2012 12th IEEE-RAS International Conference on Humanoid Robots
Nov.29-Dec.1, 2012. Business Innovation Center Osaka, Japan

Automatic Face Replacement for a Humanoid Robot with 3D Face Shape Display

Akinobu Maejima*, Takaaki Kuratate†, Brennand Pierce†, Shigeo Morishima* and Gordon Cheng†
* Waseda University, Tokyo, 169–8555, Japan
Email: akinobu@mlab.phys.waseda.ac.jp, shigeo@waseda.jp
† Institute for Cognitive Systems, Technical University Munich, Munich, 80333, Germany
Email: {kuratate, bren, gordon}@tum.de

***Abstract*—In this paper, we propose a method to apply any new face to a retro-projected 3D face system, the Mask-bot, which we have developed as a human-robot interface. The robot face using facial animation projected onto a 3D face mask can be quickly replaced by a new face based on a single frontal image of any person. Our contribution is to apply an automatic face replacement technique with the modified texture morphable model fitting to the 3D face mask. Using our technique, a face model displayed on Mask-bot can be automatically replaced within approximately 3 seconds, which makes Mask-bot widely suitable to applications such as video conferencing and cognitive experiments.**

I. Introduction

To provide smooth interaction between humans and robots, it is important to provide robots with the capacity to communicate naturally – similar to how people communicate in daily life – in appearance and behaviour, especially in face-to-face communication. To achieve such a goal, many researchers developed various types of humanoid robotic faces in recent years[1], [2], [3], [4], [5].

Kuratate et al. developed the humanoid robotic face called Mask-bot[5], which consists of a 3D life-size face-shaped screen, a data projector with fisheye lens and a pan-tilt unit, and whose expressions can be controlled by projecting facial animations synthesized by a computer graphics technique onto a 3D face-shaped screen (Fig. 1). Similar retro-projected 3D faces have been developed by various researchers[6], [7], and although all give better appearances to users compared to conventional flat panel displays or stereoscopic 3D displays, Mask-bot has a great advantage in its ability to express not only animated abstract faces, but realistic faces as well. Moreover, in principle, it can easily switch faces by changing not only to a differently shaped screen but also by changing the face texture of the facial animation. However, to replace a robotic face with other individual faces while preserving the individual 3D facial geometry, we have to perform a calibration between the individual 3D face model and the face-shaped screen for each face. This calibration process requires significant efforts.

To solve this problem, we proposed an automatic face replacement method for such retro-projected 3D face systems, especially for Mask-bot. Our method can automatically generate an appropriate individual texture which is aligned to a standard face model or pre-defined face model already calibrated to a face-shaped screen using a single frontal image of a new person. A part of the texture which is impossible to acquire from a single snapshot is complemented by using our texture morphable model. Using our method, we can replace a retro-projected robot face with anybody in approximately 3 seconds. As a result, it is possible to apply Mask-bot to various applications, such as video conferencing and cognitive experiments.

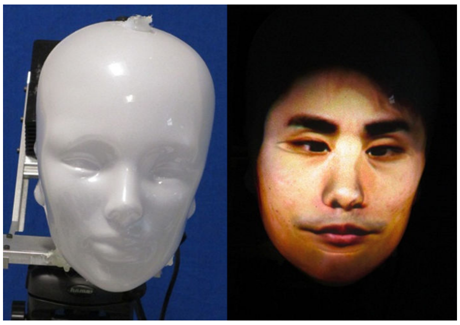


Fig. 1. The Mask-bot without projection (left) and with the image synthesized with the proposed method (right).

