Image and Vision Computing 58 (2017) 193–203

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Examplar coherent 3D face reconstruction from forensic mugshot database ☆

Dan Zeng, Qijun Zhao, Shuqin Long, Jing Li*

College of Computer Science, Sichuan University, Chengdu, 610065, China

ARTICLE INFO

Article history:
Received 8 October 2015
Accepted 17 March 2016
Available online 31 March 2016

Keywords:
3D face reconstruction
Forensic mugshot images
Examplar coherent

ABSTRACT

Reconstructing 3D face models from 2D face images is usually done by using a single reference 3D face model or some gender/ethnicity specific 3D face models. However, different persons, even those of the same gender or ethnicity, usually have significantly different faces in terms of their overall appearance, which forms the base of person recognition *via* faces. Consequently, existing 3D reference model based methods have limited capability of reconstructing precise 3D face models for a large variety of persons. In this paper, we propose to explore a reservoir of diverse reference models for 3D face reconstruction from forensic mugshot face images, where facial examplars coherent with the input determine the final shape estimation. Specifically, our 3D face reconstruction is formulated as an energy minimization problem with: 1) shading constraint from multiple input face images, 2) distortion and self-occlusion based color consistency between different views, and 3) depth uncertainty based smoothness constraint on adjacent pixels. The proposed energy is minimized in a coarse to fine way, where the shape refinement step is done by using a multi-label segmentation algorithm. Experimental results on challenging datasets demonstrate that the proposed algorithm is capable of recovering high quality 3D face models. We also show that our reconstructed models successfully boost face recognition accuracy.

1. Introduction

Three dimensional (3D) face models have been extensively used in face recognition task under unconstrained environment, thanks to their capability of addressing the problem of pose, illumination, and expression variations that commonly exist in natural images. It is, however, both expensive and tedious to collect 3D face data by using 3D scanners. On the other hand, there are already plenty of two dimensional (2D) face images available from various sources, such as social media and forensic databases. Moreover, the number of 2D face images keeps increasing rapidly every day. Therefore, it is of significant importance to develop methods that can reconstruct 3D face models from these 2D face images. In this paper, we focus particularly on the problem of reconstructing 3D face models from one frontal and two profile face images. Such uncalibrated 2D face images widely exist in forensic mugshot databases and are routinely used by police officers.

A number of 3D face reconstruction methods have been proposed in the literature. Most of prior arts use existing 3D face models as reference, which provides scale of the face or depth of the facial components that are missing in the input uncalibrated 2D face images. Reference 3D face models are thus needed to serve as a constraint on the estimated 3D face model. Although impressive results have been reported using such prior knowledge, these approaches have difficulties to accurately reconstruct the 3D face models for persons with tremendous appearance/shape difference, because they use only a generic reference model or a few ethnicity specific reference models. But different persons, even those of the same ethnicity, usually have considerably different face shapes, which forms the basis of person recognition *via* faces. Fortunately, it is possible that different persons may share some similar facial components, though their faces are different in overall appearance. This motivates us to propose an example coherent 3D face reconstruction method as presented in this paper.

Unlike previous reference model based methods, our proposed method explores a reservoir of diverse reference models and searches for each component in the input face appropriate depth guesses from the reservoir (Fig. 1). This way, it is of high probability for us to find the most alike candidate for each component of the input face. Our proposed method has several advantages: (i) Compared with single reference model based approaches, our method can effectively exploit more abundant reference models to avoid heavy homogeneity of the reconstructed 3D face shape models;

☆ This paper has been recommended for acceptance by Michele Nappi.

* Corresponding author.
E-mail address: lijing712@scu.edu.cn (J. Li).

http://dx.doi.org/10.1016/j.imavis.2016.03.001
0262-8856/

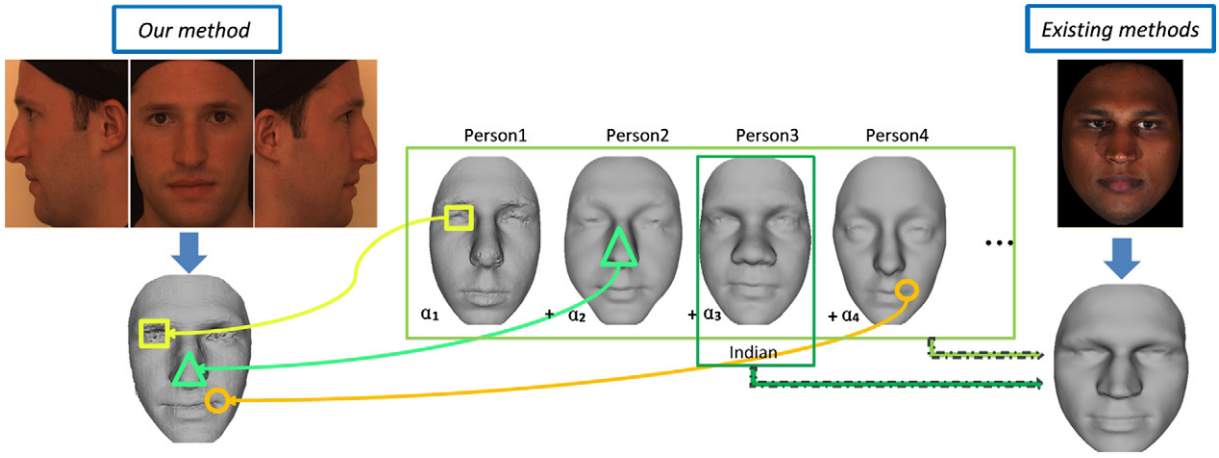

Fig. 1. Existing methods linearly combined multiple reference models to reconstruct 3D face model from the input 2D image, or only use one generic model or a gender/ethnicity specific average model [1]. Our proposed examplar coherent method uses multiple reference models, each for some component in the input face. This way, our method can obtain more accurate reconstructed 3D face models.

(ii) While expressing a novel face as a linear combination from multiple basis 3D face models (*i.e.* 3D Morphable Model) may suffer from fine detailed reconstruction, our method is able to overcome this issue thanks to pixel-wise optimization.

Specifically, we formulate the reconstruction problem by minimizing an energy function under the following constraint: shading consistency between the desired shape and the input face images, distortion-preserving color consistency between adjacent views, as well as uncertainty based depth smoothness within local neighborhoods. The proposed energy is solved from coarse to fine, by first estimating an initial shape from shading constraint, and then reformulate the problem of searching for reference models and regularizing the refined 3D face model into one unified optimization framework under Markov Random Field (MRF). Thus, the reconstruction problem is converted into a multi-label segmentation problem and is solved using sophisticated algorithms.

The rest of this paper is organized as follows. Related work is shown in Section 2. Section 3 describes the problem at hand. The proposed energy is described in Section 4, together with the optimization method. Experimental results and discussions are shown in Section 5. And concluding remarks are drawn in the end.

