Model-based Face Reconstruction using SIFT Flow Registration and Spherical Harmonics

Fanzi Wu, Songnan Li, Tianhao Zhao, King Ngi Ngan
Department of Electronic Engineering
The Chinese University of Hong Kong
{fzwu, snli, thzhao, knngan}@ee.cuhk.edu.hk

***Abstract*—In this paper, we propose a robust method for face reconstruction using a single color image. A 3D morphable model is used to reconstruct a smooth 3D face shape. To find the correspondence between model vertices and image pixels, landmarks are updated using SIFT flow which is illumination and rotation invariant. To reconstruct more detailed information, depth values are refined using a shape from shading method which approximates lighting condition by spherical harmonics. We test the proposed method on a set of real world images and compare reconstructed results with depth maps captured by a depth camera. The average error is around 3.3 mm.**

I. Introduction

3D face reconstruction has been widely used in face recognition [1] [2] and face tracking [3]. It also plays an important role in augmented reality, virtual reality, 3D printing and etc. Accurate face reconstruction requires information of shape, texture, illumination and camera property. Thus recovering a 3D face is ill posed and unconstrained when only a single color image is available. Under this circumstance, prior information and assumption are essential. Blanz and Vetter [5] [6] fitted a 3D morphable model (3DMM) based on correspondences using optical flow between the reconstructed face and the input image. This method could reconstruct faithful human faces but is highly computationally expensive because of the non-linear optimization. Romdhani *et al.* [7] proposed a linear approach to compute an incremental update to the shape and texture parameters given dense measurements of residual errors from optical flow. Besides optical flow, there are many other methods to find correspondences [8]. Moghaddam *et al.* [9] used silhouettes computed from a large number of input images. Since pixel intensities of a face image can be dramatically influenced by expression, occlusion and complicated illumination, model fitting methods are restricted by these variations. Zhu and Yi [10] proposed a robust method using multi-feature framework which included SIFT feature, pixel intensity and contours. Since SIFT feature is proved to be invariant to uniform scaling, orientation, and partially invariant to affine distortion and illumination changes, it can locate the facial components successfully [11]. However, face external boundary was not considered in this method, which may have significant influence to the reconstruction quality. Keypoint is another popular feature which provides sparse correspondence [12]. However, this sparse correspondence was manually marked which may cause matching error. Therefore we combine face boundary landmarks and SIFT features, to improve the model fitting result on both inner facial components and face external boundary.

Moreover, the shape description ability of 3DMM is limited. High frequency details such as wrinkles and nevi cannot be synthesized by 3DMM. This restriction of model itself can be rectified by shape from shading methods. Pixel intensities from the input image provide information about the luminance, albedo and shape. Basri and Zhang [13] [14] proposed a method to estimate the illumination of images using spherical harmonics approximation. The spherical harmonics are related to normals which are gradients of depth values. Some researchers used the average shape obtained from prerequisite images as a reference shape, and reconstructed a detailed face from a single image using spherical harmonics [3]. The average shape highly relies on the collected images, e.g., if the collected images were occluded by glasses, the average shape will show artifacts of glasses. Raw depth captured by a depth camera was used as a reference shape in [15]. Besides, [16] [17] proposed using depth from 3DMM as a reference shape, which inspired our work.

In this paper, we propose a robust method which combines 3D morphable model and shape from shading to reconstruct a highly detailed 3D face. The proposed method combines facial landmarks and SIFT flow to fit the 3DMM. To improve the fitting accuracy, we use normals of landmarks in the objective function. Then based on the reconstructed 3D face, depth values are refined to recover details using spherical harmonics approximation.

This paper is organized as follows: In Section II, we briefly introduce the 3DMM and spherical harmonics approximation. In Section III, the proposed method is explained in details which include the model fitting and shape from shading. In Section IV, we present experimental results and give both subjective and objective evaluations. The final section draws the conclusion and discussion the future work.

