the results shown in Fig. 2 (c), the better performance is achieved when using 5 neighboring poses in the gallery than 3 and 1 for all pose clusters.

3.2 Pose-Clustered Recognition with SRC

Because the Sparse Representation-based Classification (SRC) is proven effective for face recognition, especially in handling illumination, expression and occlusion [16]–[18], but rarely applied for tackling pose, it is explored in this study with different algorithms, features and parameter settings. We imposed a pose-oriented mask on each image to obtain the region of interest for feature extraction. The pose-oriented masks, made based on 12 subjects selected from

the FRGC database, are for blocking out non-facial regions in the probe images at recognition phase. The subjects selected for making the masks are different in gender and age for diversified contours in different poses. The mask for a specific pose is obtained by averaging the contours of all 12 subjects in the same pose, except for poses 67.5° and 90°. Because the contours at 67.5° and 90° are affected by nose, one of the most prominent features viewed from the sides, we selected the contour of the subject with the closest spatial distribution of the fiducial points to those in the probe.

To apply SRC, we first form a matrix $A = [A_1, A_2, \ldots, A_s]$ from the training set, where A_i denotes the subset formed by all pose-oriented clusters of Subject-i and s is the number of subjects. Each column in A_i is a normalized downsampled feature vector extracted from a training image, and the features can be pixel intensities or others. The Local Binary Pattern (LBP) is chosen as another feature in the experiments for comparison purpose. An extensive experimental study on these features is presented in Sect. 4.

Given a probe q^*, the core part of SRC considers the linear representation of q^* in the span of A, i.e., $q^* = Ar^* + \mu^*$, where r^* is a sparse vector and μ^* is a noise with bounded energy, i.e., $\|\mu^*\|_2 < \epsilon$. Following the rules in compressing sensing [16], r^* can be obtained by solving the following l_1-minimization:

$$\hat{r^*} = \text{argmin} \|r\|_1, \quad \text{subject to } \|q - Ar\|_2 \leq \epsilon \tag{8}$$

A comprehensive discussion on the solutions for the above l_1-minimization is given in [19], where five fast algorithms were evaluated on the face recognition performance under illumination variations. We select the best two, the TNIP (Truncated Newton Interior-Point) and Homotopy, from the five to evaluate their performance against pose variations. The TNIP exploits gradient projection (GP), and searches for the sparse vector r along certain gradient direction with fast convergence speed. It reformulates the problem (8) into the following form:

$$\hat{r^*} = \underset{r}{\text{argmin}} \frac{1}{2}\|q - Ar\|_2^2 + \lambda\|r\|_1 \tag{9}$$

where λ is the Lagrange multiplier. Such a formulation enables the solution using quadratic programming.

A different solution scheme, known as Homotopy, finds a solution path X_h that varies with λ,

$$X_h = \{r_\lambda^* : \lambda \in [0, \infty)\} \tag{10}$$

When $\lambda \to \infty$, $r_\lambda^* = 0$, and when $\lambda \to 0$, r_λ^* converges to the solution. The Homotopy algorithm considers the fact that the objective function in (9) changes as a homotopy from the l_2 constraint to the l_1 objective as λ decreases. It can be shown that the solution path X_h is piece-wise constant as a function of λ [17], [18]. Therefore, when constructing a decreasing sequence of λ, it is only necessary to identify the "breakpoints" that lead to changes of the support set of r^*. See [17] for more details on the computation and implementation. The Matlab programs that solve (8) using the TNIP and Homotopy are available in the SparseLab Toolbox at http://sparselab.stanford.edu/.

4. An Extensive Experimental Study

All experiments were run on a Linux PC with 2.6GHz and 4G DDR3. The reference models were taken from the FRGC database [12] and normalized to 250×300 in size. The face images in the training and testing sets were all scaled to 128×128 and resized to 24×24 when comparing performance variations with different scales. The performance was evaluated on the PIE pose subset, which has 68 subjects and 13 poses. The frontal pose of each subject was used in the gallery for enrollment and the rest poses in the probe for testing. This protocol is common for 3D-based methods. Most 2D-based methods need a "pose training set", which is often composed of all poses of half of the subjects, i.e., 34 subjects, for learning the relationship between the 13 poses. The frontal of the other 34 subjects are used as the gallery and the rest poses used as the testing set.