2. Related work

Reconstructing 3D face shape from 2D face images has long been explored. As the problem is ill-posed, usually additional information is needed to reach the solution. Motivated by the importance and unique properties of faces, many researchers utilize prior information to regulate the final face model, such as shape prior in terms of reference models, or illumination assumption of the face. On the other hand, adding additional input (*e.g.*, multi-view face images) also improves the performance of 3D face reconstruction. In fact, 3D face reconstruction from forensic mugshot images including frontal and profile 2D face images has potential applications on public security. In this section, we will review existing work on prior based 3D face reconstruction, as well as face shape estimation methods from forensic images.

2.1. Prior based approach

The prior based method has advantages of reducing the solution space and simplifying the problem at hand.

The traditional Shape from Shading (SFS) methods [2–4] required Lambertian reflectance property and an unknown light source direction to produce accurate models. These impractical constraints are discarded by introducing an additional reference model [5,6], which was based on the observation that different faces look similar globally but vary considerably across individuals in detail. Given an arbitrary reference face, its known lighting, albedo and shape information can be used as prior and be optimized iteratively in the reconstruction process. One single reference model can also help reconstructing 3D faces from unconstraint images [7], by successively estimating the coarse model from landmark driven 3D warping, and further refining the model using photometric normal constraint. This method can produce fine detailed geometry; however, its computational time is unaffordable. The severe condition required in SFS can also be overcame using learning-based approaches. Refs. [8–12] reconstructed a 3D face by modeling the relationship between intensity and depth, using statistical learning techniques such as PCA, partial least squares, or canonical correlation analysis. One drawback of these methods is that the pixel by pixel alignment between intensities and depthmaps is required.

For the single reference model based approaches, Ref. [13] reconstructed the face shape from an arbitrary reference model, by optimizing the displacement between the reference and the final estimation through joint depth-appearance similarity. This arbitrary reference model based method is able to produce visually pleasant results, but is sensitive to the reference model in use and may suffer from inaccurate reconstruction. To reduce the effect of arbitrary reference model, Ref. [14] synthesized an identity-preserving reference model *via* photo collections of the same person. The reconstruction process in Ref. [14] first estimated a dense 3D flow from 2D face image, and then optimized an objective function based on shape from shading constraint. The identity-preserving model was produced using technique from Ref. [15], by leveraging multiple pose-and-expression-normalized image shading.

3D Morphable Models (3DMM) [16,17] based methods fall into multiple reference model category. 3DMM is a crucial and widely used method to estimate 3D shape of the face that fits a morphable model to a 2D image. Linear combination of basis models is estimated in terms of both shape and texture, and the reconstructed shape is represented by model parameters that minimize the texture residual errors between the rendered model image and the input. However, some limitations need to overcome, for example, the convergence time [18]. Simplified versions of 3DMM [19-23] usually use a sparse set of shape models to reduce computational cost. These sparse

models are constructed by statistically learning a set of 3D facial feature points, indicating salient features of a face, and in this way, a large number of dissimilar faces are ignored. The state of the art [24] proposed to reconstruct the face shape in real-time by separately fitting rigid and non-rigid part of the face. However, density of the reconstructed 3D point cloud is the key issue of this method, hindering its way to dense detailed reconstruction. In fact, the morphable model based methods all lack the ability to describe rich geometric details of the surface, and some even require frontal view of the face with homogeneous illumination and neutral expression [22].

2.2. Forensic image based approach

An additional profile face usually helps to reconstruct more accurate models in the 2-view-based method. Ref. [25] obtained the 3D shape by deforming a generic model in accordance with the extracted facial features, from both frontal and profile face images. Ref. [26] first reconstructed an initial 3D face from the frontal image, and then depth of the profile face is used to refine the result. Nonetheless, notable deformations can be captured especially when the reconstructed model is rotated under large view point changes.

The forensic mugshot database usually contains one frontal and two profile face images. In this experimental settings, Ref. [27] utilized sparse bundle adjustment to reconstruct 3D landmarks from multi-view images, which are further used to deform a generic 3D face model to the final shape. To better explore multi-view constraint on texture images, Ref. [28] reconstructed 3D face model from 5 face images with approximately 45° apart. Their method followed the pipeline of multi-view stereo, by first calibrating the cameras through feature matching, and then obtaining dense face reconstruction based on voxels. However, as face images are textureless, the calibration process may fail or inaccurate calibration results may be obtained that may severely influence the final shape.

3. Problem statement

In this paper we aim at solving a particular multi-view 3D face reconstruction problem in the presence of multiple reference models. Input of our algorithm contains three face images $\Psi = \{lp, f, rp\}$ of the same person, including one frontal and two profile images that are captured for forensic mugshot databases, and our goal is to recover pointcloud-based 3D model of that person's face. Framework of our proposed algorithm is illustrated in Fig. 2.

Theoretically, any novel face model can be synthesized from a database covering infinite number of faces, by assembling consistent facial parts to the final model. However such database is impractical and is not accessible experimentally. Fortunately, an optimal approximation can be obtained when 1) a large set of reference models is given, and 2) these prior models are of large variety of shapes.

Suppose the reference model pool is denoted as $\{D_l^r, l = 1, 2, \ldots, K\}$, where D_l^r is the l-th reference model that is collected using modern laser scanners. Examples of the reference modes are shown in Fig. 3. Each reference model D_l^r is disassembled into irregular parts $D_l^r = \bigcup_k S_{k,l}^r$, so the final estimation can be obtained by combining these parts $X = \bigcup_{k,l} S_{k,l}^r$. Here k is index for each facial part and l is index of each model.

4. Proposed method

Consistent facial part is defined by measuring coherency between the output 3D face model and the input face images. In our formulation, the coherency is measured by exploring shading information of the input image, multi-view constraint between different view points, as well as smoothness shape priors of the surface. On basis of the three cues, we propose to solve the reconstruction problem by minimizing the following energy function

$$E = E_{shad} + \lambda_m \cdot E_{mview} + \lambda_s \cdot E_{smooth}. \quad (1)$$

The shading term E_{shad} ensures that our desired estimation be consistent with the input images using shading information by solving the Irradiance Equation. The second term E_{mview} is a large distortion and occlusion based multi-view constraint, imposing color consistency between two adjacent views. And the uncertainty based depth smoothness term E_{smooth} penalizes depth discontinuity between neighboring pixels. The parameters λ_m and λ_s are weighting parameters controlling importance of each term.

4.1. Energy function

4.1.1. Shading term E_{shad}

The shading term utilizes SFS theory, where 3D reconstruction problem can be solved using shading information of the input image.

