2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Expression Invariant 3D Face Modeling from an RGB-D Video

Donghyun Kim, Jongmoo Choi, Jatuporn Toy Leksut and Gérard Medioni
Institute for Robotics and Intelligent Systems,
University of Southern California,
3737 Watt way PHE 101, Los Angeles, CA 90089, United States
{kim207, jongmooc, leksut, medioni}@usc.edu

***Abstract*—We aim to reconstruct an accurate neutral 3D face model from an RGB-D video in the presence of extreme expression changes. Since each depth frame, taken by a low-cost sensor, is noisy, point clouds from multiple frames can be registered and aggregated to build an accurate 3D model. However, direct aggregation of multiple data produces erroneous results in natural interaction (e.g., talking and showing expressions). We propose to analyze facial expression from an RGB frame and *neutralize* the corresponding 3D point cloud if needed. We first estimate the person's expression by fitting blendshape coefficients using 2D facial landmarks for each frame and calculate an expression deformity (expression score). With the estimated expression score, we determine whether an input face is neutral or non-neutral. If the face is non-neutral, we proceed to neutralize the expression of the 3D point cloud in that frame. To neutralize the 3D point cloud of a face, we deform our generic 3D face model by applying the estimated blendshape coefficients, find displacement vectors from the deformed generic face to a neutral generic face, and apply the displacement vectors to the input 3D point cloud. After preprocessing frames in a video, we rank frames based on the expression scores and register the ranked frames into a single 3D model. Our system produces a neutral 3D face model in the presence of extreme expression changes even when neutral faces do not exist in the video.**

I. INTRODUCTION

We aim to produce an accurate 3D face model in the presence of extreme expression changes for non-cooperative subjects. Due to the development of low-cost depth cameras, such as Kinect [1], prior researches demonstrate how to reconstruct an accurate 3D model with a low-cost depth sensor. Since only a single raw frame is not enough to make an accurate 3D model, due to the noisy input, multiple depth frames are needed to be registered using iterative closest point (ICP) [2], [3] to make an accurate 3D model [4]–[6]. Recently, Hernandez *et al.* [4] show that a near laser scan quality 3D face model can be reconstructed by registering multiple depth frames.

The ICP algorithm is widely used for registration of two point clouds by minimizing the difference of the two point clouds [3]. The key limitation of using standard ICP, which requires two rigid point clouds, for 3D face modeling is that the person should keep the same expression and stay as still as possible until a 3D model is reconstructed. However, in natural environments, people can change their expressions such as talking and smiling. Especially, reconstruction of an accurate model is challenging when the mouth opens due to large expressions since the topology of the face is changed.

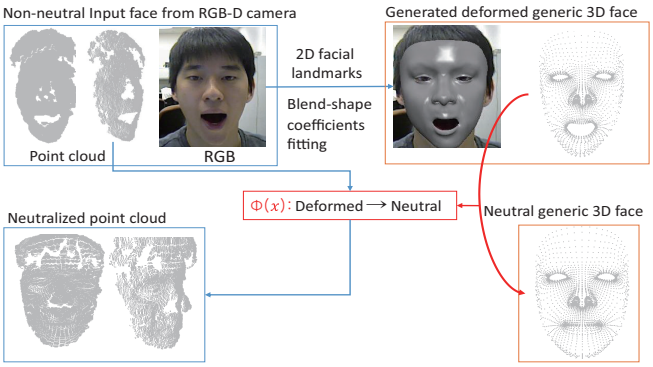


Fig. 1. Overview of the proposed neutralization method.

In this paper, we propose a method that aims to make an accurate neutral 3D face model from a sequence of RGB-D frames containing drastic expression changes. Although we only produce a neutral 3D face model with many uses (e.g., face recognition), this method can be applied to generate other non-neutral 3D face models as well.

