Example-Based 3D Face Reconstruction from Uncalibrated Frontal and Profile Images

Jing Li, Shuqin Long, Dan Zeng, Qijun Zhao
College of Computer Science, Sichuan University
Chengdu, 610065, China

Abstract

Reconstructing 3D face models from multiple uncalibrated 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 using faces. Consequently, existing 3D reference model based methods have limited capability of reconstructing 3D face models for a large variety of persons. In this paper, we propose to explore a reservoir of diverse reference models to improve the 3D face reconstruction performance. Specifically, we convert the face reconstruction problem into a multi-label segmentation problem. Its energy function is formulated from different cues, including 1) similarity between the desired output and the initial model, 2) color consistency between different views, 3) smoothness constraint on adjacent pixels, and 4) model consistency within local neighborhood. Experimental results on challenging datasets demonstrate that the proposed algorithm is capable of recovering high quality face models in both qualitative and quantitative evaluations.

1. Introduction

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 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 frontal and profile 2D face images widely exist in forensic databases and are routinely used by police officers.

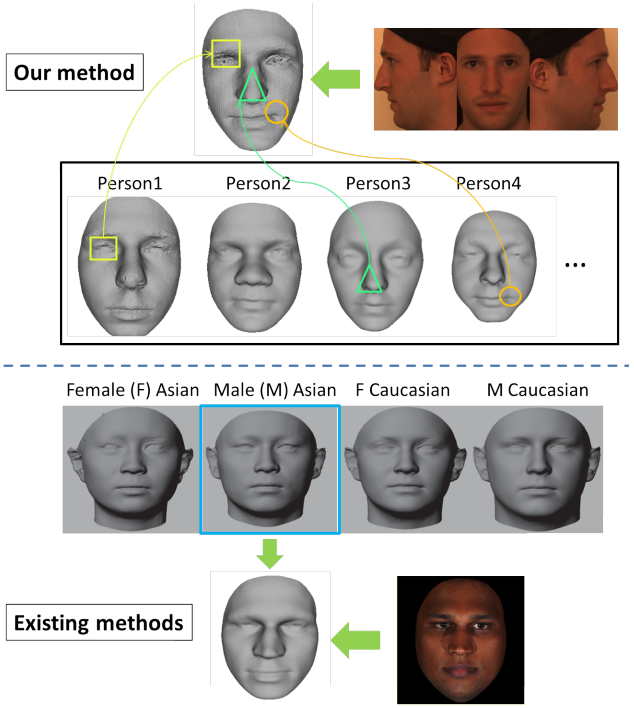


Figure 1. Existing methods utilize one single reference model to reconstruct 3D face model from the input 2D image. The reference model could be a generic model or a gender/ethnicity specific average model [7]. Instead, our proposed example-based method uses multiple reference models, each for some component in the input face. This way, our method can obtain more accurate reconstructed 3D face models.

A number of 3D face reconstruction methods have been proposed in the literature. Most of them require some reference 3D face models. This is because prior knowledge about the scale of the face or the depth of the facial components is missing in the input uncalibrated 2D face im-

ages. Reference 3D face models are thus needed to serve as a constraint on the reconstructed 3D face model. Although impressive results have been reported by using existing reference model based methods, they have difficulties to accurately reconstruct the 3D face models for a large variety of persons because they use only a single reference model or a few gender/ethnicity specific reference models. But different persons, even those of the same gender or ethnicity, usually have considerably different faces, which forms the basis of person recognition using faces. Fortunately, it is possible that some persons may share some similar components in their faces, though their faces are different in overall appearance. This motivates us to propose an example-based 3D face reconstruction method as presented in this paper.