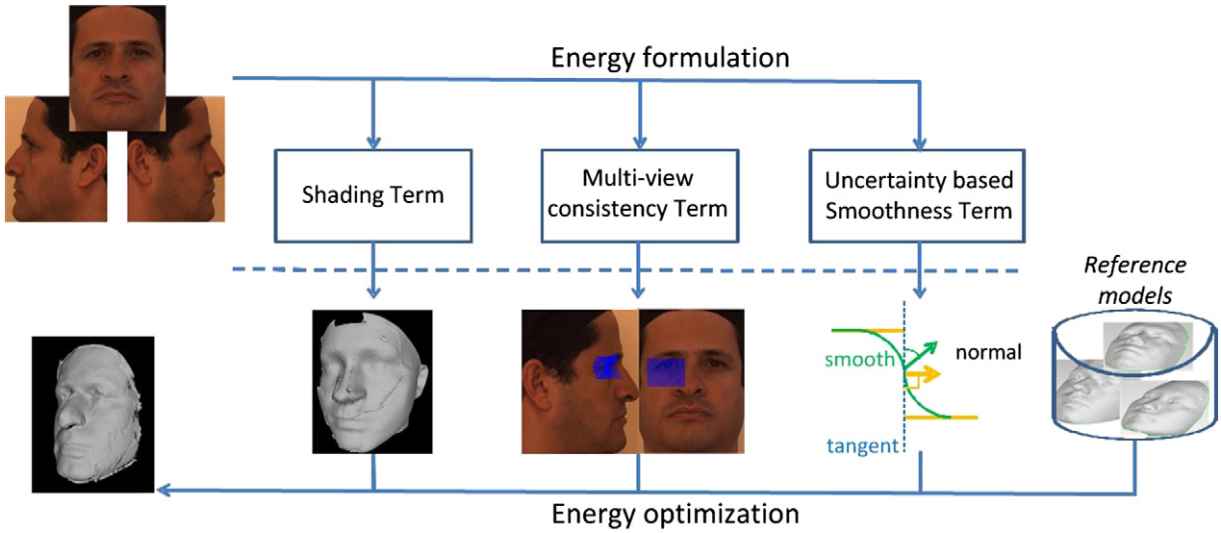

Fig. 2. Framework of our algorithm.

Fig. 3. Example of the reference models selected from BU3D (left) and Bosphorus (right) database.

Usually, we represent the irradiance equation of a Lambertian surface using $I = \rho \overrightarrow{l}\,\overrightarrow{n}$. Here $\overrightarrow{l}$ is lighting coefficient representing the direction and intensity of a single point light source placed at infinity, n is the surface normal and ρ surface albedo. Consider a face image $I(x,y)$, its corresponding surface is given by $z(x,y)$, then surface normal at point (x,y) is denoted as

$$\overrightarrow{n}(x,y) = \frac{1}{\sqrt{p^2+q^2+1}}(p,q,-1)^T \tag{2}$$

where $p(x,y) = \partial z/\partial x$ and $q(x,y) = \partial z/\partial y$.

The above theory motivates us to design the shading term as

$$E_{shad} = \sum_{v\in\Psi}\sum_{i\in\Gamma} |I(\pi_v^{X_i}) - \rho(\pi_v^{X_i})\overrightarrow{l}_v n(X_i)|. \tag{3}$$

Here Ψ is the set of input images as mentioned in Section 3 and Γ is the set of 3D point indexes within the face model. The symbol $n(X_i)$ denotes surface normal centered at point $X_i = (x,y,z)$. The subscript v in Eq. (3) represents each of the view point, thus $\pi_v^{X_i}$ denotes the corresponding 2D projection on view v with respect to X_i. Suppose P_v is the projection matrix to view v, we have

$$\pi_v^{X_i} = P_v X_i. \tag{4}$$

The shading term encourages the final estimation X_i be consistent with shading information in each of the three views, by simultaneously estimating albedo and lighting.

4.1.2. Multi-view consistency term E_{mview}

In the shading term we explore shading properties of each input image, but the multi-view constraint hasn't been fully considered. The view-point difference provides additional information by using epipolar geometry, and helps further refine the final shape. Our multi-view color consistency term is designed to measure color consistency between pairs of images. This term states that, the projections from the same 3D point onto different views should have similar appearance. However, textureless property of natural face image makes per-pixel similarity measurement among different views ambiguity and insufficient. One possible way to address this problem is to employ local image patches instead of single pixels for feature representation. Thus, for patch R_i centered at point i in the frontal view, we can easily find its corresponding pixel j in the profile view along with its corresponding patch R_j centered at j. However, in fact, R_i and R_j are not semantically matched pixel-by-pixel, mainly due to the existence of large distortion and self-occlusion between frontal and profile views. In this paper, we don't directly compare two square patches, but find correspondence between two compact sets in the frontal and profile view. In this scene, it is possible that two or more pixels in the frontal view match with the same counterpart on the profile view, and verse vice.

Suppose $C_i = \{X_i^1, \ldots, X_i^M\}$ is a compact set of 3D points centered at i, we define the multi-view color consistency term as

$$E_{mview} = \sum_{(u,v)}\sum_{i\in\Gamma}\sum_{m=1}^{M} \| I(\pi_u^{X_i^m}) - I(\pi_v^{X_i^m})\|_2. \tag{5}$$

Here (u,v) denotes a pair of neighboring views. In our special case, the two profile views are not neighbors because they share no common regions. Note that in Eq. (5) the extracted compact regions from u and v are semantically registered.

4.1.3. Smoothness term E_{smooth}

The depth smoothness term ensures smooth transition in depth and penalizes sharp depth edges. The smoothness prior is reasonable because human faces can always be described using smooth surfaces. We define the depth smoothness term as

$$E_{smooth} = w_{ij} \sum_{(i,j)\in\mathbf{N}} \| X_i - X_j\|_2. \tag{6}$$

Here $\mathbf{N}$ is the set of neighboring points. The weighting parameter w_{ij} describes uncertainty of the depth guesses and is further defined as

$$w_{ij} = \frac{w_i + w_j}{2}. \tag{7}$$

The smoothness term states that geometry of neighboring points should change smoothly, especially when large uncertainty occurs on depth estimation X_i and X_j. Meanwhile, we tend to relax the smoothness constraint when depth estimation is accurate. Uncertainty of the depth guess is measured to enforce geometric constraint on the energy, by assuming that surface normal and its tangent should be perpendicular. Thus for each point i, we define uncertainty w_i as

$$w_i = |Nor(X_i)\cdot Tan(X_i)| \tag{8}$$

where normal $Nor(X_i)$ and tangent vector $Tan(X_i)$ of point X_i can be estimated using universally acknowledged equations.

4.2. Energy minimization

There are several unknowns in the proposed energy in Section 4.1, such as the 3D face shape, lighting parameter and albedo. To make this problem well-posed, we use an approximate approach by first estimating an initial model from the shading term and then reformulating the proposed energy into a discrete multi-label segmentation problem.

4.2.1. Initial model estimation

We would first estimate albedo, lighting parameters and an initial 3D face model according to the shading term base on the assumption that the face is Lambertian with albedo while ignoring the effect of cast shadows and inter-reflections. Motivated by the previous

works [5,6] that are able to recover depth from a single face image in the presence of a single reference model, we estimate shape of the face model by iteratively optimizing light, albedo and shape. However, Refs. [5,6] focused mainly on the frontal view. In this work, we modify this algorithm to adapt both frontal and profile views. For profile images, we first rotate the reference model to the profile view and then go through the pipeline of depthmap estimation.