II. Related Work

A. 3D morphable model

A 3DMM is constructed from ω registered 3D faces with d vertices. Each face is represented by a shape vector $S = [x_1, y_1, z_1, ..., x_d, y_d, z_d]$ in x, y, z coordinates and a texture vector $T = [r_1, g_1, b_1, ..., r_d, g_d, b_d]$ in RGB channels. By applying PCA to ω face samples, principal components can be extracted. Then new shape and texture can be synthesized as a linear combination of principal components.

$$S(\alpha) = \mu_s + U_s\alpha \qquad T(\beta) = \mu_t + U_t\beta \tag{1}$$

where U_s and U_t are principal components for shape and texture model, $\alpha = [\alpha_1, \alpha_2, ..., \alpha_\omega]$ and $\beta = [\beta_1, \beta_2, ..., \beta_\omega]$ are parameters for U_s and U_t, μ_s and μ_t are the mean shape and texture.

B. Shape from shading

Besides the correspondences between 3DMM and the input image, pixel intensities are also used to estimate depth values. Human face is assumed to be a Lambertian surface and the luminance is uniform for all points belonging to the surface. The Lambertian surface only has low frequency reflectance. We use spherical harmonics to approximate the lighting condition. The spherical harmonics are a set of functions that form an orthogonal basis for the set of all square integrable functions defined on the unit sphere. They are analogue on the sphere to the Fourier basis on the line or circle. Under this circumstance, the reflectance of a image could be approximated as:

$$R(x,y) \approx \sum_{j=0}^{J}\sum_{v=-j}^{j} l_{jv}Y_{jv}(x,y) \tag{2}$$

where j is the order of spherical harmonics and v represents different components of j^{th} spherical harmonic function, $R(x,y)$ is the reflectance at point (x,y), l_{jv} are coefficients of lighting and $Y_{jv}(x,y)$ are the surface spherical harmonic functions evaluated at the surface normal. To reduce the complexity, we use the first order spherical harmonic functions with

$$Y(x,y) = (1, n_x, n_y, n_z) \tag{3}$$

Then pixel intensity from a face image is represented by:

$$I(x,y) = \rho(x,y) \times l \times Y(x,y) \tag{4}$$

where $\rho(x,y)$ is the albedo, i.e, the ratio of reflection to incidence, $Y(x,y)$ denotes the surface spherical harmonic functions and l represents coefficients of $Y(x,y)$.

III. Method

The proposed method consists of two main stages: 3DMM fitting and refinement of depth values. In the model fitting stage, key points of 3DMM are manually marked and landmarks are detected from the input image. Shape parameters are estimated by minimizing the sum of squared distances between vertex projections and landmarks. Landmark positions are updated using SIFT flow computed from the reconstructed face and the input image. Then shape parameters are updated by fitting 3DMM to the updated landmarks. This process iterates until its convergence. Using depth values rendered from the reconstructed face, lighting condition and albedo map are estimated and in turn depth values are refined accordingly. Fig. 1 displays these main steps in the flow chat.

