3D face reconstruction using similar facial components

Robert A. Smith*
Smith Research

Abstract

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Citations can be done this way [Jobson et al. 1995] or this more concise way [1995], depending upon the application.

Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit [Pellacini et al. 2005] in vulputate velit esse molestie [Landis 2002] consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. [Park et al. 2006]

CR Categories: I.3.3 [Computer Graphics]: Three-Dimensional Graphics and Realism—Display Algorithms I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Radiosity;

Keywords: radiosity, global illumination, constant time

Links: DL PDF

1 Introduction

2 Related Work

3 Data Format

Our dataset consists of 1204 Caucasion individuals ranging from 3 to 40 with males and females. The high-resolution 3D facial meshes are obtained from a 3dMD digital stereophotogrammetry imaging system(Atlanta,GA). These systems are outfitted with multiple CCD cameras mounted at fixed angles and distances, to capture overlapping views of the faces and heads. Each of the 3D heads comprises of 30-40,000 points and subjects are all in neutral expressions. Subjects were required to wear caps to get rid of the hairs that obscure their faces. Before the experiment, the meshes were cleaned and pose-normalized using a method described in [**?**]. 22 anatomical facial landmarks were labeled manually on each surface by a single trained medical expert as shown in Fig.1. With a generic face mesh and 22 landmarks, deformable registration is applied to all the meshes so that they all have dense correspondences to each other [Allen et al. 2003]. We manually placed 83 landmarks on the generic face mesh G and transfer them to all the meshes accordingly. These 83 landmarks are very crucial in determing different facial component regions as well as helping to warp the dataset towards Kinect input face.Also, we divide the generic face into 5 facial components, namely Eyes, Nose, Mouth, LeftCheek and RightCheek as shown in Fig. 2.

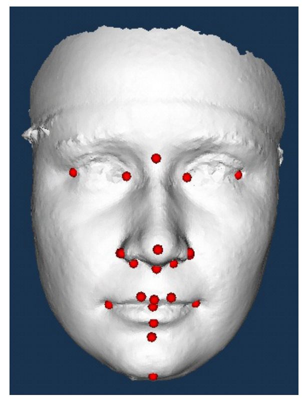

Figure 1: *High resolution 3dMD facial mesh with 22 anatomical landmarks*

4 Overview

The pipeline of our method is show in Fig. 2. Given a single RGBD frame of a person's frontal face, we first use the toolkit FacePlusPlus [Inc. 2013] to generate 83 facial landmarks on the color image. Then, we transfer the 2D landmarks to 3D points P according to the depth image using the caliberated camera projection matrix. The face region is cropped out with the help of P and the 3D point cloud is connected to form a mesh O. The Kinect raw input is smoothed using curvature flow to reduce noise in the original mesh, denoted as S. Since landmarks P on face contour can be inaccurate if just inferred the frontal view, we pick 15 anatomical landmarks P_{15} to align the input to our generic face model G. With the help of the 15 landmarks, we deform our generic face model to S to generate a dense correspondence. In case that the exact same face for S doesn't exist in our dataset, we warp all the face meshes in our dataset towards S using the 83 landmarks P. We retrieve for the most similar facial component for each of the 5 facial parts in S among the deformed dataset. After we get the most similar high-resolution facial component, we transfer the vertex normals from the HR mesh to the Kinect smoothed mesh S and use the normal-correct-position method to generate a final high resolution 3D mesh. Note that here, we just use the corresponding information to find masks for each facial components and we do not want to use the deformed generic shape to represent the input shape. Our further steps of similarity comparison and combining normals with points are still performed in the Kinect input space.