For each input view, a depthmap corresponding to that view is estimated. To merge these depthmaps into a full model, we assume rigid transformation between different view models. We further make a rough assumption that by rotating 90° around the vertical line, thus the 3D model of the profile view is in accordance with the frontal view. Thus, given landmarks commonly seen in both frontal and profile views, *e.g.* eye corners, we are able to merge an initial full 3D model by first rotating the profile model to the frontal view, and then estimate the translation *via* corresponding landmarks.

4.2.2. Problem reformulation

To make the problem well-solved, we convert the continuous minimization problem into a discrete minimization one, by formulating the assembly process as a multi-label segmentation problem. In our formulation, labels refer to index of multiple candidate shape priors, and correspond to 3D point in space.

An external 3D face database is a realistic and meaningful shape prior that can help regulate and refine the final geometry. To make the assembly process meaningful, the prior database should first be registered. We denote the training database as $\{D_l^r, F_l^r\}$, where D_l^r is the l-th 3D face model and F_l^r is its predefined landmark points. With the known initial face model D and its corresponding landmarks F, we register candidate of the l-th reference model D_l^r to D*via*

$$D_l = T_{F_l^r,F} \cdot S_{F_l^r,F} \cdot D_l^r \tag{9}$$

with D_l the registered shape. The symbol $S_{F_l^r,F}$ is a scaling matrix that makes the candidate model be of the same size as the initial one, and $T_{F_l^r,F}$ denotes the rigid transformation matrix including both rotation and translation. Note that the semantic ordering of F_l^r and F should be the same.

With the estimated initial model D and the registered prior models $\{D_l\}$, we reformulate the discrete energy as

$$E(l) = E_{shad}(l) + \lambda_m \cdot E_{mview}(l) + \lambda_s \cdot E_{smooth}(l) \tag{10}$$

where the discrete label l refers to each of the candidate model.

Discrete Shading Term E_{shad}. The discrete shading term encourages that the final estimation be consistent to the initial model, thus penalizes dissimilarity between the estimated output and the initial guess. Suppose D_{l_i} is depth of point i in the l-th candidate model, we rewrite this term as

$$E_{shad}(l) = \sum_{i\in\Gamma} |D_{l_i} - D_i| \tag{11}$$

where D_i denotes coordinate of 3D point i in the initial model.

Discrete Multi-view Consistency Term E_{mview}. We rewrite the multi-view consistency term according to Eq. (5)

$$E_{mview}(l) = \sum_{(u,v)} \sum_{i\in\Gamma} \sum_{m=1}^{M} \| I(\pi_u^{D_{l_i}^m}) - I(\pi_v^{D_{l_i}^m}) \|_2 \tag{12}$$

and $D_{l_i}^m$ is depth of the m-th point in the l_i-th candidate model, within the compact set centered at i.

Discrete Smoothness Term E_{smooth}. For the discrete smoothness term, we rewrite Eqn. 6 as

$$E_{smooth}(l) = \frac{w_{l_i} + w_{l_j}}{2} \sum_{(i,j)} \| D_{l_i} - D_{l_j} \|_2. \tag{13}$$

The corresponding weighting parameter w_{l_i} is related to normal and tangent of the surface, which are estimated using higher-order constraint. To make this term available pair-wisely, we approximate w_{l_i} within a 4-neighbor lattice network that

$$w_{l_i} = \sum_{j\in N_i} |Nor(D_{l_i}) \cdot Tan(D_{l_i}, D_{l_j})|. \tag{14}$$

Here j is within i's neighborhood $\mathbf{N}_i$ and the tangent vector of D_{l_i} is estimated *via* its neighboring point D_{l_j} that

$$Tan(D_{l_i}, D_{l_j}) = D_{l_j} - D_{l_i}. \tag{15}$$

As pair-wise constraint is not sufficient to estimate normal, we approximate $Nor(D_{l_i})$ as

$$Nor(D_{l_i}) = (D_{l_i^{x+1}} - D_{l_i^{x-1}}) \times (D_{l_i^{y+1}} - D_{l_i^{y-1}}). \tag{16}$$

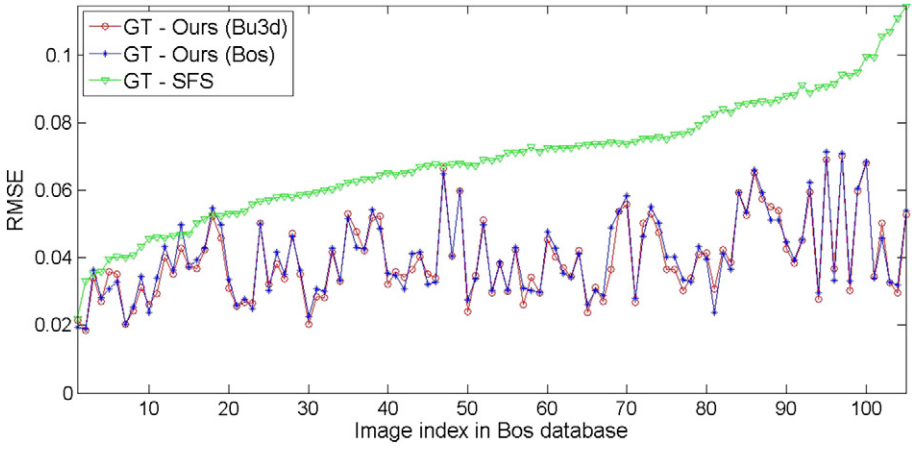

Fig. 4. RMSE calculated over each of the 105 subjects from Bos database. The green square shows error between SFS and the groundtruth. Our reconstruction error is illustrated in blue star and red circle, each of which corresponds to different reference pools, namely the Bos pool and the Bu3d pool respectively.

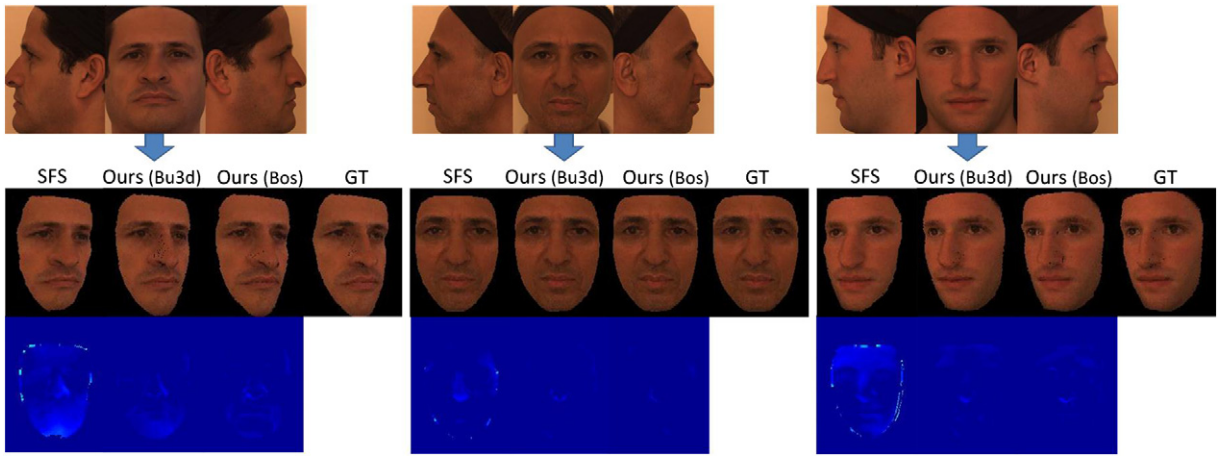

Fig. 5. From left to right, three examples of the reconstructed models from Bos database. The first row is the input images, the second row shows reconstructed results using different methods, and the last row shows error maps of the corresponding models in the second row. For each example, the models from left to right come from: SFS method, our method using Bu3d pool, our method using Bos pool, and the groundtruth model (GT).