Given an RGB frame in a video, we extract 2D facial landmarks, estimate the facial expression parameters, and compute the expression deformity (expression score). The expression parameters are coefficients of the blendshapes. If the level of deformity represents non-neutral, we neutralize the non-neutral face, as shown in Fig. 1. We first generate a deformed generic face based on the estimated expression parameters. We then find a set of displacement vectors from points in the deformed generic face to the corresponding points in the neutral generic face. The displacement vectors are applied to the point cloud of the input face in order to neutralize the non-neutral face. At the end of video, we rank all frames based on the expression scores and register the ranked frames sequentially. Fig. 2 illustrates our approach.

Our contributions are as follows:

- We estimate each input face's expression parameters with 2D facial landmarks by fitting coefficients of 3D blendshapes, score a deformity of the expression, and neutralize the 3D point cloud of the input face depending

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

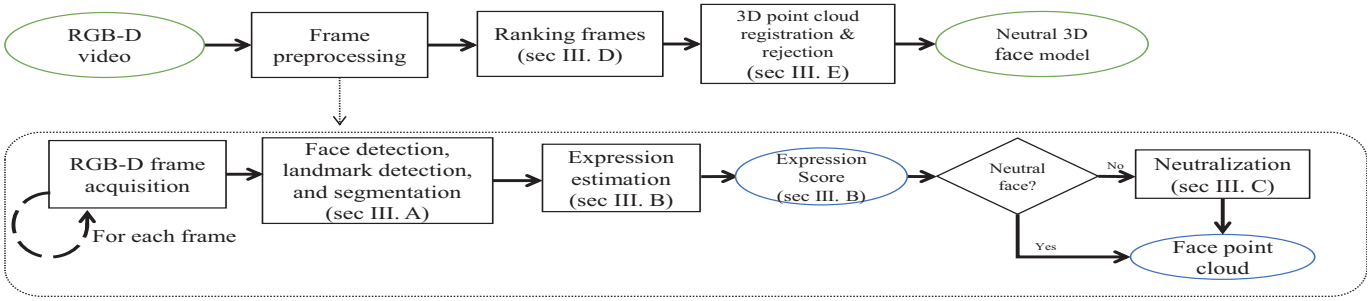


Fig. 2. Overview of the proposed system.

on the estimated deformation.

- We rank frames based on the estimated expression scores in order to smooth drastic changes. We sequentially aggregate the ranked frames of neutral faces or neutralized non-neutral faces to make an accurate neutral 3D face model.
- We can generate a neutral 3D face model even when there are only non-neutral faces in an RGB-D video.

The remainder of this paper is organized as follows. Section II summarizes the related work. Section III shows our proposed method and section IV presents experimental results. Section V concludes our paper.

II. RELATED WORK

We review 3D modeling, expression fitting, and expression deforming methods.

3D reconstruction and modeling method: We review prior methods of 3D reconstruction for static and dynamic objects with a low-cost depth camera. In here, static objects mean that it can move but do not change its shape.

Static objects. In order to make a 3D model from multiple depth frames for static objects, depth frames of different views can be aligned by using ICP registration such as rigid ICP [4], [5]. Rigid ICP makes point pairs from two different views and finds transformation by minimizing the distances of corresponding pairs. Since static objects do not differ in their shape, initial alignment of source and target matters for registration. Izadi *et al.* [5] propose the KinectFusion which provides a 3D reconstruction of a static scene by moving a Kinect sensor around with ICP based tracking. Hernandez *et al.* [4] propose a method that multiple depth frames are registered and integrated to generate a near scan 3D quality face model using rigid ICP registration. But these approaches cannot apply to non-rigid shape and need a subject to stay still.

Dynamic objects. There are several proposed methods to reconstruct a 3D model in a dynamic environment [7]–[9]. Izadi *et al.* [7] propose a method that reconstructs a static background by removing a moving object. It uses dense ICP tracking to detect the moving object. But, it cannot deal with continuous moving. Li *et al.* [9] present a system where a user can scan themselves by rotating in front of a depth camera with the same posture. The system uses global non-rigid registration method that can handle tiny movements. However, large deformation of movement causes registration to fail. Newcombe *et al.* [8] present a SLAM system which extends KinectFusion. The system reconstructs an object with a wide range of moving and scenes by non-rigidly deforming scenes. However, it is limited to a scene with slow moving and it is hard to reconstruct a motion that moves from a closed to open topology.