*e-mail:rsmith@gmail.com

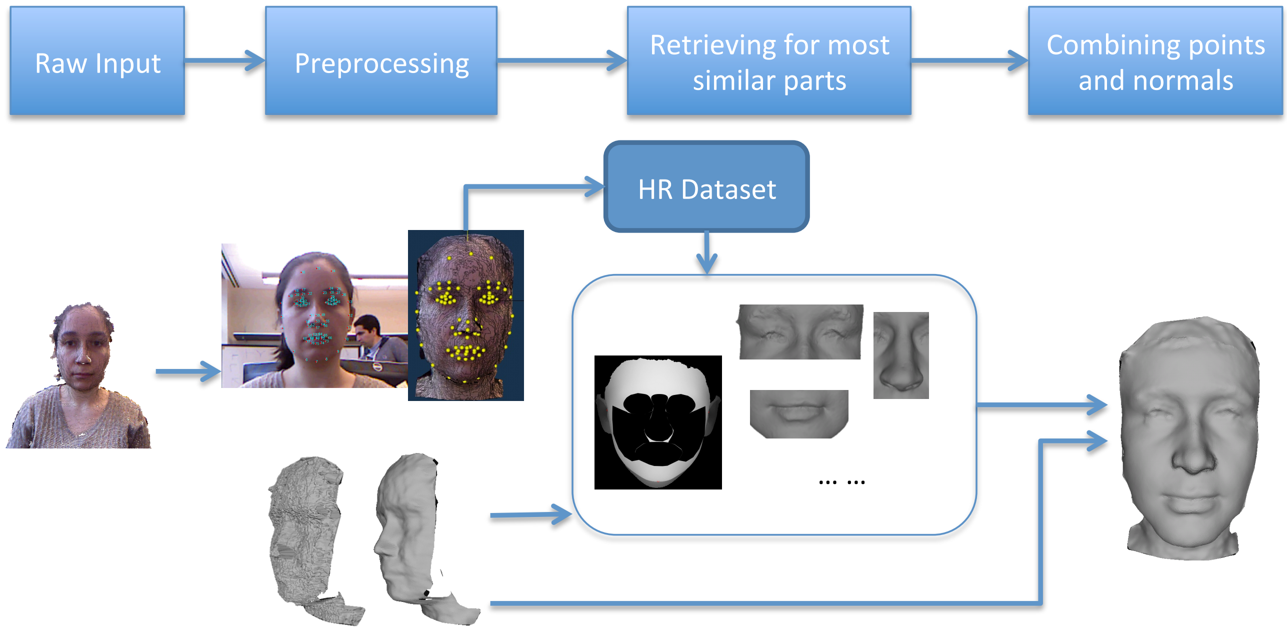

Figure 2: *Overview of the pipeline*

5 Preprocessing of Kinect raw input

Our subject is required to stay right in front of Kinect and keep a neutral expression for a few seconds with pose nearly normalized. Then a single RGBD frame is collected and sent to our reconstruction system.

The purpose of preprocessing Kinect raw input is to a) align it to our dataset b) generate facial component masks c) initialize it for HR facial components retrieval.

Landmark Labeling Since the Kinect raw input O is noisy and consists of hole regions, it is not easy to generate 3D landmarks directly from 3D shapes. We then apply 2D face alignement method to the color map to get 2D landmarks and transfer them to 3D space. Any face alignment method can be used in this step, here we use FacePlusplus [Inc. 2013] to generate 2D landmarks as well as extracting face region out. These landmarks are transferred into 3D space using caliberated camera projecton matrix of Kinect. These 83 landmarks include 19 contour points and 64 facial points. Since facial contour points depend on the face pose and can not always be correspondent on 3D meshes, we just use 15 anatomical 3D landmarks P_{15} for pose alignment as shown in Fig. 3.

Mesh Smooth Because of the inaccuracy of Kinect depth sensor, the reconstructed mesh O is very noise and it is hard to infer too much human feature from that. To retrieve for similar high resolution facial components, a mesh smooth method is performed on Kinect raw input O. Rather than do the filtering on 2D depth map, we apply curvature flow smooth method directly on O. The curvature flow smooth method is good at keeping low frequency shape meanwhile smoothing out high frequency noises. Note the Figure in Fig. 4. After the curvature flow smooth, we can see the general shape of the face more clearly.