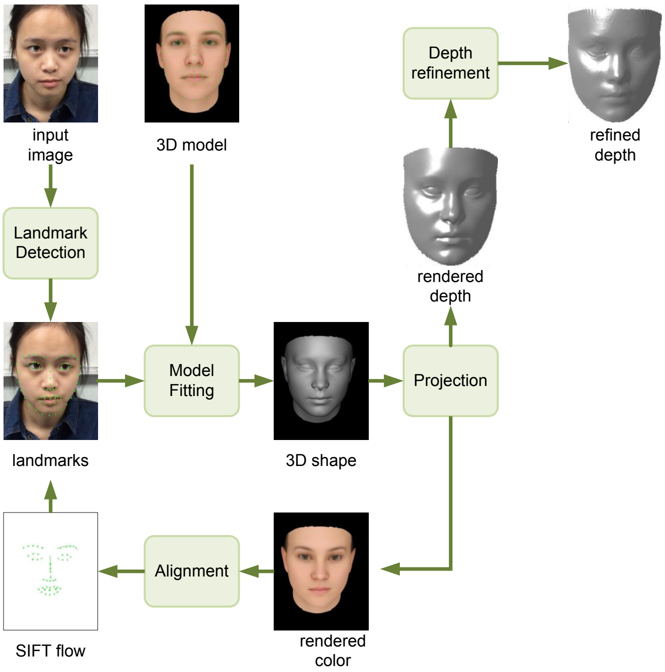

Fig. 1. Flow chart of the proposed method

A. Model fitting

In the model fitting stage, we estimate shape parameters $\alpha = [\alpha_1, \alpha_2, .., \alpha_m]$ for shape reconstruction. Landmarks are detected in the input image using the algorithms in [18] [19] which then are updated using SIFT flow computed from the rendered image and the input image. In the cost function, we take normal of landmarks into consideration. The shape parameters, transformation and focal length can be estimated by minimizing the sum of squared distances between vertex projections and detected landmarks:

$$\min_{T,R,f,\alpha} \sum_i \|p_i(T,R,f,\alpha) - l_i\|^2 + \lambda\alpha^T Q\alpha \tag{5}$$

where l_i is the 2D position of the i^{th} landmark, T, R, f and α are the translation vector, rotation matrix, focal length and face model parameter vector, respectively. Q is a diagonal matrix whose elements contain the reciprocal of variances for the principle components. This regularization term prevents the reconstructed shape drifting too far away from the mean face shape. $p_i(\cdot)$ calculates the perspective projection of the i^{th} vertex of the face model

$$p_i(T,R,f,\alpha) = \pi_f(R(\mu_i + U_i\alpha) + T) \tag{6}$$

where $\pi_f(\cdot)$ calculates the perspective projection which involves the focal length f.

The detailed model fitting process consists of two steps: registration and multi-PCA modeling, which will be introduced in the following section.

1) Registration: We manually mark corresponding vertices on 3DMM by observation which then is applicable for any input face image. Notice that the corresponding vertices on

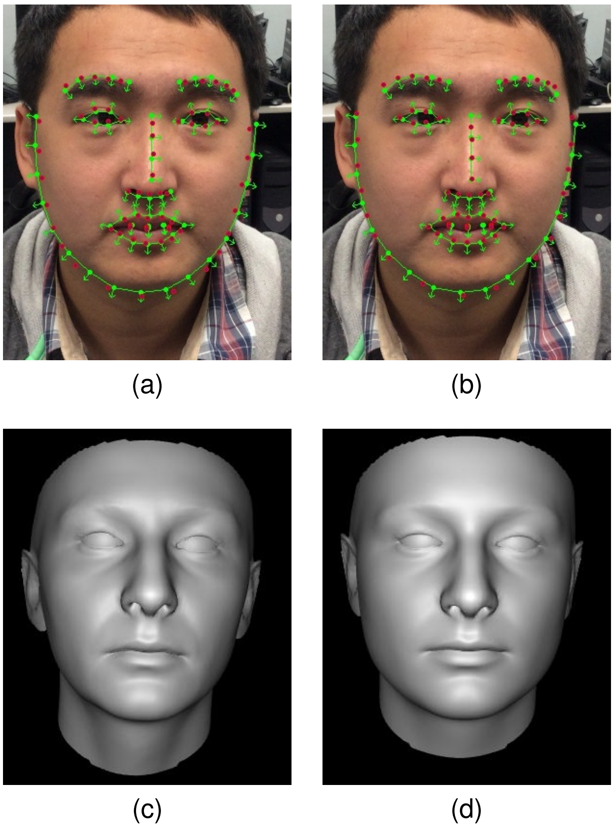
(a)
(b)
(c)
(d)

Fig. 2. Comparison between two methods. Green points are the detected landmarks and red points are vertex projections. (a) and (c): Fitting results without using line normals; (b) and (d): Fitting result using normals.

the face boundary may change as the shape and pose change. To ensure that corresponding vertices are visible, we mark a set of points as the potential boundary vertices as in [20] . When shape parameters and the pose change, corresponding vertices are updated by checking their normals.