2D Expression fitting method: Traditional expression analysis methods focus on pixel value information extracted from an image of a frontal face. Recently, 3D models have been adopted to help analyze expressions in non-frontal faces. Chu *et al.* [10] use extended 3D Morphable Models (3DMM) with expression variations proposed by Blanz *et al.* [11] to estimate facial expressions in a still image and normalize the expressions using an image warping technique for 2D face recognition. The system minimizes the projection error between an image and parameterized 3D face model projection. Such optimization technique is known to be computationally expensive. In addition to 3DMM, Blendshape [12], a computer animation technique used to construct desired 3D models using a combination of different basis shapes, has also been adopted. Cao *et al.* [13] shows an expression regressor for videos using a tensor of identity and expression blendshapes. The expression regressor pioneered in using blendshapes to track facial expression in 2D. The system, however, requires multiple frames to stabilize expression tracking.

Expression deformation method: There are several proposed methods to deform face expression [14]–[16]. Lu *et al.* [14] get synthesized deformations by learning from control groups where each subject made 7 different expressions in order to make a deformable model. Then, they fit the deformable model to a given test scan 3D model. Al-Osaimi *et al.* [16] use a PCA for modeling the expression deformation and obtain a generic expression deformation model trained by PCA with pairs of non-neutral with neutral scans of people. Mpiperis *et al.* [15] propose a method to use a bilinear model for jointly addressing a 3D face and facial expressions. They can produce a neutralized face by manipulating expression control parameters in the bilinear model. There are also other deformation methods of non-neutral 3D face model. Most of

the methods have an assumption that accurate 3D non-neutral face models are acquired. But, in the presence of dynamic expression changes, it is hard to reconstruct any non-neutral 3D face model. Therefore, we focus on generating an accurate neutral 3D face model from dynamic expression changes even when there is no neutral face at all.

III. PROPOSED METHOD

We take a sequence of RGB-D frames from a low-cost depth camera, PrimeSense [17], and reconstruct a neutral 3D face model. We detect a face and 2D facial landmarks using dlib library [18] from an RGB frame. In order to estimate the deformation of an input face, we fit the coefficients of the blendshapes using the 2D facial landmarks and compute the expression deformity (expression score). We extract a point cloud within the face bounding box of the corresponding depth frame and neutralize it depending on the expression score. At the end of the video, we rank all point clouds with increasing order of the expression scores and register the point clouds sequentially to generate a single neutral 3D face model.

A. Face detection, 2D facial landmark detection, and segmentation

We first use the dlib [18] module to detect a face bounding box and 68 2D facial landmarks from each RGB frame. The dlib face detector uses a histogram of oriented gradients (HOG) [19] for training and a method of an ensemble regression trees [20]. We then segment a face region more precise using the 2D facial landmarks and convert the corresponding depth frame to a point cloud.

B. Face deformation estimation

Fitting expression: We estimate how much a face deforms from its neutral state by fitting n expression blendshapes, represented by $\mathbf{B} \in \mathbb{R}^{3m \times n}$, to m inferred 3D facial landmark points $\mathbf{x} \in \mathbb{R}^{3m \times 1}$. We obtain 3D expression coefficients $\hat{\mathbf{w}} \in \mathbb{R}^n$ by solving a simple linear equation

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} ||\mathbf{B}\mathbf{w} - \mathbf{x}||. \quad (1)$$

To obtain inferred 3D landmark points $\mathbf{x}$, we first estimate the 3D head pose by solving for the camera matrix using the corresponding rigid 2D and 3D landmark points. Then we revert project our detected 2D landmark points onto our pose-aligned 3D model to obtain revert-projected 3D landmark points.

