Landmark Based Facial Component Reconstruction for Recognition Across Pose

Gee-Sern Hsu*, Hsiao-Chia Peng and Kai-Hsiang Chang
Department of Mechanical Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan
Email: *jison@mail.ntust.edu.tw

***Abstract*—Different from previous 3D face modeling approaches that consider the whole facial area, the proposed method reconstructs 3D facial components for handling cross-pose recognition. It has two phases, component reconstruction and component-based recognition. In the reconstruction phase, we first extract four component regions, namely two eyes, nose and mouth, from each gallery face using the pose-invariant landmarks obtained by a modified version of a landmark detection algorithm. A 3D model of each component region is reconstructed using a constrained minimization scheme with a gender and ethnicity oriented 3D model as the reference. In the recognition phase, the pose of a given probe is determined by a set of landmarks which guides the rotation of the reconstructed components so that the reconstructed can be aligned to the probe components. The match is determined by the components instead of the whole faces so that different components can be considered at different poses. Experiments on the PIE and Multi-PIE databases show that the proposed component-based approach does not just outperform its holistic counterpart, but is also competitive to many contemporary methods.**

***Index Terms*—face recognition; face reconstruction; 3D facial component;**

I. Introduction

Face recognition across pose is generally tackled by either 2D based or 3D based approaches [1]. 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 proposed approach is 3D in nature.

Quite a few 3D-based methods have been developed in the last decade. Morphable model [2] uses the prior knowledge, including the 3D face shapes and textures collected from hundreds of 3D scans to build a 3D model for a 2D face. Although regarded as a good solution for cross-pose recognition, it is expensive in storage and computation because of the huge amount of dense 3D scan data required. Recently, the Generic Elastic Model (GEM) [3] claims that the depth of a gallery face can be accurately reconstructed by a generic depth map with 2D dense meshes built on the landmarks on both the gallery face and generic model. The landmarks for the frontal pose are obtained by the MASM (Modified Active Shape Model), but those for the non-frontal poses are obtained by a commercial tool. It is further improved in an extension work [4]. Although the performance in [4] appears better than the original GEM [3] for a limited range of poses on Multi-PIE [5], both works ignore the performance handling large rotations, such as yaw angle 70° or larger. Arguing that many 3D face models based on the Lambertian assumption ignore specular and diffuse reflection, a Heterogeneous Specular and Diffuse (HSD) 3D surface approximation is proposed in [6], and is experimentally proven effective handling extreme poses in the PIE database [7]. Nevertheless, the fact that multiple frontal images with various illumination conditions are required for the HSD surface approximation substantially weakens its practical applicability. The method in [8] exploits the view-based Active Appearance Model, landmark detection and regression to estimate the pose of a probe. The probe face with the estimated pose is then aligned to a 3D head model, and the aligned face model is rotated back to the frontal view to match against the faces in the gallery. A pose adaptive filter is proposed in [9] that uses a deformable model for pose estimation. The pose correction is applied in the filter space rather than the regular image space, making this method less affected by the precision of the 3D model. It combines the holistic pose transformation and local Gabor filtering to make the extracted features robust to pose.

Component-based methods are popular for 2D face recognition [10], [11], they are, however, rarely attempted in 3D modeling for cross-pose recognition. Since the visible region of a face varies as pose changes, part of the face appears clear and good to recognize at some pose, while the other part becomes hidden or invisible. The proposed method aims to identify the component regions visible for a certain pose range, and unify these component regions for recognition. It has two phases, component reconstruction and component-based recognition. In the component reconstruction phase, we extract four component regions, including the eyes, nose and mouth, from each gallery face using the pose-invariant landmarks obtained by a modified version of an automatic landmark detection algorithm [12]. The 3D model of each component region is reconstructed using a constrained minimization scheme with a gender and ethnicity oriented 3D model as the initial reference. In the component-based recognition phase, the landmarks of a probe face are first detected and used for pose estimation and component extraction. Given the estimated pose of the probe, the component images with the same pose

are generated based on the reconstructed component models and matched against the probe. The overall workflow is shown in Fig. 1.

The novelty of this work is on the component-based 3D reconstruction and recognition. It integrates and improves several independent techniques, including pose estimation, 3D surface reconstruction and SRC based recognition. None of these techniques has implied such an integration. To the best of our knowledge, this is the first component-based approach for 3D based recognition. Many use facial landmarks for pose alignment and size normalization, we extended their application to 3D component segmentation.

