Kinect-Based Automatic 3D High-Resolution Face Modeling

Qi Sun[1], Yanlong Tang[1], Ping Hu[2], and Jingliang Peng[2†]

Taishan College, Shandong University, School of Computer Science and Technology, Shandong University

Emails: nowhereman.sq@gmail.com, yanlongtang@gmail.com, peggyhu0315@gmail.com, jingliap@gmail.com

***Abstract*—Microsoft Kinect can be used to capture both depth and color information and has been increasingly used for 3D modeling purposes. However, prior facial modeling methods either are computationally intensive or they generate rough results limited by the low resolution and instability of Kinect. In this paper, we propose a novel scheme for automatically and efficiently constructing a life-like textured 3D high-resolution model for the face of any user in front of a Kinect. Specifically, this scheme is composed of a sequence of steps including head region segmentation, depth and color image registration, resolution enhancement and 3D model fairing. Compared to prior methods, our scheme has a set of distinctive advantages. It can be robust even when the user is in a noisy environment; all the processes are automatic, which means that users need not interactively select feature points, and the energy optimization step is more efficient for fast processing of large-scale dynamic images.**

***Keywords*—Kinect; super-resolution; face modeling**

I. INTRODUCTION

Microsoft Kinect has become one of the most popular 3D sensors for both academic and entertainment uses because of its versatility, portability and affordability. It can be used to capture both depth and color information at the same time and has been increasingly used for 3D modeling purposes in recent years. However, its low resolution, which is up to 640 ×480 of depth image still limits the smoothness and vividness for high resolution 3D modeling tasks and, in particular, high resolution 3D facial modeling.

Techniques on 3D modeling, especially 3D facial modeling, have been extensively investigated. Abate, et al. [1] and Blanz and Vetter [2] presented a morphable model to reconstruct a 3D face from 2D images of different parts of a human face and it was extended for Kinect by Zollhöfer, et al. [3]. However, this method has to search from an existing database and hence, it is hard for the reconstructed model to be realistic and unique enough for a particular user.

Weise, et al. [4] used a data alignment algorithm to model an avatar face. However, the effect is limited by the low resolution of Kinect images and it can only work well in static and tidy indoor environments. Otherwise, the face cannot be detected correctly. By contrast, our method for face region detection works well even in a noisy room, as shown in Figure 3(a).

Although Pajdla, et al. successfully enhanced the resolution and smoothness of the data captured by Kinect [5], their method needs two Nikon D60 cameras, which increases the system expenses.

Tong, et al. [6] and Cui and Stricker [7] presented two different methods to reconstruct the full human body. But the resolution is not promoted. Furthermore, both their frameworks require a special environment such as a rotating disk or a semi-circle orbit, which limits the applicability of their method for common house use. By contrast, our method does not need any special device, making it well suited for common house use.

To overcome the problems as mentioned above with the previous works, we introduce a novel resolution enhancement framework to construct face models based on the raw data captured by Kinect. Compared to those of other traditional modeling methods, our results are smoother and more lifelike with a higher resolution and more facial details.

Figure 1. A Microsoft Kinect

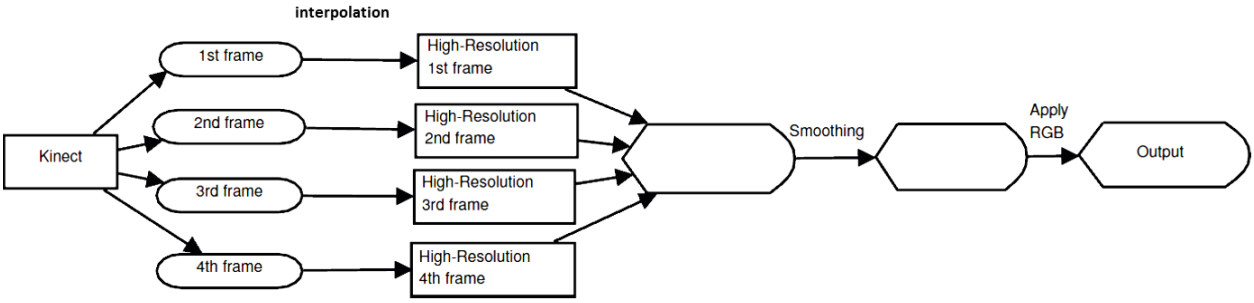
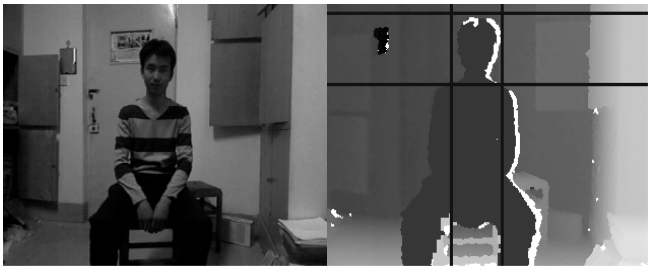