3D Mesh Alignment Although the RGBD image is taken from the frontal view, there are still some rotations in the subject's head pose. Given P_{15} and the corresponding landmarks on G, Procrustes

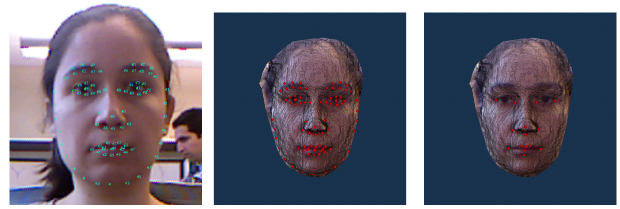

Figure 3: *Landmarks Labeling.(a) shows the 83 landmarks detected on 2D color image. (b) shows the transferred landmarks on 3D mesh O. (c) shows the 15 anatomical landmarks used for face alignment and deformable registration*

Analysis is used to align the original mesh to our dataset. To get dense correspondence from S to G, we use deformable registration method proposed in [Allen et al. 2003]. The energy to be minized for dense matching is defined as

$$E = \alpha E_d + \beta E_s + \gamma E_m \tag{1}$$

where E_d represents the data error, E_s represents the smoothness error and E_m is the landmarks error between our 15 3D landmarks P_{15} on S to the landmarks on G. The process is iterative. In early stages, the landmark error E_m contributes more to the global optimization. As the process moves on, the data error E_d dominates the optimization.
Once the deformable registration is done, the deformed generic shape G', which minimizes the point difference to S, is obtained. We can generate a mask for each facial component on S transferred from G.

Warpped HR Dataset Although our dataset consists of over 1,000 3D face meshes, it is not necessary that any of our Kinect input low resolution head can find the exact similar facial components

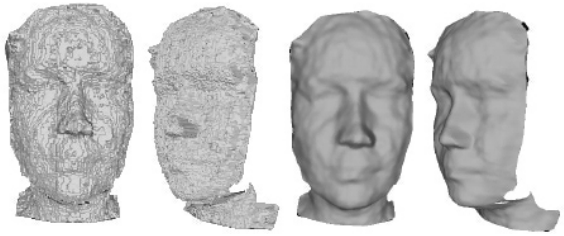

Figure 4: *Raw Kinect input mesh and smoothed mesh(frontal view and side view)*

within our dataset. To be more general, all 3D face meshes in our dataset are warpped towards S according to the 83 landmarks. Since the landmarks on S can be quite inaccurate in z-direction, we just warp our dataset using x, y-coordinate of the landmarks. Here, RBF method is used to warp all face meshes in the high resolution dataset.

6 Retrieval for the most similar HR facial components

The distance from a query facial component to any the corresponding facial component in our dataset can be defined as

$$Dist(H_q, H) = Dist(L_q, L) + \alpha Dist(hist_q, hist) \quad (2)$$

where $Dist(L_q, L)$ reprenset the distance of pseudo-landmarks and $Dist(hist_q, hist)$ is the distance between two Elevation-Azimuth Histograms.

Pseudolandmarks are a set of 3D landmarks that cover the entile facial part as shown in Fig.5. Pseudo-landmarks can be viewed as a downsampling result of face shape, which cover the whole face region and represent the general shape of one's face.

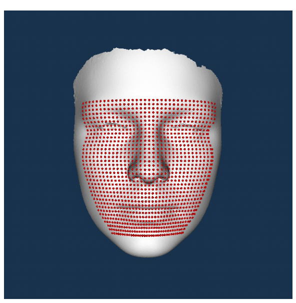

Figure 5: *Pseudolandmarks on our generic shape G*

We calculate pseudo-landmarks for smoothed Kinect mesh S using the method proposed in [**?**]. They proposed a very simple, but effective method that computes pseudo-landmarks by cutting through each 3D head mesh with a set of horizontal planes and extracting a set of points from each plane. The method starts with a 3D head mesh pose-normalized to face front. It computes two anatomical landmarks, the sellion and chin tip fully automatically and constructs horizontal planes through these points. With these two planes as base planes, it constructs m parallel planes through the head and from each of them samples a set of n points. In our retrieval, we set m as 33, n as 35. Then, 35×35 pseudo-landmarks are computed to represent a face, we will present that such resolution is enough for representing face shapes in the following sections. Pseudo-landmarks are computed on the whole face region for all the warpped HR 3D mesh and smoothed input mesh S. The similarity score for each warpped HR mesh compared to query mesh S is

$$Dist(L_q, L) = \sum_{i=1}^{1225} ||Pq_i - P_i|| \quad (3)$$

where $P_i = (x_i, y_i, z_i)$ represents a xyz-coordinate of a pseudo-landmark on a HR mesh and Pq_i is the coordinate for S.