Unlike previous reference model based methods, our proposed method explores a reservoir of diverse reference 3D face models and search for each pixel in the input face an appropriate reference model from the reservoir (see Fig. 1). This way, it is of high probability for us to find the most alike reference for every component of the input face. We formulate the problem of searching for reference models and regularizing the reconstructed 3D face model into one unified optimization framework under Markov Random Field (MRF). This allows us to jointly optimize consistency between the desired output and the initial face model, color consistency between adjacent views, as well as depth and model smoothness within local neighborhoods. Our example-based method has several advantages: (i) Compared with single reference model approach, our method can largely reduce reliant on the reference itself; (ii) While expressing a novel face as a linear combination of 3D face database may suffer from fine detail reconstruction, our method is able to overcome this thanks to pixel-wise optimization.

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

2. Related Work

Existing work on 3D face reconstruction [10, 1, 8, 14, 3, 12, 6, 4] can be classified as either single-view approach that recovers the face shape from only one image or multiple-view approach where multiple face images are used as input.

Single-view Approach 3D Morphable Models (3DMM) [1] is a crucial and widely used model to estimate 3D shape of the face. Linear combination among different models are estimated in terms of both shape and texture. However, some limitations need to overcome, for example, the convergence time, the diversified 3D database and the correspondence establishment. Based on the ideology of 3DMM, Ref [8] proposed an analysis-by-synthesis framework for face recognition with variant pose, illumination and expression. For the reconstruction process, one input image is enough, which averted the heavy enrollment work, and they overcame the difficult problems like complex PIE in face recognition. Whereas, the input face must be frontal. Therefore, their reconstruction accuracy was sensitive to pose variation of the input image. To address this problem, Ref. [14] combined the fitting techniques and a sparse 3D deformable model to recover the 3D geometry.

Ref. [9, 10] recovered a 3D face model by introducing an arbitrary face model. This method was based on the observation that different faces look similar globally but vary considerably across individual in detail. The objective function was defined on basis of Structure from Shading (SfS) technology and was optimized iteratively. However, this method relied heavily on the reference face model. On the other hand, when the accurate 3D face model is not necessary, the single-model-based method is able to produce a visually pleasant result. For this purpose, Ref. [5] optimized the displacement between the reference model and the final estimation through joint depth-appearance similarity. To reduce the effect of arbitrary reference face model, Ref. [16] synthesized a reference model specifically to each person via photo collections of the same person [11].

Multi-view Approach Ref. [4] reconstructed a 3D face using only one frontal and a profile view. The frontal image was used for initialization and the profile for refinement. The reconstructed model was further tested on face recognition task across various poses. Nonetheless, notable deformation can be captured especially when the reconstructed model is under large view point changes. Ref. [2] utilized sparse bundle adjustment to reconstruct 3D landmarks, which are further used to deform a generic 3D face model. However, in these methods, multi-view constraint on the whole face is not fully explored.

Ref. [13] reconstruct 3D face model from 5 face images with approximately 45 degree apart. Their method followed the pipeline of multi-view stereo, by first calibrating the cameras through feature matching, and then obtaining dense face reconstruction base on voxels. However, as face images are lack of features, the calibration process may fail or errorness results may be obtained that can severely influence the final reconstruction result.

3. Problem Statement

Input of our algorithm contains 3 face images of a certain person, with one frontal view and two profile images that captured freely, and our goal is to recover pointcloud-based full 3D model of that person's face.