Suppose D_{l_i} denotes depth of the l-th candidate at pixel i, the depth of its right neighbor along x direction is denoted as $D_{l_i^{x+1}}$ and $D_{l_i^{y+1}}$ is depth of its lower neighbor along y direction. Thus, normal estimation is independent of neighboring point j, but searches for the local neighborhood centered at i.

4.2.3. Optimization algorithm

With the proposed algorithm we have converted the 3D face reconstruction problem into a MRF labeling one, and solve the discrete energy function in Eq. (10) that combines different cues and constraints in order to achieve the optimization goal. To simplify the optimization process, the dimensional of the solution space is reduced from 3D to 2D in order to use mature optimization algorithms in image processing. To achieve this goal, we would first compute the corresponding 2D depthmaps from these registered 3D models.

Pointcloud–Depthmap Convention. Similar to [28], we sample the 3D point in a cylindrical coordinate system which yields a compact representation of the 3D surface, and represent geometry of the face using depth image d. Suppose a Cartesian coordinate system on a 3D face whose original point $O = (C_x, C_y, C_z)$ is at the center of the face model, with its X–Y coordinate parallel to horizontal and vertical lines, and the Z axis points frontward. In the converted depthmap, the value $d = d(x,y)$ at pixel (x,y) denotes depth of the corresponding 3D point $\mathbf{X}_{\theta,\phi,d}$

$$\boldsymbol{X}_{\theta,\phi,d} = (d \cdot sin\theta + C_x, \phi + C_y, d \cdot cos\theta + C_z) \tag{17}$$

with

$$x = k_x \cdot \theta * 180/\pi, y = k_y \cdot \phi. \tag{18}$$

Here θ is the angle between $O\mathbf{X}_{\theta,\phi,d}$ and the Z axis, and ϕ denotes point projection along Y axis in the Cartesian coordinate. Symbols k_x and k_y are parameters controlling density of the depthmap.

Graph-cuts Optimization. We choose to use graph-cuts to minimize the proposed energy function in Eq. (10) because: 1) the max-flow-based optimization algorithms are proven to achieve a global minimum solution and meanwhile, 2) their complexities remain in the order of polynomial time in terms of the number of the underlying graph nodes and edges. In this paper we use α-expansion to solve the converted multi-label segmentation problem.

5. Experimental results

In this section we describe our experiments. We evaluate our method from two different perspectives. We first show qualitative

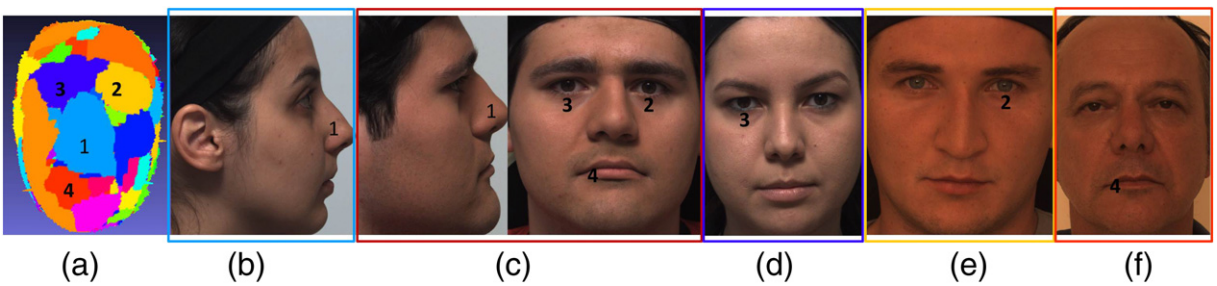

Fig. 6. (a) Example of the final labeling results. Semantic part comes from the same candidate model that is encoded in the same color. (c) Input face images. The other images are final selected faces from the reference pool. The facial parts marked '1' at the nose area in (b), '2' around the right in (e), '3' around the left eye in (d) and '4' mouse area in (f) are assembled to form the final model. The assembled areas are visually similar in both appearance and shape.

Fig. 7. From left to right, three examples of the reconstructed models from Bos database. For each set of examples, the first row shows the input frontal and profile face images. The following rows show untextured mesh from: SFS method, our method using Bu3d pool, our method from Bos pool, and the groundtruth scanning.

and quantitative results of the reconstructed shape. We then show performance of our results in 3D face recognition.

We evaluate our proposed method on two different datasets. The constraint Bosphorus (Bos) dataset[1] contains a total of 105 subjects with different poses, including one frontal (0°) and two profile ($\pm 90^\circ$) face images for each subject. Bos also provides the groundtruth models that are captured *via* laser scanner. The unconstraint Colorferet [29] dataset is much more challenge, because 1) the input frontal face is not exactly 0° facing forward, and 2) the input profile faces are not 90° rotation from the frontal one. Furthermore, Colorferet also contains images with different expressions. To show effectiveness of our algorithm, we manually selected 349 subjects from Colorferet for evaluation. The selected subjects are of near neutral expression, near frontal, and the input profile faces are facing near 90°.

Unless otherwise indicated, all experiments were run with the same parameters. We manually set the parameters in Eq 1 as $\lambda_m = 2$ and $\lambda_s = 30$. The 3D compact region used in multi-view color consistency term is set to be 7×7 in the cylindrical coordinate system (*e.g.* $M = 49$). We implemented the proposed algorithm using C++ on a 64-bit windows workstation with Intel *i*5 CPU and 4GB memory.

5.1. 3D face reconstruction evaluation

We first ran our method on Bos dataset. To show that our method is not sensitive to the prior reference models, we choose two sets of references for comparison (Fig. 3). The Bu3d database[2] contains 100 subjects, ranging races from White, Black, Indian, East Asian, Middle-east Asian to Lation-Hispanic, with 44 males and 56 females. The other reference dataset is Bos dataset itself. For each subject be reconstructed, it is automatically withdrawed from the reference pool, and only the other 104 models are used as references. The 3D face models in the reference pool are first registered to the initial reconstruction, and served as depth candidates for further optimization.

In our experiment, resolution of the frontal view is 205×181, and 205×150 for profile views. On average, the computational cost for each subject is about 30 s. For quantitative evaluation, we show accuracy of our algorithm in terms of Rooted Mean Square Error (RMSE). Suppose X^{es} is the estimated depthmap, and X^{gt} is the corresponding grondtruth. The RMSE for each model is computed as

$$RMSE = \frac{\sqrt{\sum_i \| X_i^{es} - X_i^{gt} \|_2}}{N} \tag{19}$$

where i is index of each pixel and N the total number of valid pixels in use.