II. Related work

In closely related work, Hayashi et al. have developed a humanoid robotic face where the 3D face-shaped display can be roughly approximated to a subject's 3D facial geometry by controlling needle positions and projecting the subject's video onto the display by front projection [4]. However, the available space is limited due to the front projection. Moreover, the system enforces subjects to maintain their head poses while taking video. This constraint is not practical nor possible for actual use in many applications, especially in video conferencing. We would have a feeling of strangeness if we omitted normal head pose information, losing an important part of the natural communication behaviors we seek. On the other hand, our method can automatically generate an appropriate individual texture which is aligned to a standard face model or pre-defined face model already calibrated to a face-shaped screen from a single snap shot and synthesize facial animations. Our system does not constrain a subject's head pose during the capturing process. Users

978-1-4673-1369-8/12/$31.00 ©2012 IEEE

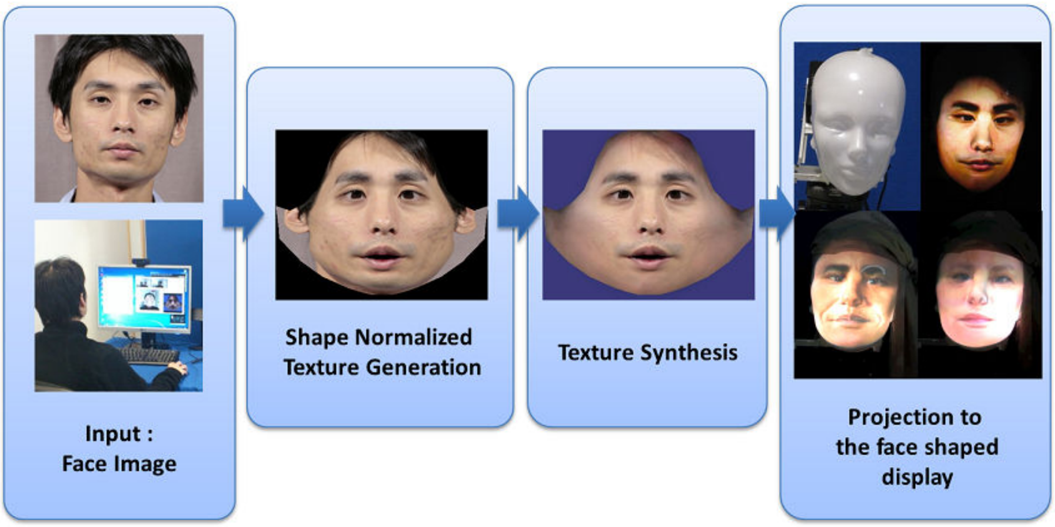


Fig. 2. System Overview

only have to look at camera for a few seconds. Therefore, our method is more convenient for various applications.

As for the texture estimation, one of the representative methods which addresses the need to augment texture captured from a single snapshot is the 3D Morphable Model proposed by Blanz et al[8]. In this paper, we employ a similar approach for the texture part of the 3D Morphable Model. However, instead of multi-resolution model fitting which requires considerable time, we introduce a fitting acceleration strategy to decimate the evaluation points while still preserving the estimated texture quality.

In this paper, we assume that the subject's head pose for the new face capture is roughly frontal with respect to the camera. To allow for large variations of the head pose, we correct for the head pose by fitting a 3D Standard Face Model (SFM) or 3D average face model to the face in the input image using the rigid body transformation and rendering the frontal-posed image of the fitted face model before the texture estimation.

III. Overview

In this paper, we define a 3D face model which has already been calibrated to the display as an Standard Face Model and utilize it to create all of the individual face textures. By generating an individual texture that has same shape of the SFM, we can change the robotic face by only replacing individual textures.

An individual face texture for a 3D face-shaped display - especially targeted to the Mask-bot system - is generated from a single snapshot as follows (Fig.2). First, from an input snapshot we create an individual face texture which is matched to the SFM. We refer to the generated texture as a shape normalized texture. Second, to complement a texture with missing parts such as the lateral face, we estimate a texture by using our texture morphable model. Finally, a complete individual face texture for the 3D face-shaped display is generated by linear blending both shape normalized and resulting estimated textures. The generated texture is immediately projected to the Mask-bot system.