The rest of the paper is organized as follows: the landmark-based component segmentation and reconstruction is presented in Sec. II. Recognition with landmark-assisted alignment of the reconstructed components to the pose of a given probe is given in Sec. III where we exploit the SRC for decision making. An experimental performance study on the PIE and Multi-PIE databases and a comparison with contemporary approaches are given in Sec. IV, followed by a conclusion in Sec. V.

II. Facial Component Reconstruction

A. Face Segmentation by Pose-Invariant Landmarks

Facial landmark detection has been advanced considerably in recent years [13], [14], [12]. The Zhu-Ramanan model [12] can simultaneously solve face detection, landmark localization, and pose estimation. It is experimentally proven better than most state of the art in all three tasks. The core part of the method is a mixtures of trees with a shared pool of parts; each facial landmark is modeled as a part and a global mixture is used to capture the topological variations across poses. The model detects 68 landmarks for yaw angle between L45 (45° to the left) to R45 (45° to the right) and 39 landmarks for yaw angle beyond this range.

We ran an experiment on the pose subsets of the Multi-PIE database [5] in which 80 subjects were randomly selected from Session 2 as the training set and tested on all subjects in Session 1. This experiment was to study the consistency of landmark localization across poses, determine those which were invariant to some range of pose variation, and use them for component segmentation. Part of the landmarks, especially those along facial contours, varied in locations with different poses, and thus were removed. Only those whose locations were invariant across a range of poses were selected as the pose-invariant subsets. Fig. 2 shows four such subsets on poses with yaw angles $\leq 45°$, and three subsets with yaw $\geq 60°$. The landmarks denoted by $+$ are obtained automatically using the Zhu-Ramanan model, and those in $\triangle$ are add-ons to the originally detected ones so that the facial component regions can be better confined. These add-ons are obtained by some ad-hoc rules: those on the bottom boundary of the eye region are the mirror reflections of the ones detected on the eyebrow, considering those on the eye as a mirror; the two at the ends of the bottom boundary of the nose region are the extrapolation of those detected at the nostrils. The segmented region of interest (ROI) of the component region is the smallest rectangle that encloses all landmarks.

B. 3D Reconstruction of Facial Components using Ethnicity and Gender Reference Model

A scheme, improved from the face reconstruction in [15], is explored for the 3D reconstruction of the 2D segmented components. Given a 2D segmented component as the target $t(x, y)$ and a 3D scan of the same component but from a different face as a reference, it recursively estimates the surface reflectance $R(x, y)$, depth $z(x, y)$, albedo $\rho(x, y)$ and surface normal $\vec{n}(x, y)$ of the target with the surface parameters of the reference for initialization. The improvements of this scheme over the original [15] include the following:

- The minimization in the core part of [15] concerns the whole face, and hence works well on the low frequency (or smooth) regions, such as cheeks, but with errors on the high frequency regions, such as eyes, nose and mouth. The errors on the high frequency regions are evened out when computing the overall error of the face to be minimized, making further error reduction difficult. The proposed scheme only minimizes the errors at the components, which are exactly the high frequency regions, and ignores low frequency regions. Because the components reveal better discriminating features, the scheme would lead to better recognition.
- Because only components are considered, the scheme comes with a better computational and storage cost.
- Instead of using one single reference model, we use a small set of reference models with gender and ethnicity same as the gallery face. As the 3D reconstruction is for a 2D frontal face, the recovery of the depth, which is missing in the 2D face, is strongly affected by the 3D reference model. Experiments show that the side view of a reconstructed component, which dominates the recognition performance at extreme poses, can be substantially improved if the reference model can be of the same ethnicity and gender. It is considered legitimate to be able to enter a subject's name tag, gender and ethnicity when enrolling his/her face to the gallery.

We select the reference faces from the 3D scan subset of the FRGC database [16]. Fig. 3 shows the average depth along the landmarks on the noses of samples randomly selected from four groups, Caucasian male (CM) and female (CF) and Asian male (AM) and female (AF). The Caucasian shows higher nose and larger depth variation than the Asian. The depth difference is primarily caused by ethnicity rather than gender.