Fig. 4 shows the overall mean reconstruction error for each of the model in Bos. We compare our method with SFS [6]. Ref. [6] reconstructed a 3D face model from a single image by employing a single reference face model. Ours is different in two perspectives: 1) our method uses both frontal and profile images as input, which provides more information compared with the single image setting, so ours can produce more accurate reconstruction; 2) our method utilize multiple reference models, which is much less model-reliant compared with the single reference model approach. In general, our reconstruction error is much smaller than that form SFS. We also show that our reconstruction error from Bos pool is comparable to that from Bu3d pool, demonstrating that our method is not sensitive

[1] http://bosphorus.ee.boun.edu.tr/Home.aspx
[2] The Bu3d data comes from State University of New York at Binghamton.

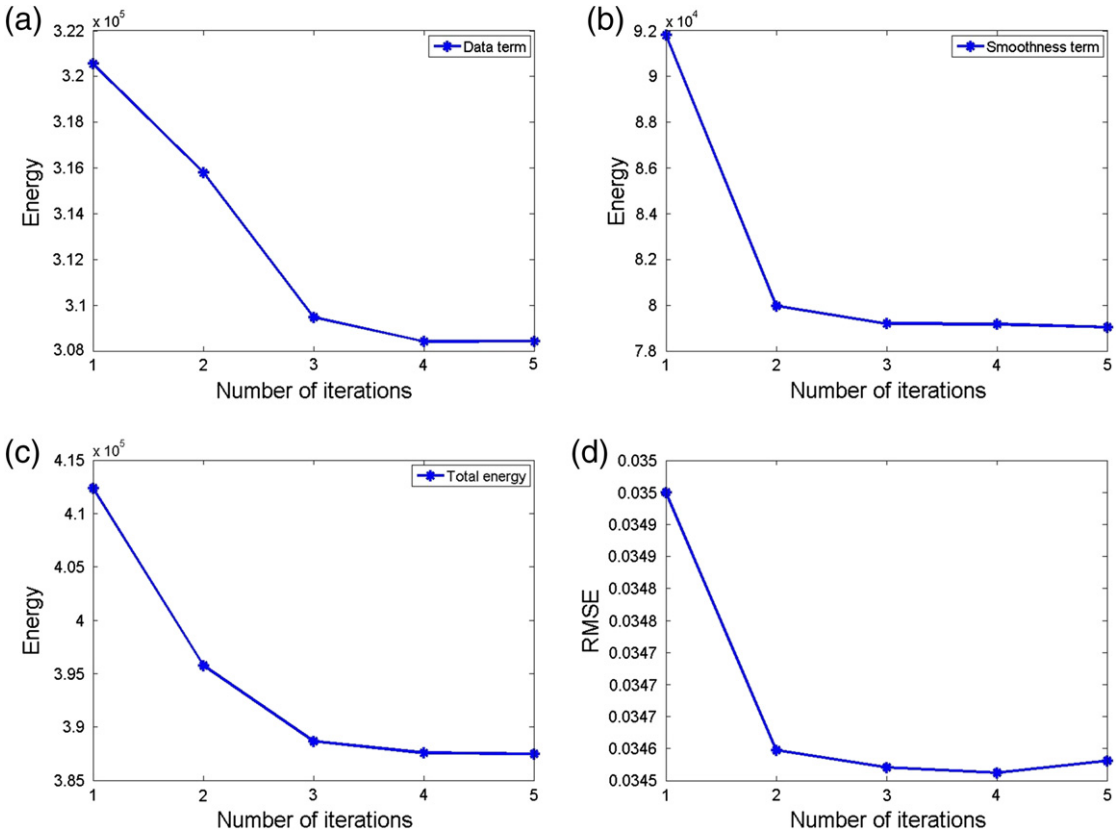

Fig. 8. Energy distribution versus number of iterations with respect to (a) data term (b) smoothness term and (c) total energy. (d) shows error distribution as number of iteration increases.

to different reference model sets, thanks to our examplar coherent pipeline.

Fig. 5 shows examples of reconstructed shape compared with the prior art [6] rendered under different view points. The error map is calculated using absolute difference between the estimated model and the groundtruth. We show that our optimization pipeline can correct large depth errors, both on the surface and at depth boundary.

Fig. 7 shows more rendering results from new synthesized viewpoint, where the results from SFS [6] are oversmoothed. Meanwhile, our method can preserve more geometric details, thus producing more visually pleasant results. In comparison with our method using different reference models, we find that the reconstruction from Bos pool can produce more detailed geometry (*e.g.* around the eye, the mouse). This is because the reference models in Bos contain more details compared with that from Bu3d (Fig. 3). Note that our example-based pipeline searches for the most coherent patchwork from the reference pool and then assembles them. So it is reasonable that detailed reference pool produces more detailed reconstruction. Fig. 6 shows the final label distribution that are encoded in color. Although we don't aim at semantic segmentation on the face, the labeling results can provide some clue that coherent parts from different persons are usually semantically distributed. Usually, dozens of faces are enough for synthesizing a desired face.

To show effectiveness of using the discrete optimization algorithm, Fig. 8 shows one sample on how energy and error vary as the number of iteration grows. Fig. 8 (a), (b), and (c) shows energy decline with respect to the data term, smoothness term, as well as the total energy, experimentally demonstrating that optimizing our energy function using α-expansion is possible. Fig. 8 (d) further shows how the reconstruction error declines during iteration, which on the other hand illustrating correctness and effectiveness of the optimization algorithm in use. Fig. 9 provides textured shapes in each of the iteration in Fig. 8. The visual artifact reduces significantly after the first iteration, which is in accordance with error distribution in Fig. 8 (d).

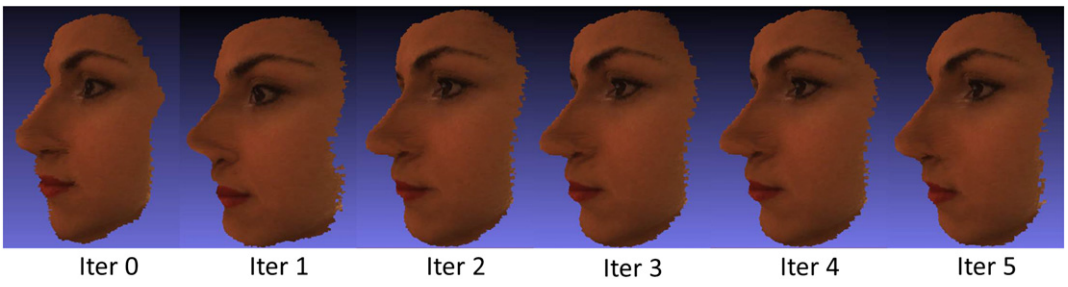

Fig. 9. The shapes reconstructed in each iteration. The initial model is shown as the 0th iteration.