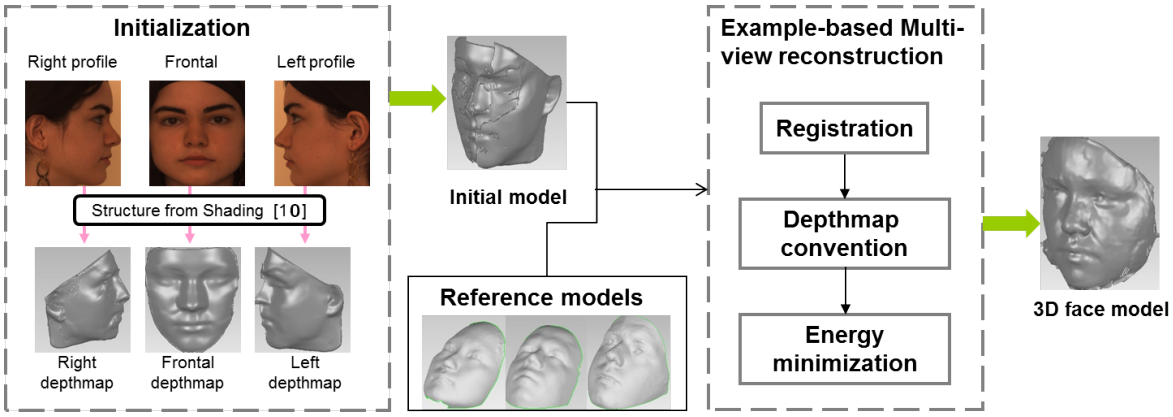


Figure 2. Framework of the proposed method.

Given the input images, we would first estimate depth map of each view. Base on the initial estimation, the proposed algorithm aims to find some candidate depths X^c of the 3D face model, and then the unknown depth X is estimated through global optimization. A framework of the proposed pipeline is illustrated in Fig. 2.

Initialization We would first estimate depth maps of each input view via SFS. This is motivated by the previous work [9, 10] that is able to recover depth from a single view. Given a $2D$ face image and a reference 3D face model, the initial estimation of the face model can be obtained by iteratively optimizing light, albedo and shape. However, Ref. [9, 10] focused mainly on frontal views. In this work, we modify their 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 depth map estimation base on the assumption that the face is Lambertian with albedo while ignoring the effect of cast shadows and inter-reflections.

To merge the aforementioned depth maps into an initial full model, we assume rigid transformation between different view models. We further make a rough assumption that the 3D model of the profile view can be obtained by rotating 90 degree around the vertical line. Thus, given landmarks commonly seen in both frontal and profile views, *e.g.* eye corner, we are able to merge an initial full 3D model by first rotating the profile model to the frontal view, and then estimate the transition via corresponding landmarks.

Candidate Calculation We formulate the 3D face reconstruction problem as a multiple label image segmentation problem. In order to make the reconstruction pipeline well defined, we employ prior shape knowledge to generate candidate depth base on the initial face model.

An external 3D face database is realistic and meaningful shape prior that can help to refine the final geometry, with less reliant on the initial model. Suppose $\{D_k, F_k\}, k = 1, ..., K$ is the training database , where D_k denotes the kth 3D face model and F_k predefined landmark points on D_k. Given the initial model D, with its corresponding landmarks F, candidate of the kth model $D_k^{register}$ is generated via

$$D_k^{register} = T(F_k, F) \cdot S(F_k, F) \cdot D_k \quad (1)$$

where $S(F_k, F)$ is a 3D face scaling process that makes the database model be of the same scale as the initial one, and $T(F_k, F)$ denotes the transformation matrix that can register the model in the database to the initial estimation. Note that the semantic ordering of F_k and F should be the same.

4. Energy Minimization

We propose to solve the above mentioned MRF labeling problem by minimizing the following energy functional

$$E = E_{data} + \lambda_c \cdot E_{color} + \lambda_d \cdot E_{depth} + \lambda_m \cdot E_{model} \quad (2)$$

where E_{data} ensures that our desired estimation resembles the initial depth recovery via SFS, E_{color} is a multi-view constraint, imposing color consistency between two adjacent views, E_{depth} penalises depth discontinuity between neighboring pixels and E_{model} encourages local geometry coming from a compact region of the same model. The parameters λ_c, λ_d and λ_m are weighting parameters controlling the importance of each term.

Data Term E_{data}: The data term penalizes dissimilarity between the estimated output and the initialization via SFS. Let i be a 3D point of the face model, we denote D_i the depth of point i recovered using SFS, and X_i our depth estimation of the same point. Thus the data term is defined

as

$$E_{data} = \sum_{i \in \Gamma} |X_i - D_i| \quad (3)$$

where the symbol Γ denotes the 3D face model.