2D Elevation-Azimuth Histogram Given the normal vector $n(n_x, n_y, n_z)$ at a 3D point, the azimuth angle θ is defined as the angle between the positive x-axis and the projection of n to the xy plane, n'. The elevation angle ϕ is the angle between x-axis and n.

$$\theta = arctan(\frac{n_z}{n_x}), \phi = arctan(\frac{n_y}{\sqrt{(n_x{}^2 + n_z{}^2)}}) \quad (4)$$

Here $\theta \in [-\pi, \pi]$, $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

With azimuth and elevation angles, any unit vector in 3D space can be defined. With a 2D histogram, the orientation variaty of a surface can be represented according to the normal direction of all its vertices. On relatively flat regions of a head, all point normal vector point in the same direction, which will cause a strong signal in some bins of the 2D histogram as show in Fig. 6. In our dataset, a 32×32 histogram is used for each facial component. So each facial component can be represented by a 32×32 matrix M. Then the angle distance E_{angle} can be defined as

$$E_{angle} = ||M - M_q|| \quad (5)$$

Image for A-E HISTOGRAM

Figure 6: *Elevation-Azimuth Histogram of a 3D face*

Pseudo-landmarks and 2D Elevation-Azimuth histogram can be computed for the whole face region as well as for separate facial components. Rember that we divide the whole face region into 5 components: nose, eyes, mouth, left cheek and right cheek. We retrieve for the most similar part for them separately. For nose region, where the depth of each point varies a lot, we choose a small α value. For eyes and mouth regions, where normal angles vary a lot, we enlarge α value to let the normal angle feature contributes more.

7 Combining HR details with Kinect input

Once we get the most similar HR facial components for each facial part, we can use the normal information from the HR parts to improve our smoothed mesh S.

Now that all faces in our dataset are in dense correspondence to the deformed generic model G', we simply copy the vertex normals from HR facial components to replace the original normals of G' part by part. Although G' is deformed to minize the shape difference to S, the mesh structures are different. Simply transferring HR normals by copying the normal of nearest neighbour will cause more smoothed result. For each vertex V in S, we search for the nearest triangle $\triangle ABC$ in G', the normal vector of V can be interplated as the weighted combination of the normal direction of $\triangle ABV$, $\triangle VBC$,$\triangle VCA$.

After we compute new normals for each vertex in S, we fuse the depth information and normal information together using the method in [**?**].

Fig. 7 shows the difference of directly copying normals from G' and using interplated normals. The directly copy method will produce more smoothed regions.

Image to be generated.

Figure 7: *Different results with normals copied in different ways*

8 Experiment and results

8.1 Similarity measurement

To valiadate our similarity measurement, we add 7 more 3dMD head meshes to our dataset and retrieve for the most similar facial components using their Kinect smoothed mesh S.

We compared psedo-landmarks, 2D Elevation-Azimuth Histogram at different resolutions as well as our final combined similarity metric. $S1 - S7$ represent 7 subjects. Each number presents the ranking(the lower, the better) of the groundtruth in the retrieval resul. If any groundtruth facial part ranks first, it means that this metric picks the groundtruth as the best matching one. It still makes sense if the groundtruth doesn't rank first but the ranking is in top 10% because there are other similar facial components from another person's face.

The retrieval result for nose region and left cheek region is in Table. 1 and 2. Note that increasing the resolution of pseudo-landmarks will not have significant effect on the final result and our combined metric works best in most of the cases.