Fig. 10. Different input conditions from the constraint Bos dataset (left) and the unconstraint Colorferet dataset (right). For each dataset, from left to right shows right profile image, left profile image, initial depthmap and optimized depthmap in cylindrical coordinate system. The final estimation is not sensitive to inaccurate profile pose and the initial model.

We also test our algorithm on the Colorferet database, which is much more challenging because it is less controlled and exhibit more variations, *i.e.* it violates the assumption that 3D model of the profile view can be obtained by rotating 90° around the vertical line. In our experiment, the profile views are usually $[70, 90]^\circ$ facing left or right. Fig. 10 shows the initial SFS condition used in different datasets. The rough initial model originates from inaccurate input poses. However, our method can still improve quality of the final model (Fig. 10 (d)) under such unconstraint environment. Fig. 11 shows more results with different synthesized views from Colorferet. In such tough conditions, our algorithm can also produce visually pleasant results, demonstrating robustness of our proposed method.

5.2. 3D face recognition evaluation

One of the most important applications using 3D face models is 3D face recognition. To evaluate performance of our reconstructed model, we test it in our newly designed 3D face recognition system, and show its superiority in terms of recognition rate.

3D models can help solve pose variation problem that commonly exists in 2D face recognition. For instance, when the query face is facing 60° left, but the gallery contains only frontal faces. With 3D face model in the gallery, we can generate 2D face images of that person facing exactly 60°, and compare it directly with the input image. In such scenario, 3D face models can be used to enlarge diversity of enrolled gallery in order to improve performance of 2D face recognition under various poses. However, generating the view-dependent face images can be costly both on memory space and computational complexity. Consequently, in this experiment our 3D-aided face recognition pipeline focuses on locating corresponded semantic parts across different poses *via* 3D models. The gallery contains one frontal and two profile face images, as well as their 3D face models. And we aim at recognizing a face image facing arbitrary view.

In our experiment, only 7 landmarks are used, they are two eye centers, two eyebrow centers, nose root, nose tip and the mouth center. The 7 predefined landmarks are first located on both the 2D face images and the corresponding 3D face models. Note that the 2D and 3D landmarks are semantically corresponded within the same identity. The 3D semantic patches centered at 3D landmarks are then projected on both frontal and profile images. These projected 2D patches form the basis of gallery features. Given a face image of an arbitrary pose, these semantic patches from the input image can also be extracted *via* landmark correspondences. As a result, the cross pose face recognition problem is converted into a semantic patches feature matching problem. Similarity measure between semantic patches in the gallery and those of the input is calculated based on local binary patterns (LBP) descriptor. We compare the features extracted in the enrolled images and that of the probe, and recognition result is obtained using nearest-neighbor classifier.

We conduct our experiment on Bos. For the gallery, we choose 3 different poses for evaluation, including facing up, right up and right down. The results are shown in Fig. 12. For each pose, the recognition rate is calculated over all the 105 samples. Our reconstructed model helps boost recognition rate compared with SFS, because ours can produce more accurate results. With the aforementioned recognition method, the corresponding patches between different poses are semantically coherent when the underlying 3D model being precisely estimated. In this perspective, it is reasonable that our model produces much higher recognition rate.

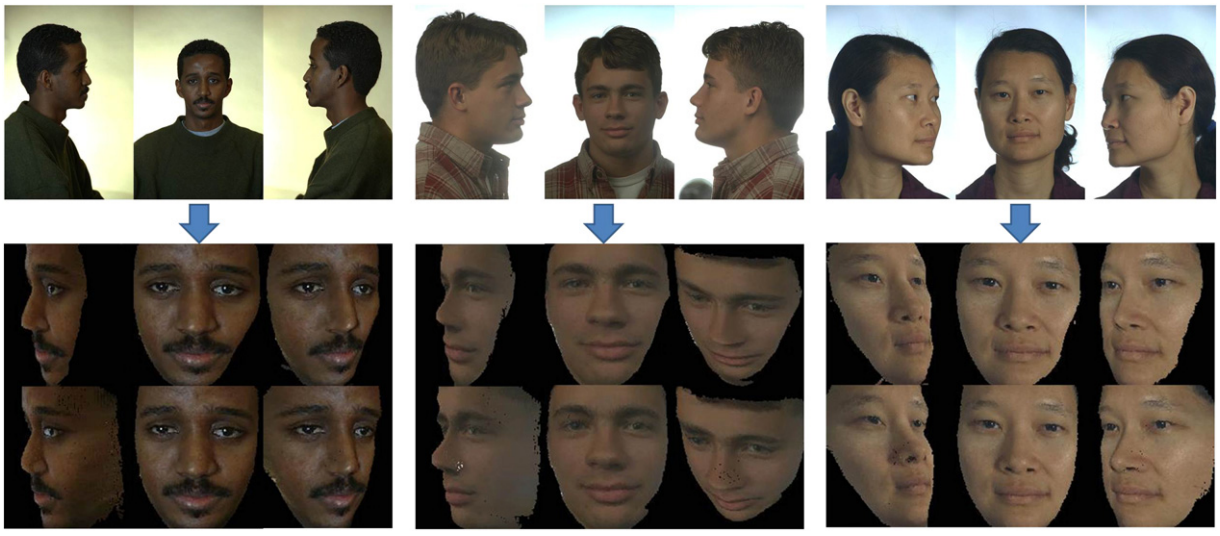

Fig. 11. Three example of the reconstruction from Colorferet database. For each subject, the first row shows input images, the second and third rows show reconstructed results using SFS and our method respectively.

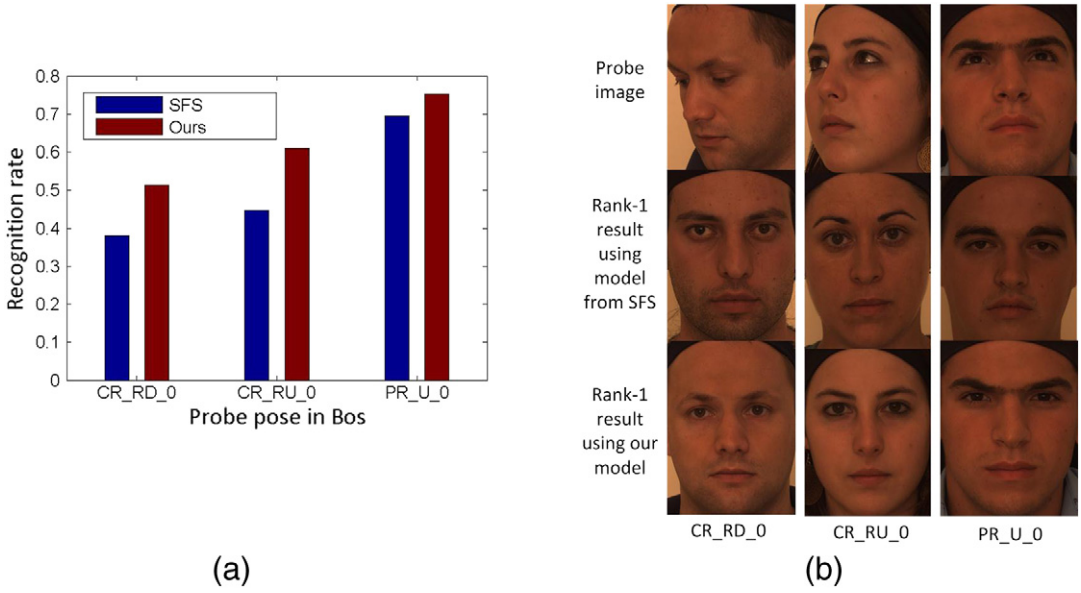

Fig. 12. (a) Recognition rate in Bos. The probe poses we use are: right down (CR_RD_0), right up (CR_RU_0) and up (PR_U_0). (b) We visualize cases where Rank-1 hits using our model while using that from SFS fails. For each sample from left to right, SFS hits at Rank-6, Rank-5 and Rank-9, respectively.