Given the depth $z_r(x, y)$, surface normal $\vec{n}_r(x, y)$ and albedo $\rho_r(x, y)$ of the component reference model, the reconstruction is to estimate the 3D shape of the 2D target component $t(x, y)$ segmented from a gallery face. Assuming that the component surface is Lambertian, $t(x, y)$ can be written as $t(x, y) = \rho(x, y)\vec{h}(x, y) \cdot \vec{n}(x, y) = \rho(x, y)R(x, y)$, where $\rho(x, y)$ is the surface albedo at (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

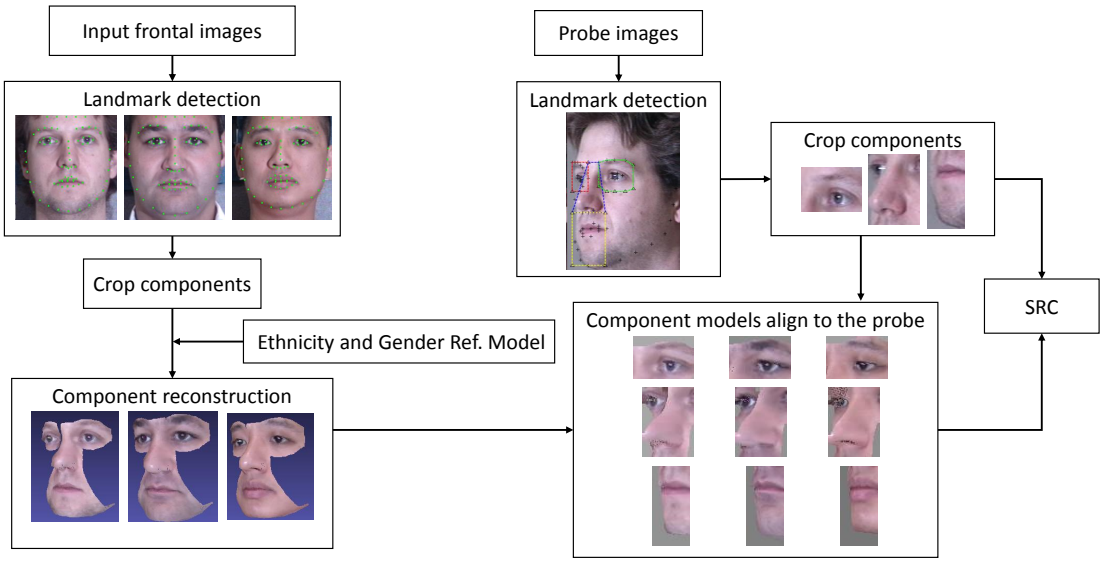

Fig. 1. Workflow of the proposed method.

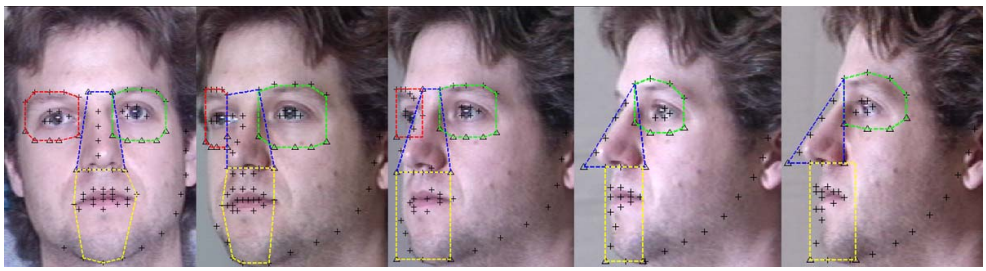

Fig. 2. Pose-invariant landmarks in poses F (frontal), L30 (30° to the left), L45, L60 and L75. Those in "+" are obtained using the Zhu-Ramanan model [12] and those in △ are add-ons so that the component regions can be better confined. The landmarks in the regions enclosed and passed by dashed lines are selected as pose invariant.

the reflectance $R(x,y) = \vec{h}(x,y) \cdot \vec{n}(x,y)$. For simplicity of notation, the coordinates (x,y) are dropped in the rest of the paper, and $\vec{n}(x,y)$, for example, is written as $\vec{n}$.