Figure 2. The pipeline of the proposed scheme.



(a) (b)

(c)

Figure 3. (a) The RGB data captured by Kinect; (b) The rectangle in the depth data containing the face; (c) The segmented 3D face depth data captured by Kinect.

II. Our Scheme

The proposed scheme is composed of several main steps: data acquisition and head region segmentation, pre-processing, combination and smoothing. The pipeline of the proposed scheme is given in Figure 2.

A. Data Acquisition and Head Region Segmentation

Kinect is able to detect the distance of the entire scene with the distance between 85cm and 4000cm from itself and returns in real time a 640×480 range image and a corresponding RGB image using a pair of IR projector and camera.

We put one Kinect device (see Figure 1) on a desk to capture the depth data. A user should sit in front of it, as shown in Figure 3(a), and rotate his or her head within a valid range.

To segment the face of the user from the range image captured by Kinect, we use the feature in the official Microsoft Kinect SDK on detecting the human body and joints in depth data. Our method is based on the assumption that the user's head is located in the area between the left shoulder and the right shoulder. Specifically, we build a rectangular area, R, as defined below and assume that R contains the face.

$$R = \left\{ (x, y) \mid x \in [LS.x, RS.x], y \in [V.y, 2 \times H.y - V.y] \right\}$$

Here, LS, RS, V and H stand for left shoulder, right shoulder, vertebrae and head, respectively. This face segmentation process is illustrated in Figure 3(b).

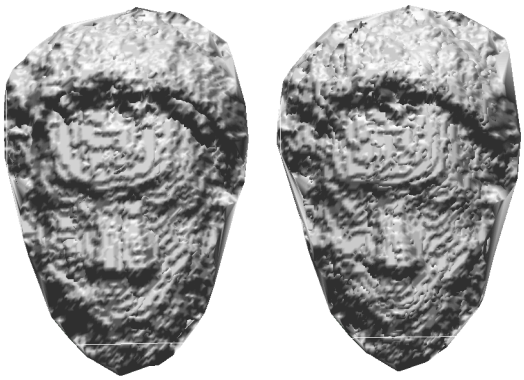
After that, we need to accurately segment the face region from this rectangular area. Specifically, denoting the depth of the head joint as $H.z$, we recognize any pixel with a depth under $H.z + \Delta$ as a facial pixel, where Δ is an adjustable system parameter. Compared to the "flood-fill-like method" of Zollhöfer, et al. [3], our method can be effective even in a noisy and messy indoor environment and the user does not have to rotate his or her head at a constant speed. That is to say, an approximately static or high-speed rotating head can be modeled equally well. After this process, a 128×128 resolution face can be segmented from the whole scene, as illustrated in Figure 3(c).

B. Pre-Processing

To use the captured depth information as input of our algorithm, a pre-processing step should be applied. We combine four neighboring original frames as a group to construct one high-resolution face. As the four frames in a given group are captured at different time instances, we should first align them to the central one. For that purpose, we use a popular method, iterative closest point (ICP) algorithm [8], to register them (see Figure 4).

After the alignment of the adjacent four frames, each of the aligned frames should be enhanced into a high resolution (512 × 512) depth image via bilinear interpolation, as shown in Figure 5. However, pure resolution enhancement does not yet provide lifelike

visual of the new facial data. In order to obtain a more realistic facial model, we combine these four high-resolution frames to create an exquisite and smooth one by energy function minimization.