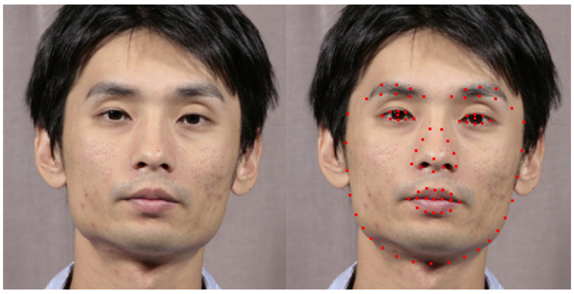


Fig. 3. The result of feature point detection. The left side is the input image and the right side is the detection result.

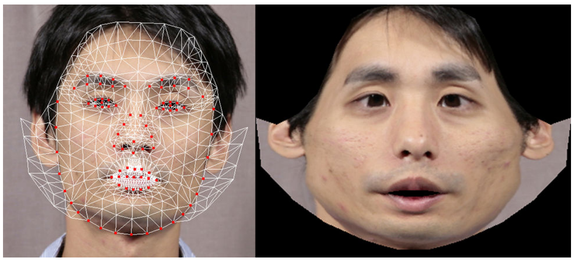


Fig. 4. The result of shape normalized texture generation. Based on the fitting result shown in the left, a shape normalized texture will be synthesized as in the right.

IV. Shape Normalized Texture Generation

A shape normalized texture is generated by the following procedures. First, 84 facial feature points are automatically detected by Zhang's detector [9] as shown in Fig.3. Then, using 84 pairs of detected facial feature points and their corresponding vertices on a 3D SFM, the SFM is fit to the input image by Radial Basis Functions (RBFs). After

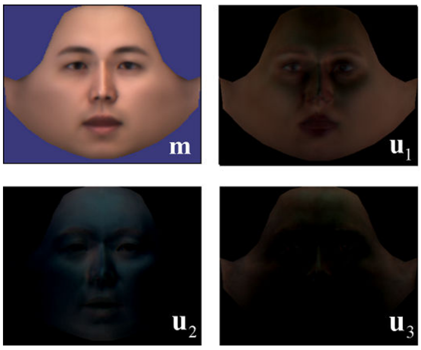


Fig. 5. The texture morphable model. The top-left is a mean vector and first three basis vectors are located in the remaining part with raster scan order.

this, image coordinates of the deformed face model can be obtained. RBFs $f_i(x)$ for $i = x, y$ elements corresponding to both feature points and unknown texture coordinates respectively is defined by following:

$$f_i(\mathbf{x}) = \sum_{j=1}^{N} w_j^i \phi(\mathbf{x} - \mathbf{x}_j) \tag{1}$$

N is the number of feature points and vertices on the SFM. $\mathbf{x} = (x, y)'$ is a 2D coordinate of an arbitrary vertex on the SFM, $\mathbf{x}_j = (x_j, y_j)'$ is also a 2D coordinate (a center of a basis function) of a j-th vertex on the SFM corresponding to j-th feature point. Additionally, w_j^i is a weight for j-th basis function of $i = x, y$ elements, ϕ is a basis function with the Hardy multi-quadrics [10] which is defined by following:

$$\phi(\mathbf{x} - \mathbf{x}_j) = \sqrt{||\mathbf{x} - \mathbf{x}_j||^2 + s_j^2} \tag{2}$$

where, s_j^2 is a closest point on the SFM for $\mathbf{x}_j$ which is obtained by

$$s_j^2 = \min_{i \neq j} ||\mathbf{x} - \mathbf{x}_j|| \tag{3}$$

Using the obtained u, v coordinates, we render the standard 3D face model with the input image as a texture to generate a normalized face texture as shown in Fig.4.

V. Texture Synthesis

To create a complete individual face texture, we estimate the texture difficult to acquire from the input snapshot (e.g. the lateral part of a face in the case of a frontal pose) using the Texture Morphable Model (TMM). The TMM is linear combination model of multiple face textures and can be built by applying Principal Component Analysis to 130 individuals' facial images aligned with the same format of the shape normalized image. Fig.5 shows the average vector and $+3\sigma$ first three basis vectors of our TMM.