With a few assumptions [15], the reflectance can be approximated using spherical harmonics, i.e.,

$$R(x,y) \approx \vec{l} \cdot \vec{Y}(\vec{n}) \tag{1}$$

where $\vec{l}$ is the lighting coefficient vector and $\vec{Y}(\vec{n})$ is the spherical harmonic vector [17], which, in the second order approximation, takes the following form:

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

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 $\vec{h} \cdot \vec{n}$ and $\vec{l} \cdot \vec{Y}(\vec{n})$ is that the lighting intensity and direction are all merged into $\vec{h}$ in the former, separated from $\vec{n}$, but in the latter 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. If the target t can be aligned with the reference model using the landmarks, the core problem can be formulated as the minimization of $||t - \rho\vec{l} \cdot \vec{Y}(\vec{n})||$ over ρ, $\vec{l}$ and $\vec{n}$, i.e.,

$$\min_{\vec{l},\vec{z},\rho} \int (t - \rho\vec{l} \cdot \vec{Y}(\vec{n}))^2 + \lambda_1 (L_g * d_z)^2 + \lambda_2 (L_g * d_\rho)^2 dxdy \tag{3}$$

where $d_z = z(x,y) - z_r(x,y)$, $d_\rho = \rho(x,y) - \rho_r(x,y)$, and L_g* denotes the convolution with the Laplacian of Gaussian (LoG); λ_1 and λ_2 are constants. LoG is used to locate large differences in depth and albedo, and force the minimization performed on the spots with the large differences. The formulation in (3) can be interpreted as the minimization of $||t - \rho\vec{l} \cdot \vec{Y}(\vec{n})||$ subject to the constraints $L_g * d_z \approx 0$ and $L_g * d_\rho \approx 0$, and λ_1 and λ_2 are the Lagrange multipliers. Assuming that the target t is aligned to the 3D reference model, the reconstruction processes the minimization in (3) by first solving for the spherical harmonic coefficients $\vec{l}(x,y)$ using the references $\vec{n}_r$ and ρ_r, then the depth $z(x,y)$ by writing $\vec{n}_r$ into the following form:

$$\vec{n}_r = (p, q, -1)^T / \sqrt{(p^2 + q^2 + 1)}$$

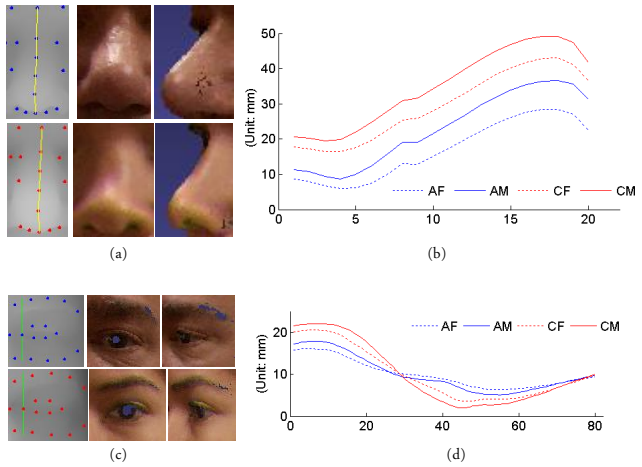

Fig. 3. (a) shows the depth map and depth variation at nose, viewed in the front and 60° from the side. (b) shows the mean depth (in mm) along the landmarks on the nose (in yellow line). (c) shows the depth map and depth variation at eye sockets, and (d) shows the mean depth (in mm) along the inner side of the eye socket (the green lines in (c)). Samples include 86 CMs, 118 CFs, 38 AMs and 46 AFs randomly selected from the FRGC 3D scan subset. Eye corner is taken as the ground level.

where $p = \partial z/\partial x$ and $p = \partial z/\partial y$. If $z(x, y)$ is solved, then the albedo $\rho(x, y)$ can be recomputed. This process is repeated until the estimates of the spherical harmonic coefficients, depth and albedo converge.

III. Component-based recognition with SRC

We consider a common scenario that the gallery has a single frontal face image per subject, and the probe set contains images of other poses for recognition. The 3D facial component models of each gallery face are first reconstructed following the above approach. Because each 3D component model is reconstructed on the frontal scanned reference model, the depth points on the reconstructed model appear dense to the front, but many null spots appear when viewed from the sides. We fill the null spots by the triangle mesh computed on each depth point with two nearest neighbors. When a probe image is given, its pose is first estimated using the Zhu-Ramanan model with corresponding landmarks detected, and its facial components are cropped using the detected and add-on landmarks. The reconstructed component models built on the gallery set are then rotated to the estimated pose of the probe so that the 2D projection of the reconstructed components can be aligned with the 2D probe components.