(a) (b)

Figure 4. (a) The 1st and 20th frames combined before using ICP; (b) The 1st and 20th frames combined after using ICP.

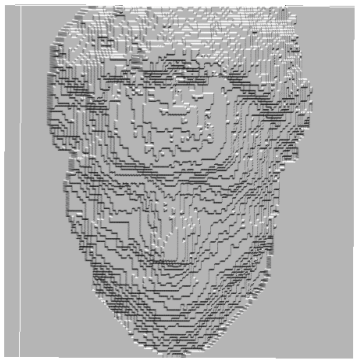


Figure 5. Result using Bilinear interpolation. Although the resolution is high, it is not acceptable.

C. Combination and Smoothing

In order to acquire a better 3D facial model of the user from four interpolated frames, we solve the following optimization problem,

$$\hat{X} = \arg\min \sum_{i=1}^{4} \left\| X - X_i \right\| \quad (1)$$

where X is the target depth image and Xi is the i-th interpolated frame. The rationale is that each frame in a group contains particular feature information which may differ from those contained in the other frames, and we should try to integrate all the features while reducing the noise in each frame caused by the Kinect hardware. This is achieved through the energy function as defined in Equation (1). It should be noted that this formula is different from that in Lidarboost [9], one of the most popular 3D super-resolution method, in that the second regular term

$$E_{regular} = \sum_{u,v} \left\| \nabla X \right\|_2 \quad (2)$$

is not considered in our energy function due to its high computational cost. Instead, we employ a smoothing algorithm that will be described below. The smoothing algorithm plays a similar role and yields very similar results as the regular term used by Lidarboost [9] .

The minimization problem is solved by Sedumi [10]. This step finally adds important facial information such as eyes and mouth to the resultant model. Furthermore, in order to smooth the resultant face model, we apply the Laplacian algorithm [11] on it:

$$\bar{x}_i = \frac{1}{N} \sum_{j=1}^{N} x_j \ . \quad (3)$$

Here, we choose the number of adjacent points N=25, which represents two square neighboring rings of a given pixel in the depth image, as shown in Figure 6.

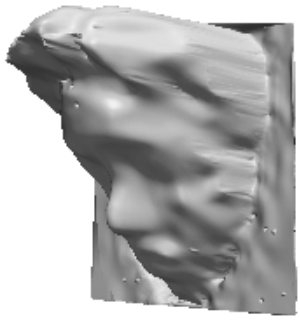
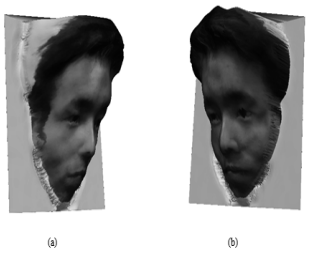


Figure 6. The smoothed face model

Finally, based on the correspondence of the RGB images and the depth images returned by Kinect [12], we add the RGB texture to the smooth 3D face model to make it more attractive and lifelike (See Figure 7).



(a) (b)

Figure 7. The right and left sides of our 3D face model.

III. EXPERIMENTAL RESULTS

In our experiments, the human subject sits in front of a Kinect device in a messy room while he rotates his head freely. For the facial region segmentation as described in Section IIA, we set the parameter Δ to 20, which successfully segments the face region from the background and the neck, as shown in Figure 3(c). However, it should be noted that different Δ values may be needed for different people and therefore, a quick

trial-and-error process may be needed for a specific subject to determine the best Δ value.

As can be observed from Figure 8, the reconstructed textured 3D facial model looks very vivid and similar to the real face image captured by a camera. The whole reconstruction process is fully automatic and no user intervention is needed. All the messiness of the background was automatically filtered out and did not affect the result.