We also test our reconstructed model on Colorferet in terms of face recognition. We manually select 200 samples in Colorferet with neutral expression and without occlusion. We choose face images facing 67.5° right as query, and our reconstructed model obtains recognition rate of 76.5%, 5% boosts compared with that from SFS.

6. Conclusion

In this paper we proposed an examplar coherent method for 3D face reconstruction from forensic mugshot database. This problem is challenging due to the uncalibrated input images with wide baseline. As face images are textureless, the traditional multi-view stereo pipeline could not work. We address this by using an external face database and generating the result through facial part composition. We proposed an energy function to formulate the 3D face reconstruction problem using cues from 1) shape from shading 2) multi-view color consistency and 3) depth smoothness prior, and solved it by first estimating shading parameters and then converting it to multi-view image segmentation problem. Quantitative and qualitative evaluation results show the effectiveness of our algorithm in reconstructing accurate 3D face models. Additional supportive experiment on face recognition further shows that our method is capable to produce accurate face models and enhance face recognition accuracy. The work can be extended to 3D face reconstruction from multiple arbitrary views, with less computational cost.

Acknowledgment

This work is supported by National Natural Science Foundation of China 61202161, and National Key Scientific Instrument and Equipment Development Projects 2013YQ49087904.

References

[1] J. Heo, M. Savvides, Gender and ethnicity specific generic elastic models from a single 2d image for novel 2d pose face synthesis and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (12) (2012) 2341–2350.
[2] R. Zhang, P.-S. Tsai, James Edwin. Cryer, Mubarak. Shah, Shape-from-shading: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 21 (8) (1999) 690–706.
[3] R. Dovgard, R. Basri, Statistical symmetric shape from shading for 3D structure recovery of faces, Computer Vision-ECCV 2004, Springer 2004, pp. 99–113.
[4] A. Ahmed, A. Farag, T. Starr, A new symmetric shape from shading algorithm with an application to 3-D face reconstruction, Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, IEEE, 2008. pp. 201–204.
[5] I. Kemelmacher, R. Basri, Molding face shapes by example, Computer Vision-ECCV 2006, Springer 2006, pp. 277–288.
[6] I. Kemelmacher-Shlizerman, R. Basri, 3d face reconstruction from a single image using a single reference face shape, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 394–405.
[7] J. Roth, Y. Tong, X. Liu, Unconstrained 3D face reconstruction, Trans. Graph 33 (4) (2014) 43
[8] M. Castelán, W.A. Smith, E.R. Hancock, A coupled statistical model for face shape recovery from brightness images, IEEE Trans. Image Process. 16 (4) (2007) 1139–1151.
[9] M. Castelan, J. Van Horebeek, Relating intensities with three-dimensional facial shape using partial least squares, Computer Vision, IET 3 (2) (2009) 60–73.
[10] M. Reiter, R. Donner, G. Langs, H. Bischof, 3D and infrared face reconstruction from RGB data using canonical correlation analysis, Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, 1, IEEE 2006, pp. 425–428.
[11] Z. Lei, Q. Bai, R. He, S.Z. Li, Face shape recovery from a single image using CCA mapping between tensor spaces, Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE 2008, pp. 1–7.
[12] J. Gonzalez-Mora, F. De la Torre, N. Guil, E.L. Zapata, Learning a generic 3D face model from 2D image databases using incremental structure-from-motion, Image Vis. Comput. 28 (7) (2010) 1117–1129.
[13] T. Hassner, Viewing real-world faces in 3D, Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE 2013, pp. 3607–3614.
[14] S. Suwajanakorn, I. Kemelmacher-Shlizerman, S.M. Seitz, Total moving face reconstruction, Computer Vision-ECCV 2014, Springer 2014, pp. 796–812.
[15] Ira. Kemelmacher-Shlizerman, Steven M. Seitz, Face reconstruction in the wild, Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE 2011, pp. 1746–1753.
[16] Volker. Blanz, Thomas. Vetter, Face recognition based on fitting a 3D morphable model, IEEE Trans. Pattern Anal. Mach. Intell. 25 (9) (2003) 1063–1074.
[17] Volker. Blanz, Thomas. Vetter, A morphable model for the synthesis of 3D faces, Proceedings of the 26th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co. 1999, pp. 187–194.
[18] X. Zhang, Y. Gao, Face recognition across pose: a review, Pattern Recogn. 42 (11) (2009) 2876–2896.
[19] U. Park, Y. Tong, A.K. Jain, Face recognition with temporal invariance: a 3d aging model, Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, IEEE 2008, pp. 1–7.
[20] U. Park, Y. Tong, A.K. Jain, Age-invariant face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (5) (2010) 947–954.
[21] C. Wang, S. Yan, H. Li, H. Zhang, M. Li, Automatic, effective, and efficient 3D face reconstruction from arbitrary view image, Advances in Multimedia Information Processing-PCM 2004 Springer 2005, pp. 553–560.
[22] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, W. Gao, Efficient 3D reconstruction for face recognition, Pattern Recogn. 38 (6) (2005) 787–798.
[23] M. Pamplona Segundo, Luciano. Silva, Olga Regina Pereira. Bellon, Improving 3D face reconstruction from a single image using half-frontal face poses, Image

Processing (ICIP), 2012 19th IEEE International Conference on, IEEE 2012, pp. 1797–1800.
[24] L.A. Jeni, J.F. Cohn, T. Kanade, Dense 3D Face Alignment From 2D Videos in Real-Time, 2015.
[25] W.B. Lee, M.H. Lee, I.K. Park, Photorealistic 3D face modeling on a smartphone, Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, IEEE 2011, pp. 163–168.
[26] H. Han, A.K. Jain, 3D face texture modeling from uncalibrated frontal and profile images, Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on, IEEE 2012, pp. 223–230.
[27] J. Choi, G. Medioni, Y. Lin, L. Silva, O. Regina, M. Pamplona, T.C. Faltemier, 3d face reconstruction using a single or multiple views, Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE 2010, pp. 3959–3962.
[28] Y. Lin, G. Medioni, J. Choi, Accurate 3D face reconstruction from weakly calibrated wide baseline images with profile contours, Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE 2010, pp. 1490–1497.
[29] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.