Furthermore, we found that the boundary landmarks detected by [19] is not consistent across different head poses, since there is no clear semantic correspondences for landmarks on the face boundary. To solve this problem, we add a normal projection term in the objective function:

$$\min_{T,R,f,\alpha} \sum_{i} \langle n_i, (p_i(T,R,f,\alpha) - l_i) \rangle^2 + \lambda \alpha^T Q \alpha \tag{7}$$

where n_i is the normal of the line connecting the $(i+1)^{th}$ and $(i-1)^{th}$ landmarks as shown in Fig. 2a. The resulting effect is that the vertex projections do not necessarily fall onto the exact location of the landmarks. Instead, they can slide along the face boundary without increasing the matching error. Fig. 2 shows the fitting results using objective functions eq. (5) and eq. (7). By comparison between Fig. 2c and Fig. 2d, it can be observed that using normals can generate better fitting results, especially for the face boundary.

2) Multi-PCA modeling: As the shape of the human head is complex and exhibits a large variation among individuals, single PCA cannot yield a sufficiently adaptive model [21]. Therefore, we segment the 3DMM into four regions: nose, eyes, mouth and the rest. The transformation parameters R and T are fixed, and each region r_k is assigned with specific shape parameters and associated with a set of landmarks. The objective function could be rewritten as:

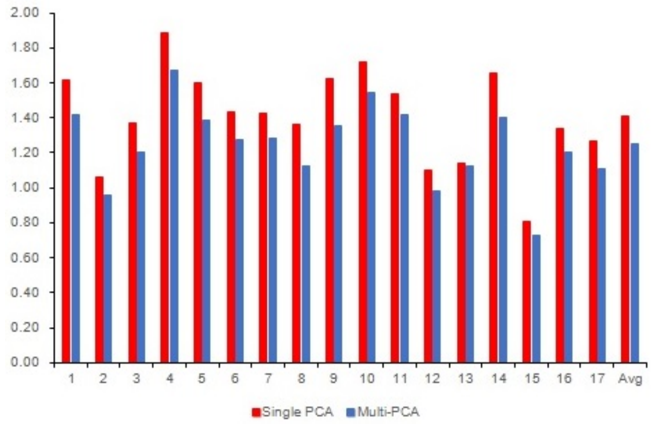

Fig. 3. Distance between landmarks and vertex projections. X-axis represents different input images and the unit of Y-axis is pixel.

$$\min_{\alpha_k} \sum_{i \in r_k} \langle n_i, (p_i(T,R,f,\alpha_k) - l_i) \rangle^2 + \lambda \alpha_k^T Q \alpha_k \tag{8}$$

We solve $\alpha_k = [\alpha_{k1}, \alpha_{k2}, ..., \alpha_{k\omega}]$ separately and reconstruct four different faces using α_k. Here ω is the dimension of principal components. For each reconstructed face, we have a weight vector which assigns larger weight to vertices belonging to its associated region. The reconstructed face is a blended shape using four PCA parameters:

$$\sum_{k=1}^{4} w_k(\mu_s + U_s \alpha_k) \tag{9}$$

where U_s and μ_s are the PCA basis and the mean shape of 3DMM, and $w_k = [w_{k1}, w_{k2}, ..., w_{k3d}]$ is the weight vector for region r_k.

Fig. 3 shows the average distance between landmarks and vertex projections for different input images. The results show that using multi-PCA has consistent improvement. The average distance is reduced by 0.15 pixel.

B. SIFT flow alignment

We render the reconstructed shape with the mean texture. By comparison between the rendered image and the input image, small displacements can be observed between these two images. By minimizing the displacement map, the reconstructed face could become realistic. SIFT features are local and invariant to the image scale and rotation. They are also robust to changes in illumination, noise, and minor changes in viewpoint. Therefore, we use SIFT flow to refine the correspondence between two images.