Table. 3 shows the retrieval result for mouth region. We can see that, for single similarity metrics, Elevation-Azimuth 2D histogram works better than pseudo-landmarks. It makes sense because in mouth region, the depth information does not vary a lot.

The retrieval result for eyes region is not so good compared to the results for other regions because Kinect actually produces really bad raw data in eye regions. But since we have 3D landmarks to warp all the faces in our dataset towards the input. The eye shape will be deformed to be similar to the person.

8.2 Reconstruction result

We compared our reconstruction result with Kinect fusion and groundtruth. Because our groundtruth heads are wearing caps, we just calculate the normal errors of facial regions. Note that here, we use normal angle differences to measure the distances between the reconstruction result to groundtruth.

9 Conclusion

Acknowledgements

To Robert, for all the bagels.

References

ALLEN, B., CURLESS, B., AND POPOVIĆ, Z. 2003. The space of human body shapes: reconstruction and parameterization from range scans. In *ACM Transactions on Graphics (TOG)*, vol. 22, ACM, 587–594.

INC., M., 2013. Face++ research toolkit. www.faceplusplus.com, Dec.

Table 1: *Retrieval Result for Nose Region*

Similarity Metrics	S1	S2	S3	S4	S5	S6	S7
PseudoLandmarks35x35	157	2	809	1	14	1	58
PseudoLandmarks65x65	157	2	813	1	14	1	38
E-A 2D histogram32x32	24	7	1	33	99	238	9
Combined metric	**14**	**1**	**3**	**2**	**14**	**1**	**2**

Table 2: *Retrieval Result for Left Cheek Region*

Similarity Metrics	S1	S2	S3	S4	S5	S6	S7
PseudoLandmarks35x35	17	64	88	**64**	49	3	89
PseudoLandmarks65x65	17	76	83	70	47	3	83
E-A 2D histogram32x32	229	98	47	314	334	11	38
Combined metric	**12**	**16**	**6**	68	**22**	**5**	**31**

Table 3: *Retrieval Result for Mouth Region*

Similarity Metrics	S1	S2	S3	S4	S5	S6	S7
PseudoLandmarks35x35	229	408	441	73	22	619	342
PseudoLandmarks65x65	227	382	478	90	22	581	276
E-A 2D histogram32x32	27	108	1	119	17	95	262
Combined metric	**20**	**94**	**1**	**60**	**2**	**83**	**229**

Table 4: *Retrieval Result for Eyes Region*

Similarity Metrics	S1	S2	S3	S4	S5	S6	S7
PseudoLandmarks35x35	92	57	543	43	102	351	475
PseudoLandmarks65x65	90	67	544	56	103	395	429
E-A 2D histogram32x32	184	617	484	713	334	11	231
Combined metric	**47**	**226**	**482**	**210**	**75**	**12**	**75**