Color Consistency Term E_{color}: As the initial estimation D_i from SFS is not always reliable, to address this problem, we introduce a color consistency term by measuring color consistency between pairs of images. The color consistency term states that, the projections from the same 3D point onto different views should have similar appearance. Because natural face images are textureless in general, the per-pixel constraint among different views is ambiguity and insufficient. To address this problem, we employ local image patches to represent feature of the central pixel. Suppose P_u is the projection matrix to view u, $m_u^j = P_u \cdot X_j$ its 2D projection from 3D point X_j, thus we have

$$E_{color} = \sum_{(u,v)} \sum_{j \in H(i)} \|I_u(X_j) - I_v(X_j)\|_2 \quad (4)$$

Here (u, v) denotes a pair of neighboring views, and $H(i)$ is a point set with a square patch centered at i.

Depth Smoothness Term E_{depth}: The depth smoothness term ensures smooth transition in depth and penalizes sharp depth edges. This term is reasonable because human faces can always be described using smooth surfaces. We also make this term sensitive to color by assuming that depth discontinuity co-occur with intensity changes. We define the depth smoothness term as

$$E_{depth} = \sum_{i} \sum_{j \in N(i)} |X_i - X_j| \quad (5)$$

Here $N(i)$ denotes the local neighborhood of point i.

Model Consistency Term E_{model}: The model consistency term emphasizes the model consistency between neighboring points on the 3D face model. It encourages local geometry of the 3D face resembles a reference model of the same part. This part-based strategy is more flexible of describing arbitrary face model from the database compared with the model-based technique, where after optimization, all points are assigned with one certain label. This term is denoted as

$$E_{model} = \sum_{i} \sum_{j \in N(i)} \delta(L_i \neq L_j) \quad (6)$$

Here $\delta(\cdot)$ is a delta function that $\delta(true) = 1$ and $\delta(false) = 0$. And L_i denotes model index of point i.

4.1. Optimization Algorithm

With the proposed algorithm we have converted the 3D face reconstruction problem into a MRF labeling one, and designed the energy function in Eqn. 2 that combining 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 models.

Pointcloud-Depthmap Convention Similar to [13], 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 XYZ on a 3D face whose original point O is at the center of the face model, with the X-Y coordinate parallel to horizontal and vertical lines, and the Z axis points across the nose tip and frontward. In the converted depthmap, the value $d(x, y)$ at pixel (x, y) denotes distance between the corresponding 3D point $X_{\theta,\phi,d}$ and the original O along Z axis

$$X_{\theta,\phi,d} = (d \cdot tan(\theta) + C_x, \phi + C_y, d + C_z) \quad (7)$$

with

$$x = k_x \cdot \theta * 180/\pi, y = k_y \cdot \phi \quad (8)$$

Here θ is the angle between OX and Z axis, and ϕ denotes Y value in the XYZ coordinate. (C_x, C_y, C_z) is the point of O in the coordinate of the initial model. k_x and k_y are parameters controlling density of the depth map.

Graph-cuts Optimization We choose to use graph-cuts to minimize the proposed energy function in Eqn. 2 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

We evaluated the proposed algorithm from two different datasets, namely the Bosphorus dataset[1] as well as the FERET [15] dataset. In the Bosphorus dataset, there are one frontal (0 degree) and two profile ($\pm$90 degree) images for each subject, together with the groundtruth model captured via laser scanner. For the FERET dataset, there contains face images with multiple face poses. In our experiment, only the frontal and profile images are used for reconstruction. For quantitative evaluation, we compare our method