Experiments were designed to investigate the performance variation with the following settings:

1. Multiple reference models considered in the reconstruction phase. The accuracy of the reconstructed face is affected by the 3D reference model. We compared the case with one single reference model and one with dual reference models of different genders and ages.
2. Different numbers of fiducial points for the alignment between the gallery face and 3D reference model. Different numbers of fiducial points yield reconstructed faces of different details, five cases with 3, 6, 12, 15 and 23 fiducial points were considered. We also investigated the cases with multiple models made with different numbers of fiducial points.
3. Different features and scales. The pixel intensities in 24×24 downsampled images, and the LBP features extracted from both 24×24 downsampled and 128×128 original scales were considered.

We also compared the performance of best ones selected from this study to the state-of-the-art approaches. Although 2D-based methods were considered limited in generic applications with unconstrained poses, they were included in our comparison to highlight the need of a different evaluation protocol. When showing the results, instead of using the PIE original pose tags, such as c02, c37, ..., we use the nominal pose angle with a letter 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.

Because reconstruction takes most of the processing time, and the larger the given image I, the longer the reconstruction takes. A few scales were tested, and although large scales generally led to better reconstruction and recognition performance, the scale factor 0.3 was selected for a balance between processing time and performance.

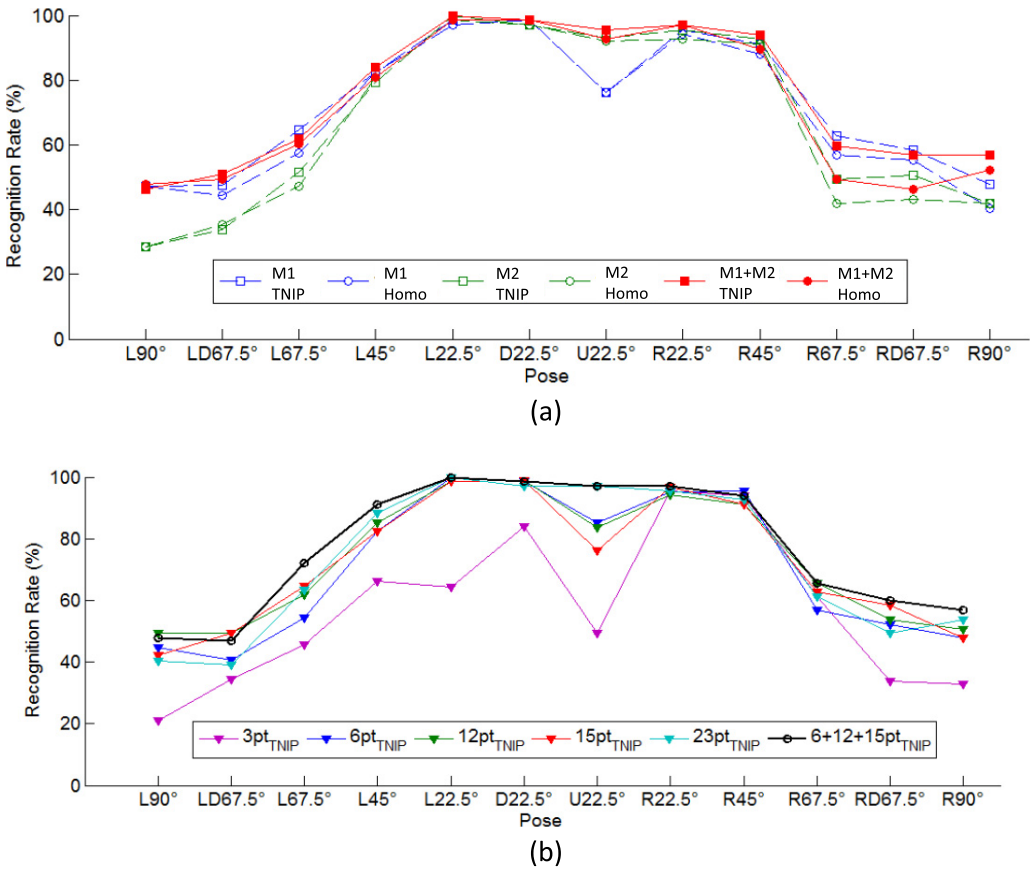

Fig. 4 (a) Comparison of single and double reference models with 15 fiducial points on 24 × 24 intensity features. (b) TNIP performance with different numbers of fiducial points on 24 × 24 images with intensity features.