By extracting dense SIFT features and compute the SIFT flow, the displacement map can be generated. Considering the i^{th} key point, the location of its projection is (x, y) and the

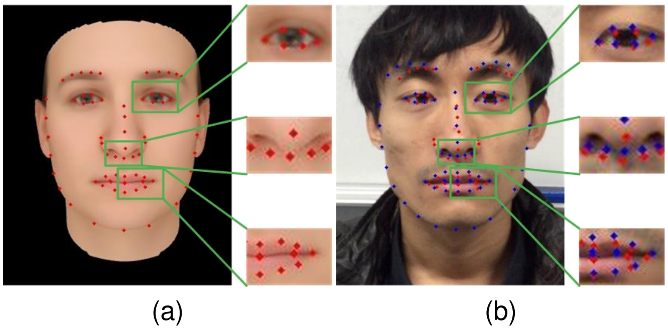

Fig. 4. Comparison between Landmarks.(a) key points on 3DMM. (b) Red: updated landmarks, Blue: detected landmarks.

displacement vector is u. Then the position of the i^{th} landmark should be replaced by $(x, y) + u$. We show updated landmarks and detected landmarks in Fig. 4. Compared with vertices of 3DMM in Fig. 4a, landmarks updated using SIFT flow has more accurate correspondence. Taking vertices of the mouth upper lip as an example, the updated landmarks are exactly at the mouth boundary, while the detected landmarks are slightly above the lips.

C. Depth refinement

3DMM only provides a smooth face shape without details like wrinkles and nevi. Based on the reconstructed face from 3DMM, we use spherical harmonics approximation to estimate the lighting condition and refine depth values.

1) Lighting condition and albedo estimation: We firstly assume that $\rho(x, y) = 1$ for all points from the image, and normal map from the reconstructed face is used as a reference. Then the lighting coefficients are estimated by solving:

$$I(x, y) = l \times Y(x, y) \tag{10}$$

where $Y(x, y) = (1, n_x, n_y, n_z)$, and n_x, n_y, n_z are three components of the normal vector. The lighting coefficients could be estimated by solving the above linear equation.

After calculating the lighting coefficients, we estimate the albedo map by solving the following linear equation:

$$I(x, y) = \rho(x, y) \times l_{est} \times Y(x, y) \tag{11}$$

Eq. (11) is also linear which can be solved using LU decomposition.

2) Shading-based depth refinement: Based on the estimated lighting condition and albedo map, we estimate the depth map by replacing $Y(x, y)$ in eq. (11) with depth value $z(x, y)$.

The normal vector can be represented by depth values:

$$(n_x, n_y, n_z) = \frac{1}{N}(\frac{\partial x}{\partial z}, \frac{\partial y}{\partial z}, -1) \tag{12}$$

where N is a normalization factor:

$$N = \sqrt{(\frac{\partial x}{\partial z})^2 + (\frac{\partial y}{\partial z})^2 + 1} \tag{13}$$

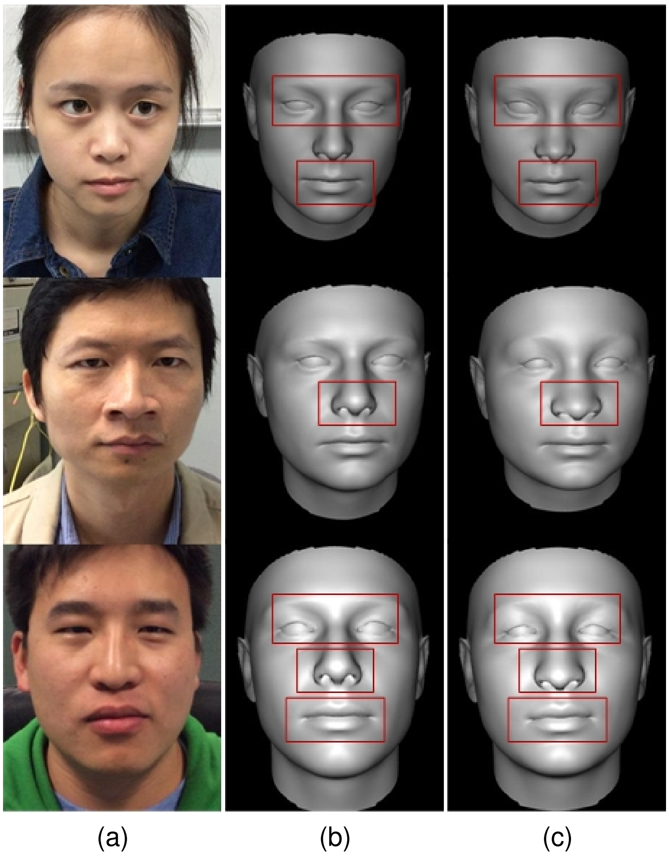

Fig. 5. Model fitting results. (a) Input images. (b)Initialization of pose and shape. (c) Final fitting results.