Compute expression score: We simply compute $L2$-norm of $\hat{\mathbf{w}}$ to measure the face deformity. The larger the expression is (e.g., big open mouth) the higher $||\hat{\mathbf{w}}||_2$ will be.

Generate deformed generic 3D model: To generate a deformed 3D face $\hat{\mathbf{x}}$ given expression coefficients $\hat{\mathbf{w}}$, we multiply $\hat{\mathbf{w}}$ back to our full blendshape basis $\mathbf{B}_{full}$ that contains all vertices of the generic model

$$\hat{\mathbf{x}} = \mathbf{B}_{full}\hat{\mathbf{w}}. \quad (2)$$

C. Neutralization

If a computed expression score, $||\hat{\mathbf{w}}||_2$, is above a threshold (e.g., 0.35), we consider the input expression as a non-neutral expression and neutralize the corresponding point cloud. Fig. 1 shows an overview of the neutralization method. After we produce a deformed generic face from an input face, we register the deformed generic face with the input face by finding a 3D similarity transformation matrix. Then we calculate displacement vectors from the deformed generic face to the the neutral generic face. We apply the displacement vectors to the input point cloud of the face to neutralize the point cloud.

Registering a deformed generic face with an input face: The point cloud of an input face and the point cloud of the deformed generic face differ in 3D shape, number of points, and scale. We first find the 3D similarity transformation between two point clouds using the corresponding 3D facial landmarks of the input face and the deformed generic face. Since the detected boundary points can be inaccurate with a low-cost depth camera, we exclude landmarks of a jaw in order to get a more reliable transformation matrix. We use 51 facial landmark points from the input face $\{\mathbf{y}_i^{input}\}$ and the deformed generic face $\{\mathbf{y}_i^{generic}\}$. The similarity transformation matrix $[\hat{c}, \hat{\mathbf{R}}, \hat{\mathbf{t}}]$, computed as

$$[\hat{c}, \hat{\mathbf{R}}, \hat{\mathbf{t}}] = \underset{c,\mathbf{R},\mathbf{t}}{\operatorname{argmin}} \sum_{i=1}^{51} ||\mathbf{y}_i^{input} - c\mathbf{R}\mathbf{y}_i^{generic} - \mathbf{t}||_2^2, \quad (3)$$

where c is a scale factor, $\mathbf{R}$ is a 3×3 rotation matrix, and $\mathbf{t}$ is a 3×1 translation vector. We use the SVD [21] algorithm to get $\mathbf{R}, \mathbf{t}$ with varying scale c. We apply the estimated similarity transformation matrix to the generic face model, as shown in Fig. 3.

Finding displacement vectors: The generic 3D face model is a point cloud which consists of n number of 3D points (e.g., $n = 3440$). A displacement vector is a translation vector from a point of the deformed model to the corresponding point of the neutral generic face, as shown in Fig. 4. In our experiments, we compute 3440 displacement vectors from all 3440 points.

Applying displacement vectors: After we align the input face with the deformed generic face, we find the nearest point from a point of the input to the point of the deformed generic in order to get a corresponding displacement vector. For every point in the input face, we find a corresponding displacement vector and apply it to the point of the input face.

D. Ranking frames

After estimating expression scores from all frames, we rank all facial point clouds of the frames in an RGB-D video. A low score corresponds to a more neutral face. We sort the list of point clouds in order of closeness to a neutral face. This method changes the order of frames in the RGB-D video. This method makes the 3D face modeling system independent of the order of frames.