[1] http://bosphorus.ee.boun.edu.tr/Home.aspx

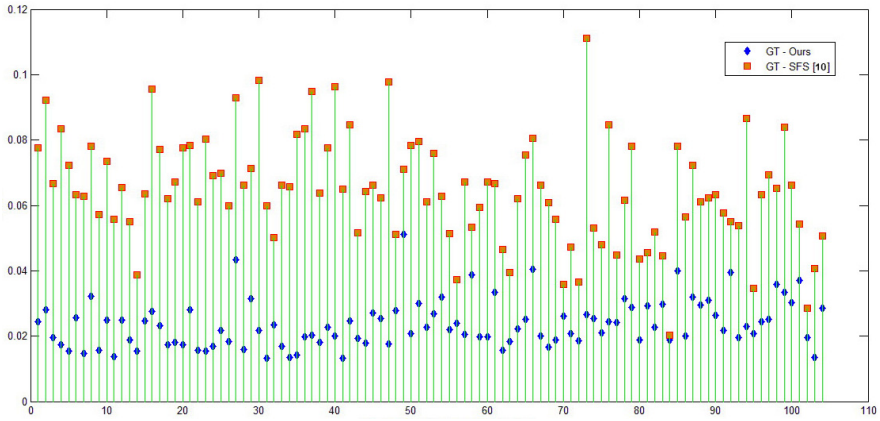


Figure 3. RMSE calculated over each of the 104 subjects from the Bosphorus database. The error between the groundtruth and SFS [10] is marked orange, and blue diamond indicates the error between the groundtruth and our result. The average error of our result is 0.0236, in comparison with SFS [10] of value 0.0648.

with the benchmark method from Ref. [10]. To demonstrate effectiveness of our proposed method, we further evaluate our results on face recognition in terms of similarity.

5.1. Implementation Details

Unless otherwise indicated, all experiments were run with the same parameters. We manually set the parameters in Eqn 2 as $\lambda_c = 1$, $\lambda_d = 30$, and $\lambda_m = 1$. In the initialization step, we set parameters $\lambda_1, \lambda_2, \sigma_x$ and σ_x be 30, 30, 5 and 5 respectively. We implemented the proposed algorithm using C++ on a 64-bit windows workstation with Intel $i5$ CPU and 4GB memory.

Our prior 3D face models come from the BU-3DFE database that contains 100 subjects containing races of White, Black, Indian, East Asian, Middle-east Asian and Lation-Hispanic, which largely reduces reliability on the prior models. The 3D face models are first registered to the initial reconstruction, and served as depth candidates for further optimization.

5.2. Results and Discussion

The proposed algorithm is mainly compared with method [10]. Ref. [10] 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 3D face models, which is much less model-reliant compared with the single reference model approach. Thanks to our example-based pipeline, our method is capable of producing accurate 3D faces especially on overlapping face regions from different views.

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

$$err = \frac{\sqrt{\sum_i |X_i^{es} - X_i^{gt}|}}{N} \tag{9}$$

where i is index of each pixel and N the total number of valid pixels in use. Fig. 3 shows the overall mean reconstruction error of each of the models in the Bosphorus dtabase. In all cases, the reconstruction errors are much smaller than the differences between the initial reconstruction and the groundtruth.

Fig. 4 shows examples from Bosphorus dataset compared with the prior art [10]. The error map is calculated using absolute difference between the estimated model and the groundtruth. Our optimization pipeline can correct large depth errors, both on the surface and at depth boundary, thanks to the multi-view setting of our algorithm. Fig. 5 shows rendering results from new synthesized viewpoint, where the results from SFS [10] are oversmoothed. Meanwhile, our method can preserve more geometric details, thus producing more accurate and visually-pleasant results. Fig. 6 shows qualitative results from FERET dataset, which is much more challenging because the profile faces are less controlled and exhibit more variations. However, by using the profile images, our method can largely improve quality the final model.

Similarity Measure across Pose One potential application using 3D face models is face recognition, which pro-