and

$$\frac{\partial x}{\partial z} = z(x + 1, y) - z(x, y) \qquad \frac{\partial y}{\partial z} = z(x, y + 1) - z(x, y) \tag{14}$$

By replacing (n_x, n_y, n_z) in eq. (12) with eqs. (13) and (14), and substituting them into eq. (11), we obtain a least square problem which can be solved analytically. Since pixel intensities are used, details of the human face as shown in the experimental section can be recovered.

IV. Experiments

For 3DMM, we used the Basel Face Model [4] which is derived from 3D face scans of 100 male and 100 female subjects. It is PCA-based and can morph to a wide range of shapes and textures by adapting parameters. To reduce the complexity, we used the first 30 principal components to reconstruct the 3D shape. The proposed method was tested on real life images. We collected 17 human faces under different poses and lighting conditions with $640 * 480$ RGB images. For each image, we detected 49 inner landmarks and 17 boundary landmarks to initialize the 3D shape. Then the inner landmarks were updated by SIFT flow. Since detected inner landmarks [18] do not provide landmarks of nose wings, we added 4 landmarks of the nose boundary during the alignment.

TABLE I
AVERAGE DISTANCE WITH DEPTH MAPS CAPTURED BY STRUCTURE SENSOR(UNIT:MM)

Faces	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	avg
InitDep[1]	4.61	4.55	3.48	4.06	4.12	4.88	4.20	2.27	2.87	3.42	3.95	3.27	5.09	5.28	3.86	3.99	3.94	3.99
InitDepN[2]	4.08	4.06	2.97	3.76	3.33	4.21	4.14	2.28	2.38	3.13	3.32	3.24	4.43	4.80	3.07	3.21	3.72	3.54
RenderDep[3]	3.32	4.11	2.92	3.39	3.56	4.25	3.34	2.94	2.27	3.36	3.17	2.88	3.06	4.19	3.50	2.86	3.49	3.33
RefineDep[4]	3.37	4.01	2.88	3.43	3.47	4.21	3.29	2.88	2.26	3.34	3.13	2.76	3.05	4.26	3.37	2.84	3.42	3.29

[1] Depth values from initialized model. [2] Depth values from initialized mode using normals. [3] Depth values from final fitted model. [4] Depth values refined using shape from shading method.

In the model fitting stage, pose and shape parameters are initialized using the 66 landmarks, where single PCA model is used. Then the initialized 3D face with the mean texture is projected into a 2D image. We estimate the lighting condition and apply it to render the color image. By computing the SIFT flow between the rendered image and the input image, the projection of vertices can be refined, and the landmarks can be updated. The 3DMM is divided into 4 segments and fitted to the updated landmarks. For each segment, one 3D face is generated, and they are blended using weight vectors as introduced in Section III.A. To improve the performance, five iterations are performed.

Fig. 5 shows the model fitting results. The initialization could provide a preliminary result but have flaws in certain face attributes as shown in Fig. 5b. By comparison between Fig. 5b and Fig. 5c, it can be observed that Fig. 5c has improvements over eyes, nose and mouth. We marked face attributes which have obvious improvement with red rectangles. Since depth values around the eyes vary moderately, the shape of the eye can be further modified in the depth refinement stage.

We compare our method with the SSF algorithm in [10]. We test images rendered from 10 face scans under 3 pose conditions (yaw rotations from -15 degree to +15 degree) and 3 lighting conditions (different directions and strengths). We evaluate the precision of the shape fitting using Root Mean Square Error (RMSE) as in [10]. Fig 7 shows the RMSE for all face points (excluding ears and neck). We can see that the average error is around 3.5 mm, while the average error of SSF is around 6.5 mm.