Fig. 3 Two reference models with different gender and age.

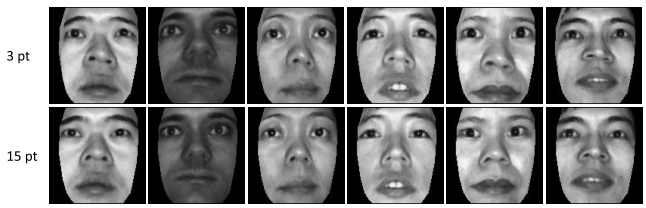

Fig. 5 Comparison of reconstruction with 3 (top row) and 15 (bottom row) fiucial points. Fewer fiducial points causes misalignment at nostrils and becomes prominent features at pose U22.5.

Multiple reference models

The default reference model was arbitrarily selected, as shown in the previous figures. We selected an additional one with different gender and age as shown in Fig. 3. Following the same approach, each gallery image had two reconstructed models, generating an additional set of pose clusters for training. Figure 4 (a) shows the comparison of single and double reference models with 15 fiducial points. The performance varies slightly for different algorithms. However, the reference model appears to make strong impacts on the performance. The model M1 performs better for large poses, while M2 is preferred for poses $\leq 45°$. But both are outperformed by their combination, M1+M2.

Different numbers of fiducial points

Different numbers of fiducial points yield reconstructed faces with different details. Five cases with 3, 6, 12, 15 and 23 fiducial points were considered. The fiducial points were used to split the face into local regions. Perspective transform and affine transform were then applied to fit a gallery image to the reference model in one region after another. A case with 12 fiducial points is shown in Figs. 2 (a) and (b). We found that the reconstruction results vary with not just the number of fiducial points, but also the locations of the fiducial points. The fiducial points at eye regions were better located right below the eyes, rather than on the center of the eyes, as shown in Figs. 2 (a) and (b). Figure 4 (b) shows the comparison of different numbers of fiducial points. Be-