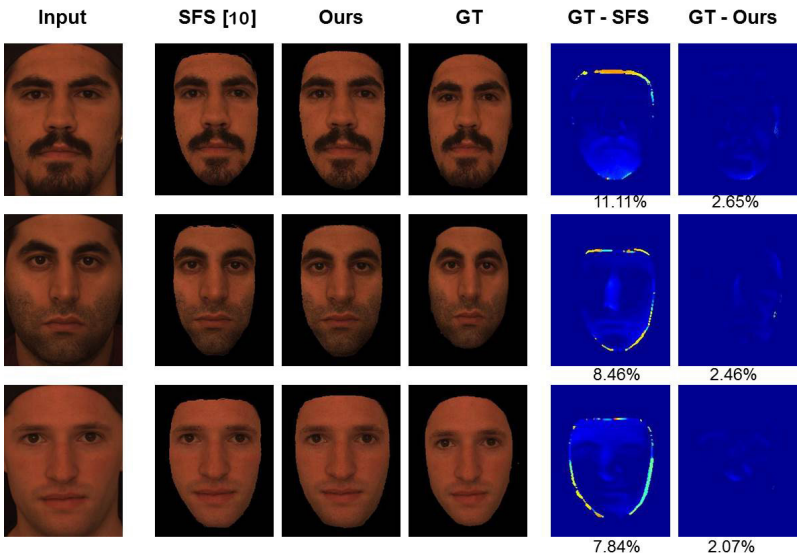


Figure 4. Example of the reconstruction results from the Bosphorus database. Each row shows one example. Each column shows, from left to right, input image of the frontal view, reconstruction result using SFS [10], our model and the groundtruth, and the last two columns are error maps from SFS [10] result and our method.

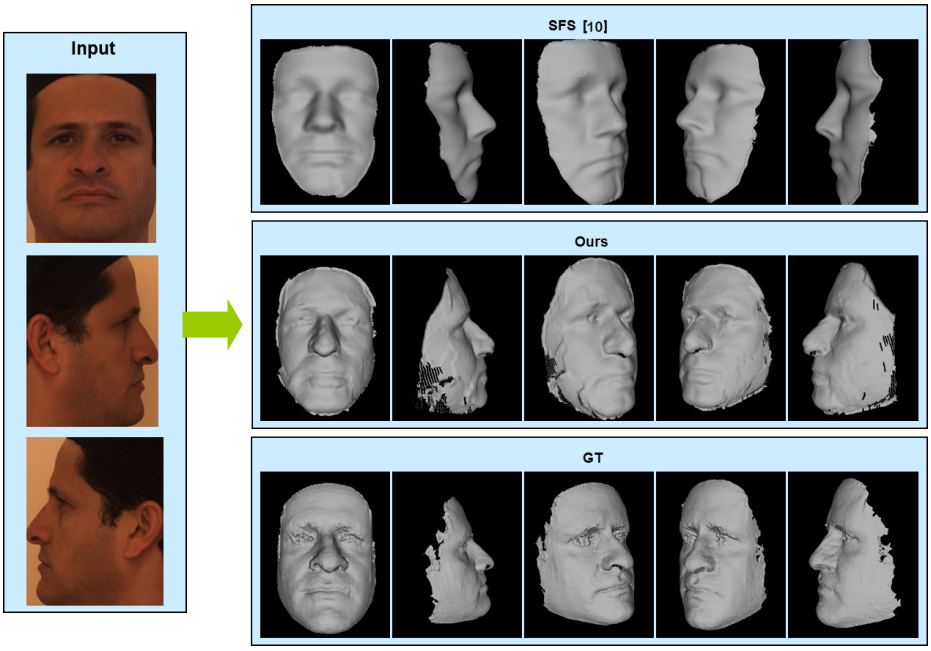


Figure 5. Example of the reconstruction results from the Bosphorus database. The first column shows both frontal and profile faces of the input. The following shows reconstructed results rendered from different viewpoints, with SFS [10] method the first row, our estimation in the second row, and the last row shows the groundtruth scanning.

motes face recognition rate with large pose variations. Given an arbitrary non-frontal face image, the identity of a subject can be more accurately located if 1) the subject is in the gallery and 2) with approximately the same pose as the query image. In this perspective, a 3D face model can rotate towards any direction you desire, which helps pose-invariant face recognition.