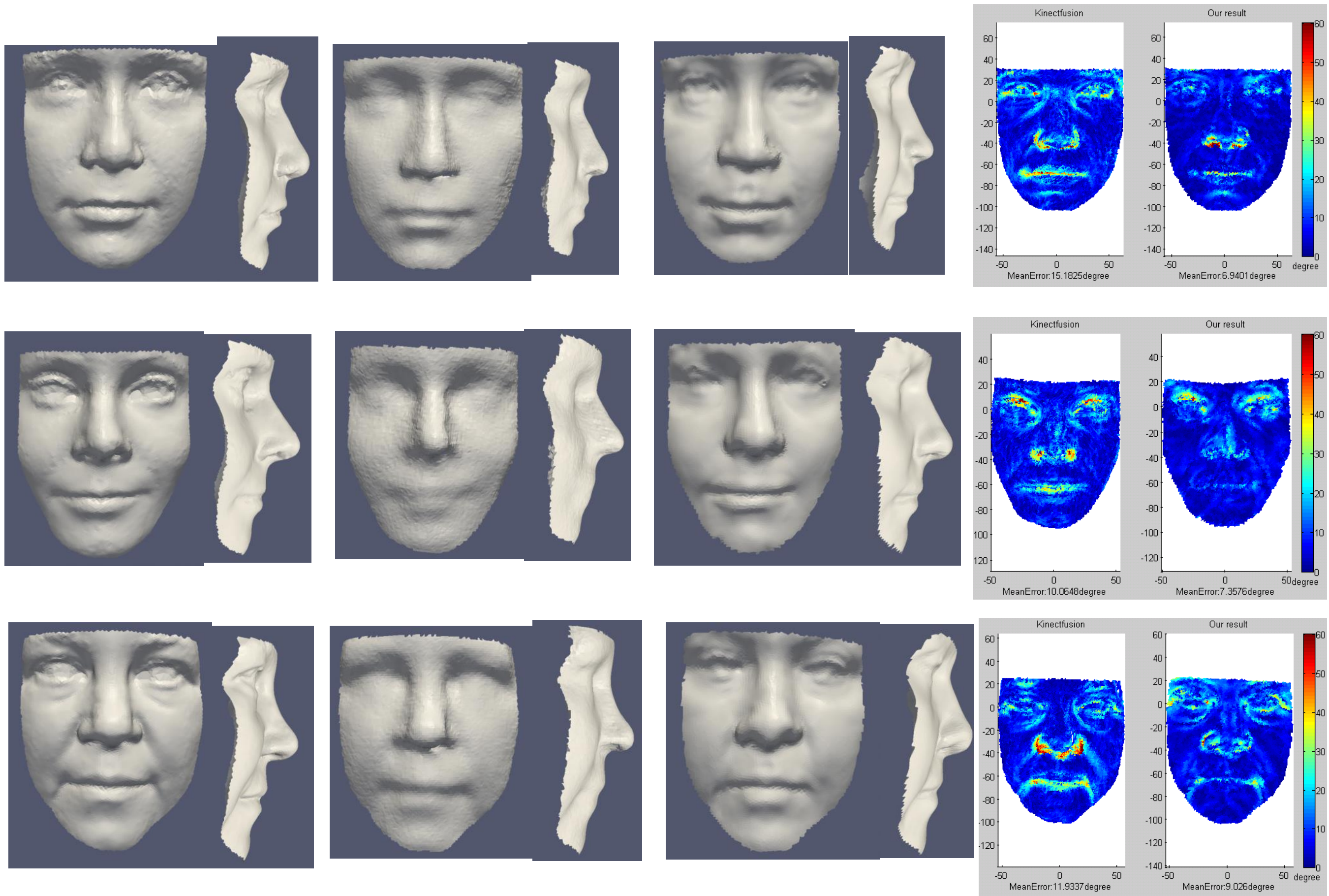

Figure 8: *Our reconstruction results compared with Kinect fusion and groundtruth. The first column is groundtruth. The second column is Kinect fusion result. The third column is our reconstruction result. We also show the normal error in column 4.*

JOBSON, D. J., RAHMAN, Z., AND WOODELL, G. A. 1995. Retinex image processing: Improved fidelity to direct visual observation. In *Proceedings of the IS&T Fourth Color Imaging Conference: Color Science, Systems, and Applications*, vol. 4, The Society for Imaging Science and Technology, 124–125.

LANDIS, H., 2002. Global illumination in production. ACM SIGGRAPH 2002 Course #16 Notes, July.

PARK, S. W., LINSEN, L., KREYLOS, O., OWENS, J. D., AND HAMANN, B. 2006. Discrete sibson interpolation. *IEEE Transactions on Visualization and Computer Graphics 12*, 2 (Mar./Apr.), 243–253.

PELLACINI, F., VIDIMČE, K., LEFOHN, A., MOHR, A., LEONE, M., AND WARREN, J. 2005. Lpics: a hybrid hardware-accelerated relighting engine for computer cinematography. *ACM Transactions on Graphics 24*, 3 (Aug.), 464–470.

WILAMOWSKA, K., SHAPIRO, L., AND HEIKE, C. 2009. Classification of 3D face shape in 22q11. 2 deletion syndrome. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, IEEE, 534–537.