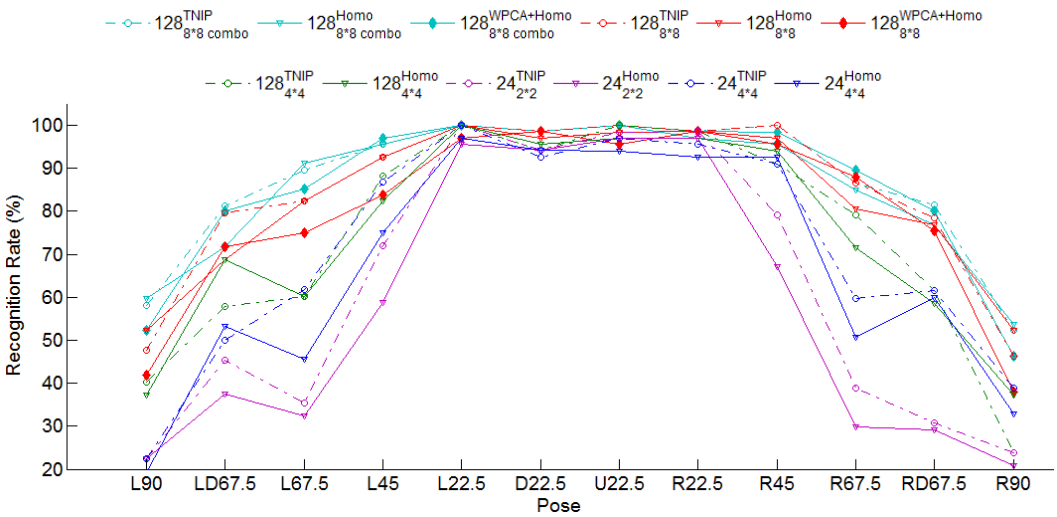

Fig. 6 Performance of LBP with different cell sizes on 24×24 and 128×128 images. Combo refers to cases with dual reference models and multiple sets of fiducial points.

cause TNIP and Homotopy algorithms perform similarly to each other, only the performance of TNIP is shown. The degraded performance at U22.5 in Fig. 4 is caused by the compound effect of imperfect reconstruction and misalignment. U22.5 refers the view from under the chin where the nostrils become prominent features as shown in Fig. 5. Because the nose is less accurately reconstructed using 3 fiducial points than the case using 15 fiducial points, the associated recognition rate of the former is also worse than the latter. Although the case with 23 fiducial points outperform others for poses $\leq 45°$, its performance degrades for large poses. As more fiducial points do not always improve the performance, an appropriate option is the combination of facial images from multiple reconstructions based on different numbers of fiducial points. It is shown in Fig. 4 (b) that the combo of $6 + 12 + 15$ fiducial points yields the best performance.

Different features and related settings

The performances of the LBP with different cell sizes and the combos of the previously reported settings are shown in Fig. 6. Those with image size 128×128 perform better than those of 24×24, reflecting the fact that texture features from higher resolution improve the performance. Because SRC with features of 128×128 pixel intensities demanded high computational cost, we consider it inappropriate and recommend the LBP features instead. However, given the facial image size 24×24, both cases with pixel intensities and LBP features perform well. Tests on 128×128 images with different LBP cell scales show that the partition with 8×8 cells performs the best. Both TNIP and Homotopy algorithms perform similarly although TNIP appears slightly better. When using multiple sets of fiducial points and dual reference models in the gallery, the size of the matrix A increases considerably and so does the recognition time. To expedite, the whitened PCA (WPCA) is exploited which is claimed to be able to effectively reduce the feature dimension [20]. Homotopy is chosen here for faster computation than TNIP. Table 1 shows the comparison on the processing time. It reduces from 2.13 to 0.18 sec using dual reference models, and from 4.81 to 0.45 sec for the combos of dual reference models and multiple fiducial point sets. Table 2 shows the dimension of WPCA features extracted from the LBP of each pose cluster†.

Table 1 Recognition time for LBP with different settings, 24_{4*4} refers to 24×24 image split in 4×4 cells and 128^{combo}_{8*8} refers to 128×128 image split in 8×8 cells handled by combo features.

	24_{4*4}	128_{4*4}	128_{8*8}	128^{combo}_{8*8}
TNIP	46	127	193	802
Homotopy	0.58	33	2.13	4.81
WPCA (average)	-	-	0.18	0.45

Table 2 WPCA dimension and recognition time for each pose cluster using dual model and multiple fiducial point sets.

Pose	L90	LD67.5	L67.5	L45	L22.5	D22.5	U22.5
Dim.	88	95	50	98	31	23	61
Time (s)	0.5	0.54	0.59	0.33	0.59	0.38	0.62

A few best ones from the previous studies were chosen to compare with the state of the art, including TNIP on 24×24 intensities (GRAY $\text{SRC}^{TNIP}_{combo}$), TNIP on LBP of 128×128 with 8×8 cells ($\text{LBP}^{TNIP}_{128d8combo}$) and Homotopy with the same LBP settings but dimension reduced by WPCA ($\text{LBP+WPCA}^{combo}_{8*8}$) as shown in Fig. 7. The combo refers dual reference models and multiple sets of fiducial points. The 3D based ones are shown in solid lines to distinguish them from the 2D based, in dashed lines, as the latter are limited in supported poses. Both $\text{LBP}^{TNIP}_{128d8combo}$ and $\text{LBP+WPCA}^{combo}_{8*8}$ perform almost equally well as the HSD [8], one of the best 3D methods but requires a few face

†The WPCA dimension can be chosen in the training phase for desired performance using the synthesized 2D images in each pose cluster.