Since the Sparse Representation-based Classification (SRC) is proven effective handling illumination, expression and occlusion in [18], [19], but rarely attempted for tackling pose and facial components, it is explored in this study. Given a set of k projections of the reconstructed components, denoted as $M = [m_1, m_2, \cdots, m_k]$, and the same component of a probe q, all labeled with the aforementioned landmarks, the core part of SRC solves for the linear representation of q in the span of A, where $A = [a_1, a_2, \cdots, a_m]$ is a matrix with its column a_i being a feature vector extracted from m_i. One can therefore write $q^* = Ar^* + \mu^*$, 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 [18], r^* can be obtained by solving the following l_1-minimization:

$$\hat{r^*} = \text{argmin}\,||r||_1, \quad \text{subject to } ||q - Ar||_2 \leq \epsilon \qquad (4)$$

We have compared different features, including pixel intensities, LBP (Local Binary Patterns) and Gabor features (obtained by the Gabor transform), and the Gabor features result in the best overall performance. We exploit the Homotopy algorithm [20] to find a solution path X_h that varies with a parameter λ, i.e., $X_h = \{r^*_\lambda : \lambda \in [0, \infty)\}$. When $\lambda \to \infty$, $r^*_\lambda = 0$, and when $\lambda \to 0$, r^*_λ converges to the solution. We used the Matlab programs available in the SparseLab Toolbox (http://sparselab.stanford.edu/) to solve (4).

The recognition by each component is determined by the Rank-1 result, and the overall recognition is determined simply by the votes from all components. When the votes are tied, which were observed in cases with four components, the Rank-2 results would be taken into account.

IV. Performance Evaluation

The experiments were carried out on the PIE database (68 subjects) and Session 1 of the Multi-PIE database (249 subjects), and the reference models were arbitrarily chosen from the FRGC database. Each subject has one single frontal face in the gallery and the rest of the poses were all in the probe set. The pose range in PIE covered up to 90° in yaw, and up to 75° on MPIE. Both sets were under the same illumination conditions. The experiments were designed to answer the following issues:

1) Impacts made by reference models of different ethnicity and gender. The frontal view of a nose and eye socket cannot reveal the depth. This fact imposes a strong constraint on how well one can estimate the depth using the scheme presented in Sec. II-B, which only uses a frontal view for depth recovery. Therefore, the depth of the reference model would play an important role as it is a dominant factor.
2) Comparison with the holistic counterpart of the proposed method in which the reconstruction is carried out for the whole face. This comparison would reveal the advantages of the component-based over the holistic one.
3) Comparison with other state-of-the-art 3D-based approaches. Since 2D-based approaches for cross-pose recognition adopt a different setup, for example, the requirement of a multi-pose training set, only those which are 3D-based in nature are considered in this comparison.

Fig. 4 reveals the impacts made by reference models of different genders and ethnic backgrounds. The best performance is observed when the reconstruction is based on the reference model of the same ethnicity and gender (E+G). The Caucasian male (CM) reference model can lead to the best performance if only one single reference model is allowed. The Caucasian female (CF) comes as the close second. However, both Asian

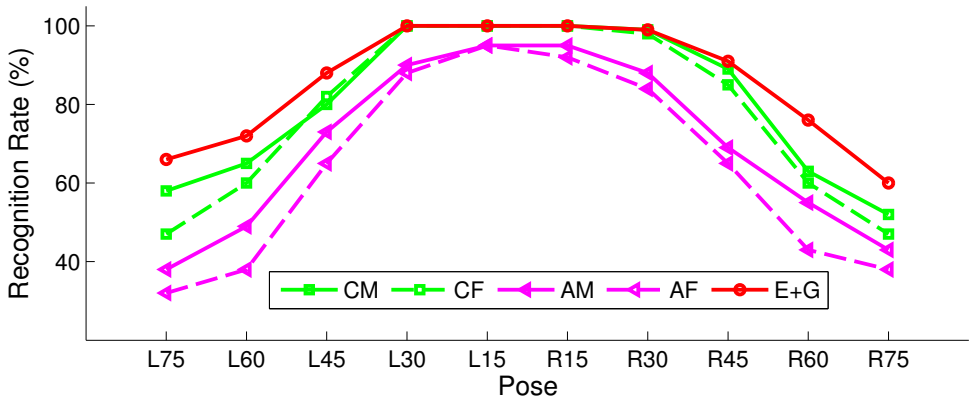

Fig. 4. Comparison of cases with specific and E+G (ethnicity and gender) oriented reference model on Multi-PIE. Four different reference models include Caucasian male (CM) and female (CM), Asian male (AM) and female (AF).