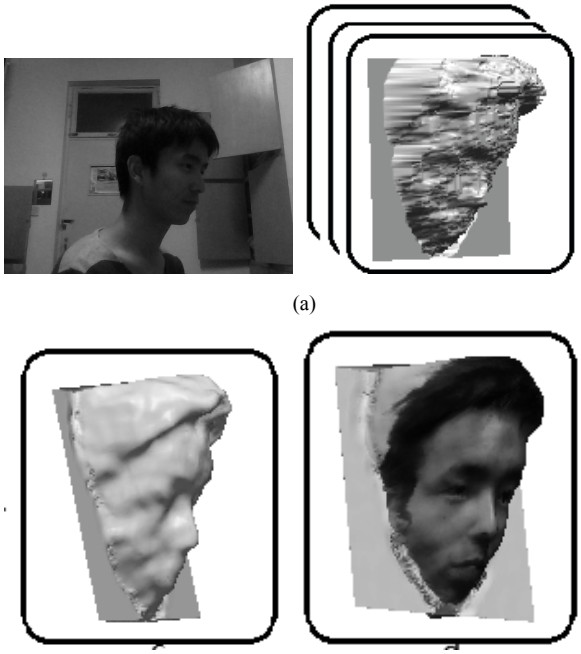


Figure 8. (a) The input face and original 3D signal; (b) The output 3D High-Resolution face model with texture.

IV. Conclusion

In this paper, we proposed a novel framework to generate a lifelike and high-resolution 3D facial model using Microsoft Kinect. We have successfully overcome the shortcomings of previous works, such as computational complexity, roughness caused by the low resolution and/or unsuitability for messy indoor environments. Moreover, the whole process does not require the user to do any facial landmark selection. Therefore, the proposed scheme is fully automatic and good for real-time 3D facial modeling. The proposed scheme should be potentially used in applications such as personal avatar generation and game character design.

† Author of correspondence

Acknowledgment

This work was supported by Shandong Provincial Natural Science Foundation, China (ZR2011FZ004), the Program for New Century Excellent Talents in University (NCET) in China and the National Natural Science Foundation of China (61070103 and U1035004).

References

[1] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, G. "2D and 3D face recognition: A survey," *Pattern Recognition Letters,* vol. 28, no. 14 (Oct.), pp. 1885-1906, 2007.

[2] V. Blanz, and T. Vetter. "A morphable model for the synthesis of 3D faces," *SIGGRAPH*, pp. 187-194, 1999.

[3] M. Zollhöfer, M. Martinek, G. Greiner, M. Stamminger, andJ. Sussmuth. Automatic reconstruction of personalized avatars from 3D face scans. Journal of Visualization and Computer Animation 22, 2-3, 195–202, 2011.

[4] T. Weise, S. Bouaziz, H. Li, and M. Pauly. "Realtime performance-based facial animation," *ACM Trans. Graph,* vol. 30, no. 4, p. 77, 2011.

[5] J. Smisek, T. Pajdla, and M. Jancosek. "3D with Kinect," In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE Computer Society, Los Alamitos, USA, A. Fossati, J. Gall, G. Helmut, X. Ren, and K. Konolige, Eds., 2011, pp. 1154-1160. CD-ROM.

[6] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. "Scanning 3D full human bodies using Kinects," *IEEE Trans. Vis. Comput. Graph*, vol. 18, no. 4, pp. 643-650, 2012.

[7] Y. Cui, and D. Stricker. 2011. "3d body scanning with one Kinect," In 2nd International Conference on 3D Body Scanning Technologies, o.A.

[8] Z. Zhang. "Iterative point matching for registration of freeform curves and surfaces," *International Journal of Computer Vision,* vol. 13, no. 2, pp. 119-152, 1994.

[9] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. "Lidarboost: Depth super-resolution for toF 3D shape scanning," In *CVPR*, pp. 343-350, 2009.

[10] J. F. Sturm. "Using SeDuMi1.02: A Matlab Toolbox for Optimization for Optimization over Symmetric Cones. Ontario: McMaster University, 1998.

[11] G. Hansen, R. Douglass, and A. Zardecki. *Mesh enhancement: Selected elliptic methods, foundations and applications.* London: Imperial College Press, 2005.

[12] D. C. Herrera, J. Kannala, and J. Heikkila. 2011. "Accurate and practical calibration of a depth and color camera pair," Retrieved from http://www.ee.oulu.fi/dherrera/kinect/2011-depth_calibration.pdf.