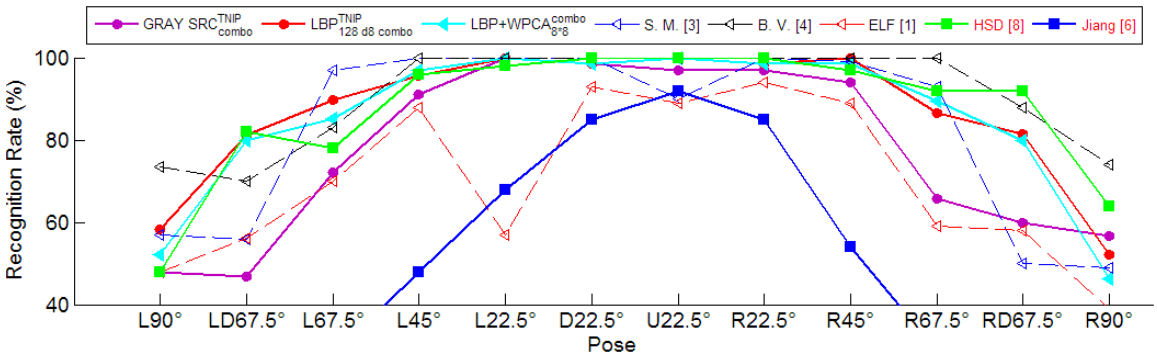

Fig. 7 Comparison with the state of the art.

images under different lighting conditions to estimate the 3-D shape. The proposed method is also competitive to B.V. (Bias Variance [4]), one of the best 2D approaches, and is more effective in the practical applications for more poses available for recognition not exist in the training set. GRAY $\text{SRC}_{combo}^{TNIP}$ performs well for poses $\leq 45°$ but degrades to some extent for larger poses, but is considered comparable to many. Although the performances of $\text{LBP}_{128d8combo}^{TNIP}$ and $\text{LBP+WPCA}_{8*8}^{combo}$ are similar to each other, the latter is better because it is much faster than the former.

5. Conclusion

3D-based approaches for cross-pose recognition deserve special attention since 2D-based ones are mostly limited to the poses same as those in the training set. This work extends a recent work on 3D face reconstruction to recognition using SRC. The smoothed surface rendering, which is an important part for reconstruction but missing in [11], has been elaborated. More importantly the impacts made by multiple reference models, multiple sets of fiducial points for alignment, various settings on the features and algorithms were also investigated, along with an extensive experimental study. This study shows that both TNIP and Homotopy algorithms perform similarly well. Multiple reference models and multiple sets of fiducial points lead to additional synthesized images and improve the recognition performance. Experiments show that SRC with downsampled images can be competitive to state-of-the-art approaches, and it can further outperform many with LBP features extracted from the original or high-resolution images.