male and female models, AM and AF, perform relatively poor. This result indicates that the impact made by ethnicity appears much stronger than that made by gender, which is agreeable to the observation in Fig. 3 about the depths of different reference models. It also reflects the demographics of the dataset, in which the majority is CM, followed by CF, then AM, and then AF.

The proposed component-based approach outperforms its holistic counterpart in both computation cost and accuracy. The holistic takes more than 5 mins for one single face reconstruction on a Windows PC with CPU 2.6 GHz and RAM 2.4 GB, while the component-based takes only 38 secs. The comparisons of the holistic and component-based with the state-of-the-art are shown in Fig. 5 and Fig. 6, on Multi-PIE and PIE, respectively. With reference models of the same ethnicity and gender (E+G), both perform better than the selected contemporary methods on Multi-PIE. Quite a few 3D approaches that have been evaluated on the PIE pose subset reveal a significant drop on the recognition rate for yaw angle larger than 67.5°, as shown in Fig. 6. This big drop is also observed on the 3D holistic (E+G); however, the 3D component (E+G) maintains its performance at 67.5°, and outperforms many of the contemporary ones.

Fig. 7 shows a few cases where the face can be recognized for rotation angle less than 45°, but failed for extreme poses. Fig. 7(b) shows the reconstructed nose with the ground-truth at 75°, and it can be seen that the reconstructed fails to capture the real nose ala. Fig. 7(d) shows a failed case with eyeglasses, which is modeled as part of the eye socket. Both failures are not seen in small poses, as shown in Fig. 7(a) and 7(c).

V. Conclusion

This can be the first attempt using a 3D component-based approach to tackle cross-pose recognition. We have experimentally proven that the component-based is better than the holistic in not just the performance but also the computational cost. This study can serve as a sample transforming other 3D holistic methods into component-based, and taking the advantages of the substantial progress made on facial landmark detection and pose estimation in recent years. We have also shown that the model-based face reconstruction is better built on reference models of the same ethnicity and gender, which can be feasible as we often enter personal information while enrolling one's face to the database. As the eyeglasses pose a threat to our approach, we would consider building it a template on the reference model in the next phase.

References

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

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

[3] U. Prabhu, J. Heo, and M. Savvides, "Unconstrained pose-invariant face recognition using 3d generic elastic models," *TPAMI*, vol. 33, pp. 1952–1961, 2011.

[4] 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," *TPAMI*, vol. 34, no. 12, pp. 2341–2350, 2012.

[5] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *IVC*, vol. 28, pp. 807–813, May 2010.

[6] 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. 1952–1961, 2012.

[7] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," in *AFGR*, 2002, pp. 46–51.

[8] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. V. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *ICCV*, 2011, pp. 937–944.

[9] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *CVPR*. IEEE, 2013, pp. 3539–3545.

[10] B. Heisele, T. Serre, and T. Poggio, "A component-based framework for face detection and identification," *IJCV*, vol. 74, no. 2, pp. 167–181, 2007.

[11] P.-H. Lee, G.-S. Hsu, T. Chen, and Y.-P. Hung, "Facial trait code," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 23, no. 4, pp. 648–660, 2013.

[12] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*. IEEE, 2012, pp. 2879–2886.

[13] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *CVPR*, 2011, pp. 545–552.

[14] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *IJCV*, vol. 91, no. 2, pp. 200–215, 2011.

[15] 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.