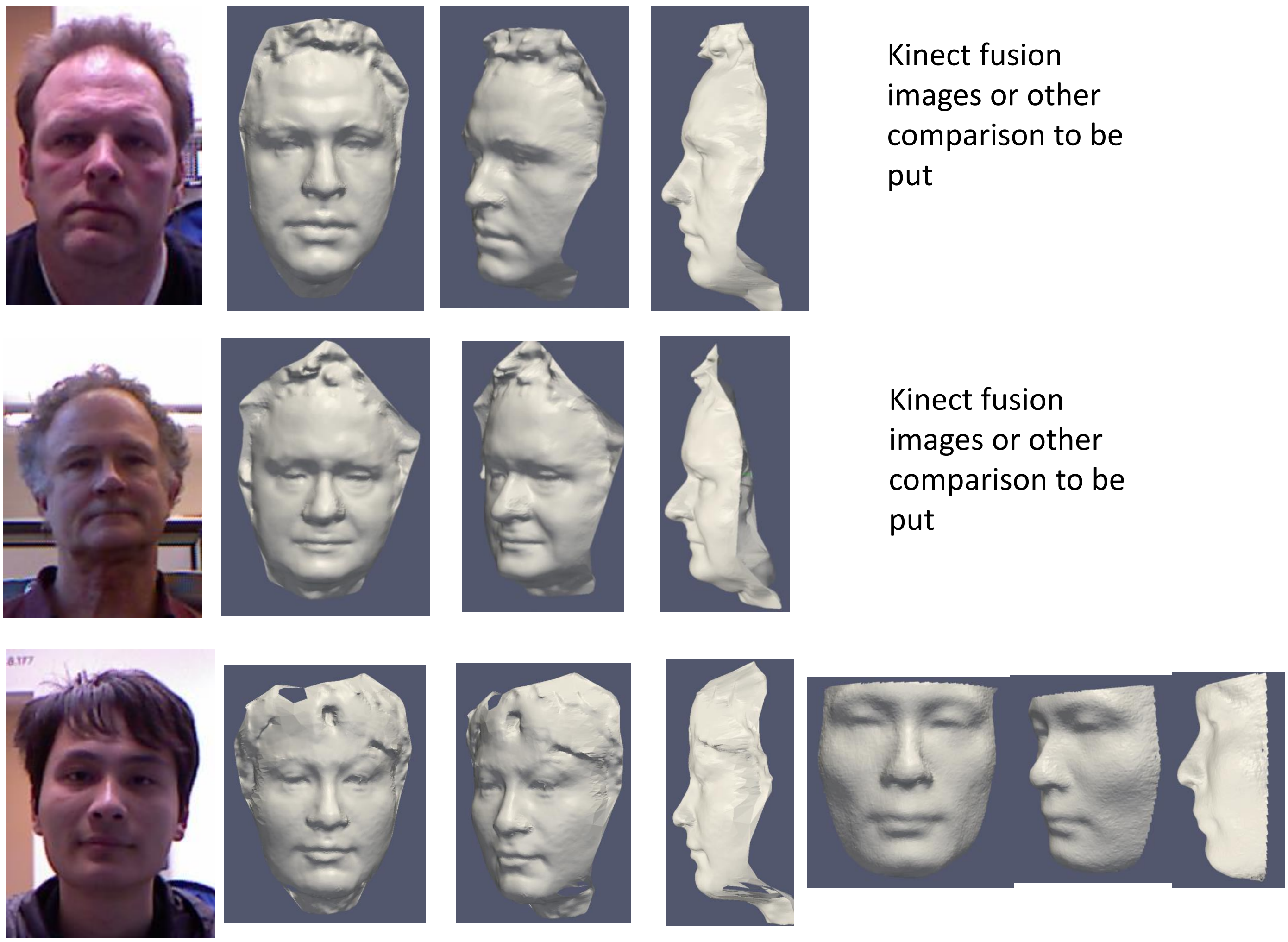

Figure 9: *More reconstruction result is shown.*