The basic but essential step towards face recognition is the similarity measure between query and gallery. In our experiment, color histogram is extracted on both the query

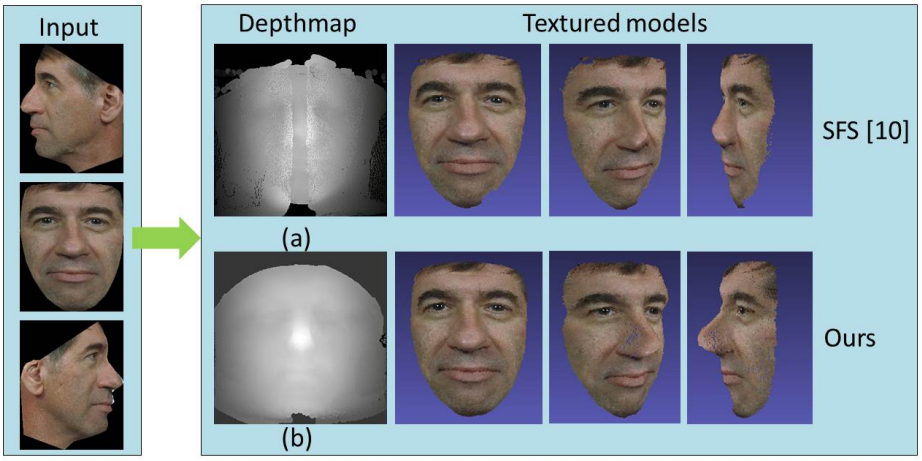


Figure 6. Example of the reconstruction results from the FERET database. The first column shows both frontal and profile faces of the input. View-dependent depthmaps from SFS [10] are registered as an initial guess (a) and the optimized depthmap model is shown in (b). The following colums shows textured models rendered from different viewpoints, with SFS [10] method the first row, our estimation in the second row.

image and the rendered one in the database. Both shape and texture are projected according to the pre-estimated pose of the query face to synthesis the rendered image. Similarity measure is then computed using Euclidean distance.

For a query face, the 3D model with the same identity is used for similarity measure. To show the power of our method, we compare the similarity from our reconstructed model to that from the model obtained using SFS [10]. Fig. 7 shows similarity measures with respect to each subject. In general, our fused model achieves higher similarity score compare with the query image, demonstrating higher recognition rate using our proposed method. Furthermore, our result is more stable across different model in the database, showing robustness of our algorithm.

6. Conclusion

In this paper we proposed an example-based method for multi-view 3D face reconstruction. Our problem is challenging by using uncalibrated the input images with wide baseline. As face images are textureless, the traditional multi-view stereo pipeline could not work for our problem. We address this by using an external face database and synthesis the result through facial part composition. We proposed an energy function to formulate the 3D face reconstruction problem and solved it via multi-view image segmentation algorithm. This work can be extended to 3D face recognition task that utilize both texture and 3D shape, but the extension is beyond the scope of this paper.

Acknowledgement

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

References

[1] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1063–1074, 2003.

[2] J. Choi, G. Medioni, Y. Lin, L. Silva, O. Regina, M. Pamplona, and T. C. Faltemier. 3d face reconstruction using a single or multiple views. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3959–3962. IEEE, 2010.

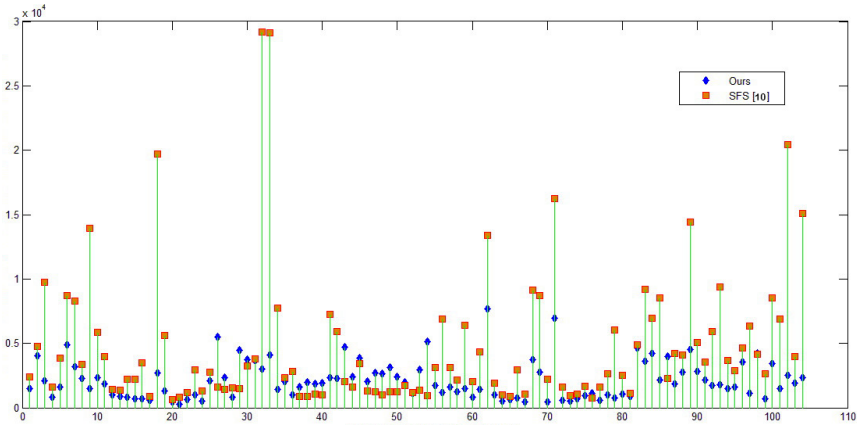
[3] J. Gonzalez-Mora, F. De la Torre, N. Guil, and E. L. Zapata. Learning a generic 3d face model from 2d image databases using incremental structure-from-motion. *Image and Vision Computing*, 28(7):1117–1129, 2010.