The refined depth values are also compared with depth maps captured by a depth camera (Structure Sensor), which has errors from 0.5 mm to 30 mm at a range from 400 mm to 3000 mm. The average distance is computed by applying rigid transformation to the refined depth values and comparison with raw depth captured by the depth camera. Table I and Fig. 6 shows the average distance per pixel with the raw depth. The distance between human face and the depth camera is about 500 mm while the average error is around 3.3 mm. The initial depth map is generated from the initialized model without using normal direction and it is reduced by 0.7 mm using the proposed method. Although the refined depth reduced the average error slightly compared with depth maps generated from the reconstructed face, the subjective quality improvement is significant as shown in Fig. 8.

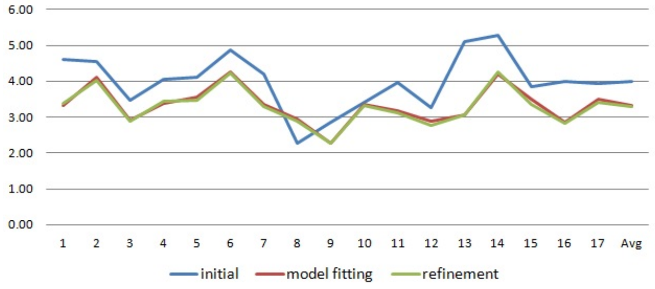

Fig. 6. Comparison with depth maps captured by Structure Sensor. X-axis represents different face images.

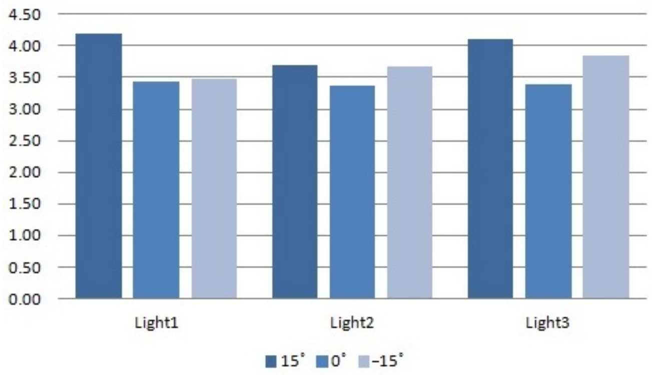

Fig. 7. RMSE of the whole face over the yaw rotations $(-15°, 0°, +15°)$ under different lighting conditions

In Fig. 8, we present the normal map of the refined depth. The details can be reconstructed well especially for the eye shape and some obvious nevi on the face. We also present the 3D views of the reconstructed face. The subjective quality of the results is substantially improved over the eyes and mouth. More details appear in the final result compared to the ones shown in Fig. 5.

V. CONCLUSION

In this paper, a novel and accurate face reconstruction algorithm using SIFT flow and spherical harmonic approximation was proposed. We detected inner landmarks and boundary landmarks, and updated them using SIFT flow. Then the depth map rendered from the reconstructed face was refined by shape from shading method which uses spherical harmonic approximation to estimate the lighting condition. We tested