References

[1] R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light fields," TPAMI, vol.26, pp.449–465, 2004.
[2] S.J. Prince, J.H. Elder, J. Warrell, and F.M. Felisberti, "Tied factor analysis for face recognition across large pose differences," TPAMI, vol.30, pp.970–984, June 2008.
[3] C.D. Castillo and D.W. Jacobs, "Using stereo matching for 2-D face recognition across pose," CVPR, pp.1–8, 2007.
[4] A. Li, S. Shan, and W. Gao, "Coupled bias-variance tradeoff for cross-pose face recognition," IEEE Trans. Image Process. vol.21, no.1, pp.305–315, 2012.
[5] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," TPAMI, vol.25, no.9, pp.1063–1074, Sept. 2003.
[6] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, and W. Gao, "Efficient 3D reconstruction for face recognition," Pattern Recognit., vol.38, pp.787–798, June 2005.
[7] U. Prabhu, J. Heo, and M. Savvides, "Unconstrained pose-invariant face recognition using 3D generic elastic models," TPAMI, vol.33, pp.1952–1961, 2011.
[8] X. Zhang and Y. 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.506–517, 2012.
[9] X. Zhang and Y. Gao, "Face recognition across pose: A review," Pattern Recognit., vol.42, pp.2876–2896, Nov. 2009.
[10] K. Niinuma, H. Han, and A.K. Jain, "Automatic multi-view face recognition via 3D model based pose regularization," IEEE 6th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, 2013.
[11] I. Kemelmacher-Shlizerman and R. Basri, "3D face reconstruction from a single image using a single reference face shape," TPAMI, vol.33, no.2, pp.394–405, Feb. 2011.
[12] P. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," CVPR, pp.947–954, 2005.
[13] M. Alexa, J. Behr, D. Cohen-Or, S. Fleishman, D. Levin, and C.T. Silva, "Computing and rendering point set surfaces," IEEE Trans. Vis. Comput. Graph., vol.9, no.1, pp.3–15, 2003.
[14] L. Ding and A.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.
[15] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," Proc. IEEE Conf. Automatic Face and Gesture Recognition, pp.46–51, 2002.
[16] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.2, pp.210–227, 2009.
[17] A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Fast l1-minimization algorithms and an application in robust face recognition: A review," Technical Report, no.UCB/EECS-2010-13, 2010.
[18] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.9, pp.1864–1870, 2012.
[19] A.Y. Yang, S.S. Sastry, A. Ganesh, and Y. Ma, "Fast $l1$-minimization algorithms and an application in robust face recognition: A review," ICIP, pp.1849–1852, 2010.
[20] W. Deng, J. Hu, J. Guo, W. Cai, and D.D. Feng, "Robust, accurate and efficient face recognition from a single training image: A uniform pursuit approach," Pattern Recognit., vol.43, no.5, pp.1748–1762, 2010.

Gee-Sern Hsu received the dual M.S. degree in electrical and mechanical engineering and the Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, in 1993 and 1995, respectively. From 1995 to 1996, he was a postdoctoral fellow with the University of Michigan. From 1997 to 2000, he was a senior research staff with the National University of Singapore, Singapore. In 2001, he joined Penpower Technology, where he lead research on face recognition and intelligent video surveillance. In 2007, he joined the Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, as an assistant professor. His research interests include object recognition, particularly face recognition, pedestrian detection, and vehicle license plate recognition. He received the Best Innovation Awards at the SecuTech Expo in 2005, 2006, and 2007, along with his team at Penpower Technology.

Hsiao-Chia Peng received the B.S. degree in mechanical engineering from National Chung Cheng University, Chiayi, Taiwan in 2007. She is a Ph.D. student in mechanical engineering from National Taiwan University of Science and Tecnnology, Taipei, Taiwan, from 2008 until now. Her research interests include image processing, pattern recognition and face recognition.

Ding-Yu Lin received the M.S. degrees in mechanical engineering from National Taiwan University of Science and Technology, Taipei, Taiwan in 2013. He is currently with the mandatory military service,Taiwan from 2013 until now. His research interests include image processing, pattern recognition, and face recognition.

Chyi-Yeu Lin is currently a full professor in Department of Mechanical Engineering and the director of Center for Intelligent Robotics in National Taiwan University of Science and Technology (NTUST). In NTUST, he had served as the director of Center of Technology Transfer and the director of Center of Innovation and Creativity both between 2007 and 2009, and vice-chairman and chairman of Department of Mechanical Engineering between 2006 to 2008 and 2009 to 2012, respectively. He received his Ph.D. degree from University of Florida in 1991. From 1991 to 2001, his major research interests were structural optimization and evolutionary methods. After 2001, he switched his interests to intelligent robotics. He and the research teams he led had created many intelligent robots including two androids and two dual-wheeled robots that performed in the world-first robot theater involving male and female bipedal humanoid robots in 2008.