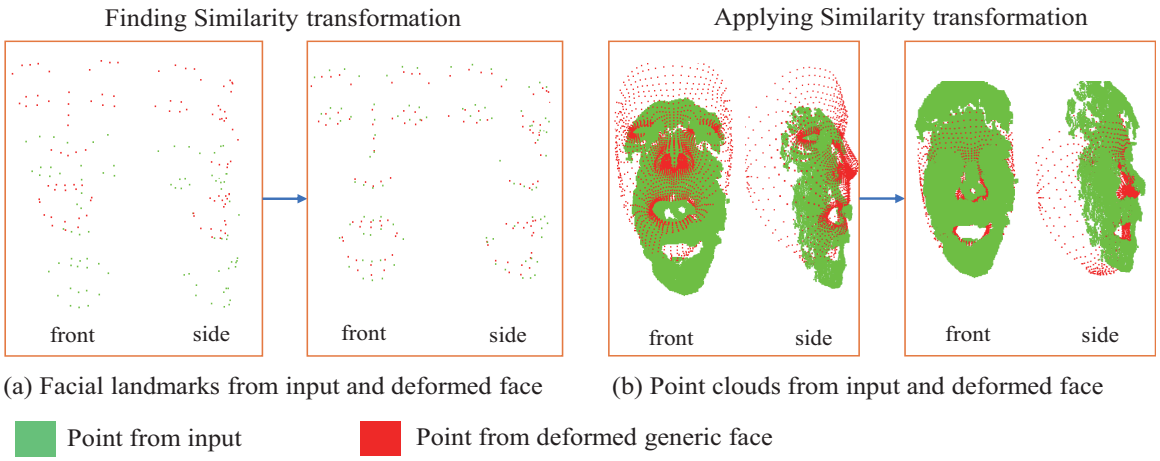


Fig. 3. An example of registering a deformed generic 3D face model to an input face. (a) We find a similarity transformation from the 51 facial landmarks and (b) apply the similarity transformation to the generic face model.

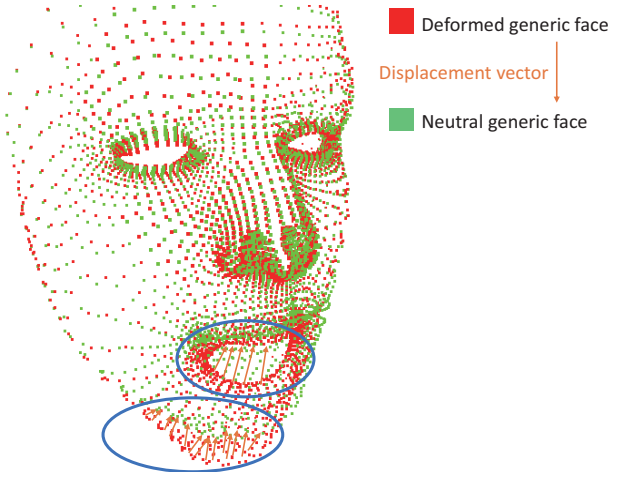


Fig. 4. Examples of displacement vectors.

E. ICP registration and rejection

To aggregate multiple facial point clouds, we use only neutral or neutralized faces instead of raw non-neutral faces. We start with the point cloud which has the lowest expression score in a video and register point clouds in increasing order of the expression scores. We use the rigid ICP method which iteratively computes a rigid transformation (rotations and translations) that minimizes the difference between the point clouds. We use the rejection strategy, described in [4], where we reject a frame if the difference between a previously reconstructed 3D model and the input is too large. We calculate the difference by averaging of the pixel-wise Euclidean distances. Because of the rejection strategy, making an accurate reconstructed model in an initial stage is more important. As the model becomes more accurate, a noisy input point cloud is more likely to be rejected. In other words, a poor registration in an initial stage is less likely to reject a noisy input point cloud.

IV. EXPERIMENTAL RESULTS

We have compared our 3D face modeling result with the state of the art 3D face modeling method (baseline) [4] on an RGB-D database containing large expression changes.