[4] H. Han and A. K. Jain. 3d face texture modeling from uncalibrated frontal and profile images. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, pages 223–230. IEEE, 2012.

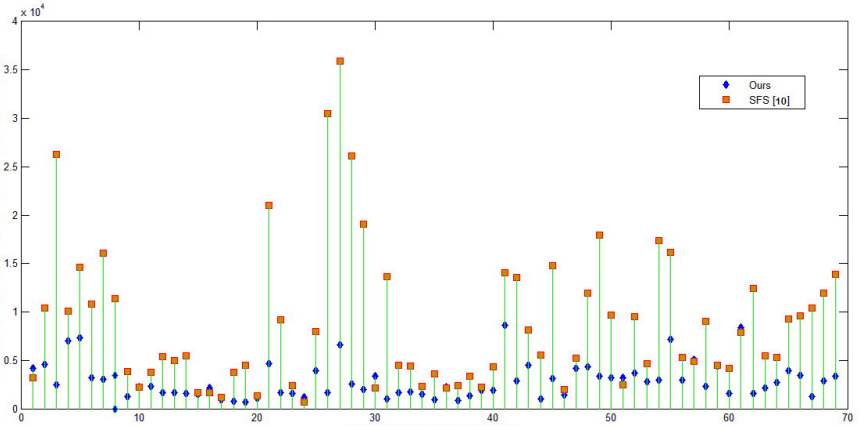
[5] T. Hassner. Viewing real-world faces in 3d. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3607–3614. IEEE, 2013.

[6] J. Heo and M. Savvides. Rapid 3d face modeling using a frontal face and a profile face for accurate 2d pose synthesis. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 632–638. IEEE, 2011.

[7] J. Heo and M. Savvides. Gender and ethnicity specific generic elastic models from a single 2d image for novel 2d



(a) Query face heading about 67.5 degree right



(b) Query face heading about 67.5 degree left

Figure 7. Similarity calculated over part of the FERET database with Fig. 7(a) query face heading 67.5 degree right and Fig. 7(b) query face heading 67.5 degree left. The x-axis is subject ID, and the y-axis denotes distance between the query image and the rendered face of the same identity. The similarity over the query face and SFS [10] rendering is marked orange, and blue diamond indicates distance between the query face and the rendered one using our reconstructed model.

pose face synthesis and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(12):2341–2350, 2012.

[8] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, and W. Gao. Efficient 3d reconstruction for face recognition. *Pattern Recognition*, 38(6):787–798, 2005.

[9] I. Kemelmacher and R. Basri. Molding face shapes by example. In *Computer Vision–ECCV 2006*, pages 277–288. Springer, 2006.

[10] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):394–405, 2011.

[11] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1746–1753. IEEE, 2011.

[12] W. B. Lee, M. H. Lee, and I. K. Park. Photorealistic 3d face modeling on a smartphone. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 163–168. IEEE, 2011.

[13] Y. Lin, G. Medioni, and J. Choi. Accurate 3d face reconstruction from weakly calibrated wide baseline images with profile contours. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1490–1497. IEEE, 2010.

[14] M. Pamplona Segundo, L. Silva, and O. R. P. Bellon. Improving 3d face reconstruction from a single image using half-frontal face poses. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1797–1800. IEEE, 2012.

[15] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(10):1090–1104, 2000.

[16] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *Computer Vision–ECCV 2014*, pages 796–812. Springer, 2014.