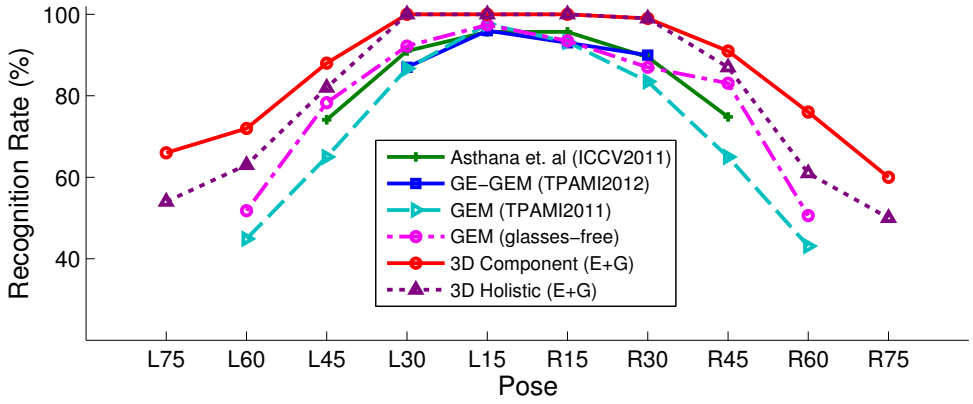

Fig. 5. Comparison with the contemporary 3D approaches, including GEM [3], Asthana et al. [8] and GE-GEM [4], on the Multi-PIE database.

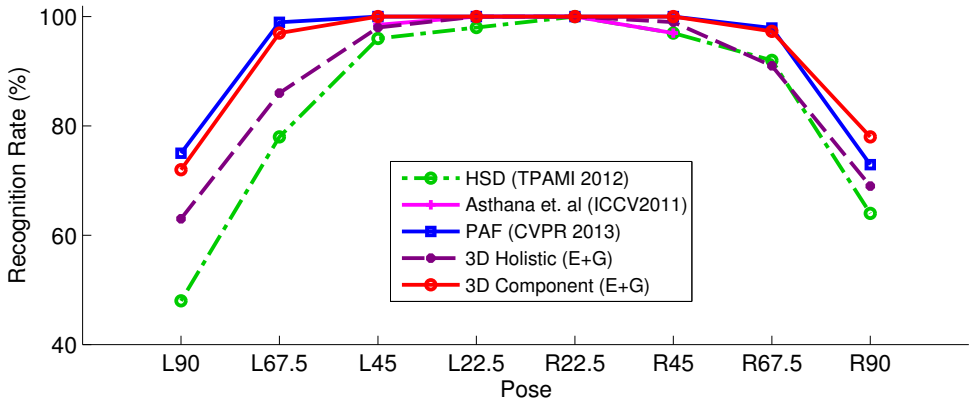

Fig. 6. Comparison with the case that the proposed scheme applied to the whole face, and with the contemporary 3D approaches including HSD [6], Asthana et al. [8] and PAF [9] on the PIE database

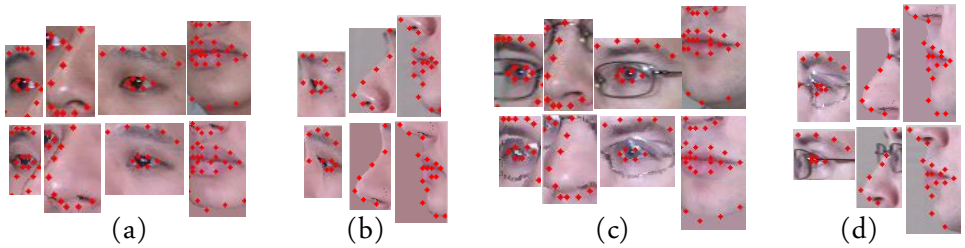

Fig. 7. Comparison of the ground-truth (top row) and the reconstructed (bottom row), (a) and (b) are from the same subject, but (a) with yaw 30° and (b) with 75°, same settings used for a different subject with eyeglasses in (c) and (d).

[16] 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," in *CVPR*, vol. 1, 2005, pp. 947–954.

[17] M. Hazewinkel, *Encyclopaedia of mathematics : an updated and annotated translation of the Soviet "Mathematical encyclopaedia*, 2001st ed. Dordrecht Boston Norwell, MA, U.S.A: Reidel Sold and distributed in the U.S.A. and Canada by Kluwer Academic Publishers.

[18] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *TPAMI*, vol. 31, no. 2, pp. 210–227, 2009.

[19] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *TPAMI*, vol. 34, no. 9, pp. 1864–1870, 2012.

[20] 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.