A. Data

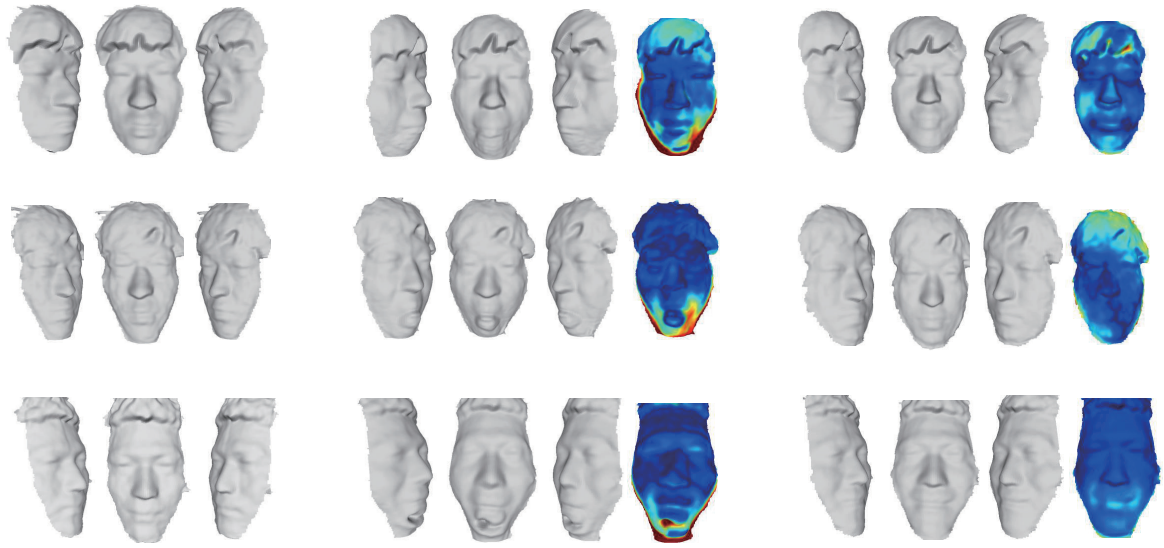
We used a fixed PrimeSense camera [17], and set its resolution 640×480. From 10 subjects, we recorded 10 videos which consist of neutral and expression parts. The neutral part contains a sequence of person's neutral faces in order to make a reference face model. The expression part contains a sequence of expression changes where a person opens the mouth at a maximum, closes the mouth again, and repeats this expression changes for 4 times. We intended to include exaggerated expressions in this database.

B. Qualitative analysis

We first set reference 3D face models (Fig. 5 (a)) reconstructed with only neutral faces. We then reconstructed 3D face models using the baseline method (Fig. 5 (b)) and the proposed method (Fig. 5 (c)) in the presence of expression changes, as shown in Fig. 5. We generated heat maps of the reconstructed 3D face models with respect to the corresponding reference models. In Fig. 5 (b), a region around the mouth shows distorted reconstruction results due to the expression changes. However, the results from our method (Fig. 5 (c)) show better results compared with the baseline method [4]. Note that a region around the mouth in the proposed method is more similar to the reference.

C. Quantitative analysis

To compare the qualities of reconstruction results of our method with the baseline method, we measure the similarities between the reconstructed 3D face models and the references as shown in Fig. 6. We use the 3D matching distance metric for face identification proposed by Min *et al.* [22]. As described in [22], we exclude unstable features (e.g. hair), register a reference with an input face using ICP, segment facial regions



(a) Reference (no expression) (b) Baseline [4] (c) Proposed method

Fig. 5. (a) is a set of reference 3D face models from only neutral faces. (b), (c) are 3D face models in the presence of expression changes.

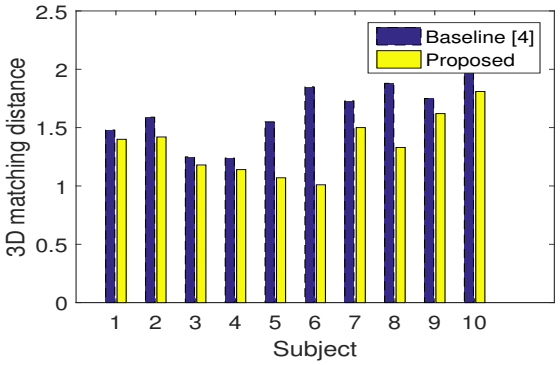


Fig. 6. 3D matching distances between a reference with only neutral faces and a 3D face model with expression changes from 10 subjects.