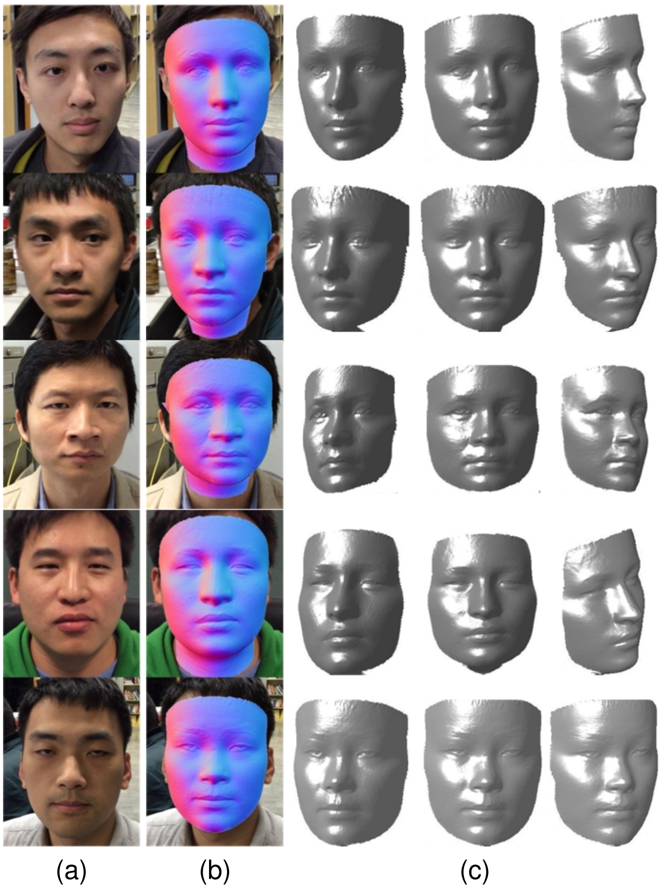

Fig. 8. Reconstruction results of the proposed method. (a) input images. (b) normal maps of final results. (c) 3D views of the reconstructed face.

our algorithm using images captured by color camera with different poses and illuminations. The obtained results showed that the subjective quality of the reconstruction results are satisfying. Furthermore, our results were compared with data captured from a depth camera, which showed that the obtained results using a single color image have a reconstruction error around 3.3 mm on average. Our future work will involve facial expression and texture model to make the algorithm more generally applicable and the reconstructed face more realistic. Moreover, ground truth obtained using devices with higher accuracy will be incorporated in the future work.

References

[1] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang, "Automatic 3d reconstruction for face recognition," in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*. IEEE, 2004, pp. 843–848. 1

[2] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, and W. Gao, "Efficient 3d reconstruction for face recognition," *Pattern Recognition*, vol. 38, no. 6, pp. 787–798, 2005. 1

[3] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz, "Total moving face reconstruction," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 796–812. 1

[4] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*. IEEE, 2009, pp. 296–301. 4

[5] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194. 1

[6] ——, "Face recognition based on fitting a 3d morphable model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1063–1074, 2003. 1

[7] S. Romdhani, V. Blanz, and T. Vetter, "Face identification by fitting a 3d morphable model using linear shape and texture error functions," in *Computer VisionECCV 2002*. Springer, 2002, pp. 3–19. 1

[8] A. Moeini, H. Moeini, and K. Faez, "Expression-invariant face recognition via 3d face reconstruction using gabor filter bank from a 2d single image," in *2014 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 4708–4713. 1

[9] B. Moghaddam, J. Lee, H. Pfister, and R. Machiraju, "Model-based 3d face capture with shape-from-silhouettes," in *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*. IEEE, 2003, pp. 20–27. 1

[10] X. Zhu, D. Yi, Z. Lei, S. Z. Li *et al.*, "Robust 3d morphable model fitting by sparse sift flow." in *ICPR*, 2014, pp. 4044–4049. 1, 5

[11] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157. 1

[12] O. Aldrian and W. A. Smith, "A linear approach of 3d face shape and texture recovery using a 3d morphable model," in *Proceedings of the British Machine Vision Conference, pages*, 2010, pp. 75–1. 1

[13] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 2, pp. 218–233, 2003. 1

[14] L. Zhang and D. Samaras, "Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 3, pp. 351–363, 2006. 1

[15] C. Wu, M. Zollhöfer, M. Nießner, M. Stamminger, S. Izadi, and C. Theobalt, "Real-time shading-based refinement for consumer depth cameras," *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2014)*, vol. 33, p. 3, 2014. 1

[16] 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*, vol. 33, no. 2, pp. 394–405, 2011. 1

[17] F. Shi, H.-T. Wu, X. Tong, and J. Chai, "Automatic acquisition of high-fidelity facial performances using monocular videos," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, p. 222, 2014. 1

[18] F. de la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, "Intraface," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8. 2, 4

[19] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models." in *BMVC*, vol. 2, no. 5. Citeseer, 2006, p. 6. 2, 3

[20] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, no. 3, pp. 413–425, 2014. 3

[21] D. C. Schneider and P. Eisert, "Fitting a morphable model to pose and shape of a point cloud." in *Vmv*, 2009, pp. 93–100. 3