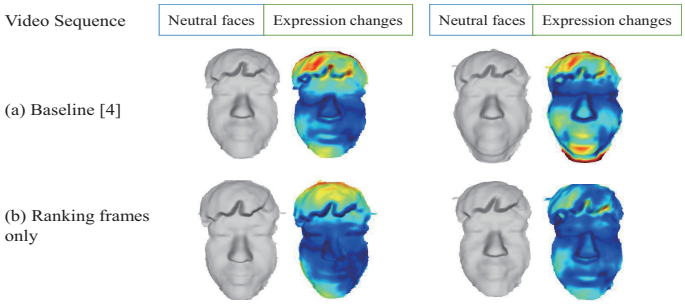


Fig. 7. 3D face model results with the two video sequences. (a) uses baseline method [4]. (b) uses ranking frames method.

for different weighting, and measure a distance between the two faces. Even though the weight for a region around a mouth is the smallest, all of the distances from the proposed method are smaller than those of the baseline method. It means that the proposed method worked well in the presence of expression changes. The proposed method outperforms Hernandez method for the 10 subjects.

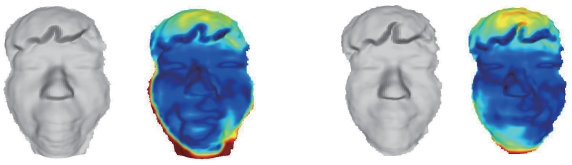
D. Further analysis of the proposed method

Since the proposed method uses both the ranking method and neutralization method, we show the effectiveness of these two methods in this section.

Ranking frames method: We use two videos. The first video consists of 30 frames neutral faces and the expression changes. The second video consists of the expression changes and 30 frames neutral faces. Due to the rejection strategy (Sec. III. E), if we can gather enough neutral faces before the expression changes, adding non-neutral faces do not change a 3D face model significantly (Fig. 7(a) left). Otherwise, adding non-neutral faces may deform the 3D face model seriously (Fig. 7(a) right). Ranking frames method selects neutral faces during expression changes and make the neutral faces be registered first. In this way, ranking frames method makes the system independent of the order of frames. As a result, ranking frames method generates a good quality of 3D face model in any order of frames (Fig. 7(b)).

Neutralization method: We use a video where there are only 16 frames of expression changes and no neutral faces. Since neutral faces do not exist in the video, ranking frames method may also fail to reconstruct a 3D face model (Fig. 8(a)). For the case of this failure, we need to neutralize non-neutral faces. Although neutralization method cannot generate the same as a real neutral input face, it still makes better neutral 3D face models (Fig. 8(b)). In short, both ranking frames and neutralization method are complementary to each other.

In order to compare improvements of each method, we mea-



(a) Ranking frames only (b) Prposed method (Ranking frames + Neutralization method)

Fig. 8. 3D face results for (a) only ranking frame method and (b) the proposed method from a video where there are no neutral faces.

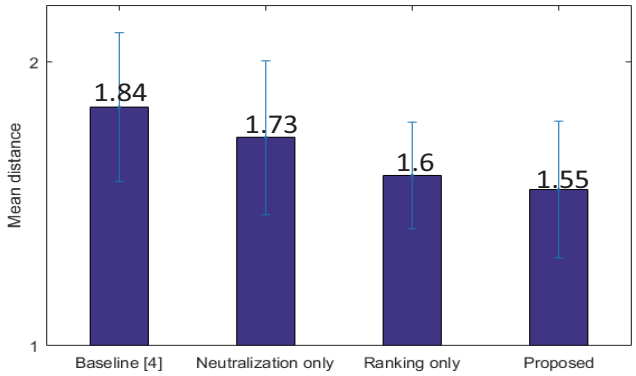


Fig. 9. Comparison of the mean of 3D matching distances from each method with 10 subjects

sured each method's mean distances of 3D matching distances between a reference and 3D face models using each method with the expression changes in the 10 videos (Sec. IV. A). Fig. 9 shows mean distances from each method. In Fig. 9, the proposed method shows the best performances. Neutralization and ranking method also generated better results than that of the baseline method.

V. Conclusion

We propose a neutral 3D face modeling system in the presence of expression changes for non-cooperative subjects. We estimate the subject's expressions and compute the expression score for each frame. Based on the estimated expression scores, we rank frames and neutralize non-neutral faces before registrations. Our method generates visually and quantitatively better 3D face models compared to the state of the art method [4]. Although we focus on generating neutral 3D face models, our method can be applied to generate a 3D face model with specific expression in the presence of expression changes.

Acknowledgments

This research is funded in part by the IT R&D program of MOTIE/KEIT (10041610, The development of automatic user information extraction and recognition technology based on perception sensor network under real environment for intelligent robot).

References

[1] J. Smisek, M. Jancosek, and T. Pajdla, "3d with kinect," in *Consumer Depth Cameras for Computer Vision*. Springer, 2013, pp. 3–25.

[2] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Robotics-DL tentative*. International Society for Optics and Photonics, 1992, pp. 586–606.

[3] U. Castellani and A. Bartoli, "3d shape registration," in *3D Imaging, Analysis and Applications*. Springer, 2012, pp. 221–264.

[4] M. Hernandez, J. Choi, and G. Medioni, "Near laser-scan quality 3-d face reconstruction from a low-quality depth stream," *Image and Vision Computing*, vol. 36, pp. 61–69, 2015.

[5] S. Izadi, R. A. Newcombe, D. Kim, O. Hilliges, D. Molyneaux, S. Hodges, P. Kohli, J. Shotton, A. J. Davison, and A. Fitzgibbon, "Kinectfusion: real-time dynamic 3d surface reconstruction and interaction," in *ACM SIGGRAPH 2011 Talks*. ACM, 2011, p. 23.

[6] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments," *The International Journal of Robotics Research*, vol. 31, no. 5, pp. 647–663, 2012.

[7] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 559–568.

[8] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 343–352.

[9] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev, "3d self-portraits," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 187, 2013.

[10] B. Chu, S. Romdhani, and L. Chen, "3d-aided face recognition robust to expression and pose variations," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1907–1914. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2014.245

[11] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating Faces in Images and Video," *Computer Graphics Forum*, 2003.

[12] J. P. Lewis, K. Anjyo, T. Rhee, M. Zhang, F. Pighin, and Z. Deng, "Practice and Theory of Blendshape Facial Models," in *Eurographics 2014 - State of the Art Reports*, S. Lefebvre and M. Spagnuolo, Eds. The Eurographics Association, 2014.

[13] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 43:1–43:10, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2601097.2601204

[14] X. Lu and A. K. Jain, "Deformation modeling for robust 3d face matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 8, pp. 1346–1357, 2008.

[15] I. Mpiperis, S. Malassiotis, and M. G. Strintzis, "Bilinear models for 3-d face and facial expression recognition," *Information Forensics and Security, IEEE Transactions on*, vol. 3, no. 3, pp. 498–511, 2008.

[16] F. Al-Osaimi, M. Bennamoun, and A. Mian, "An expression deformation approach to non-rigid 3d face recognition," *International Journal of Computer Vision*, vol. 81, no. 3, pp. 302–316, 2009.

[17] Primesense camera. [Online]. Available: https://en.wikipedia.org/wiki/PrimeSense

[18] Dlib, c++ open source library. [Online]. Available: http://dlib.net/

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[20] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[21] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 5, pp. 698–700, 1987.

[22] R. Min, J. Choi, G. Medioni, and J.-L. Dugelay, "Real-time 3d face identification from a depth camera," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 1739–1742.