Given a shape normalized texture as mentioned in Section IV, the texture estimation is performed by minimizing the error between a shape normalized texture and its estimate as defined by Equation (4):

$$\underset{\mathbf{a}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} ||\mathbf{W}_{\kappa_i}(\mathbf{d}_{\kappa_i} - \mathbf{U}'_{\kappa_i}\mathbf{a})||_2^2 \tag{4}$$

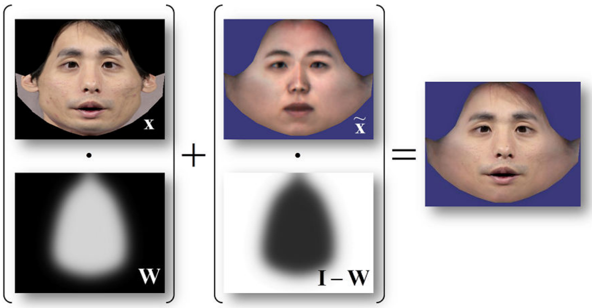


Fig. 6. Texture blending between shape normalized texture and estimated texture.

where, $'$ represents matrix transpose. $\mathbf{U}$ is a matrix including basis vectors of the TMM and $\mathbf{a}$ is a vector including their corresponding coefficients. $\mathbf{d}$ is a differential vector $\mathbf{x} - \overline{\mathbf{x}}$, $\mathbf{x}$ is a vector-form of an shape normalized image and $\overline{x}$ is a average vector of the TMM. $\mathbf{W} = \text{diag}(w_1, \cdots, w_n)$ is a diagonal matrix whose elements mean contribution weights for each pixel in the minimization. To decimate evaluation points in the texture estimation, we select evaluation points from an uniform grid with specific interval k. κ means a vector containing pixel indices selected as evaluation points. N is the total number of the selected indices. The energy function shown in Equation (4) can be minimized by the Levenberg-Marqardt algorithm with its Gradient and Hessian. After that the estimated texture $\tilde{\mathbf{x}}$ can obtained by linear blending texture bases according to the resulting $\mathbf{a}$. Finally, the shape normalized texture and the estimated texture is mixed by Equation (5) to obtain final output $\hat{\mathbf{x}}$ as in Fig.6.

$$\hat{\mathbf{x}} = \mathbf{W}\mathbf{x} + (\mathbf{I} - \mathbf{W})\tilde{\mathbf{x}} \tag{5}$$

We implemented this such that the generated texture is transferred from the image capture / texture synthesis PC to the Mask-bot facial animation PC through the network and projected onto Mask-bot.

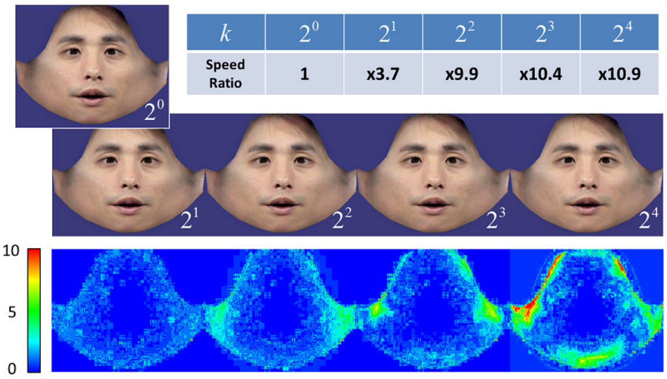
VI. Experiments

With the basic algorithms determined, it is essential to define the various parameters of the algorithm to work with reasonable output image quality and a reasonable time frame. Therefore, we evaluate parameters for grid spacing (texture estimation) and number of basis of morphable model (texture synthesis) as follows.

A. Decision of the optimal number of grid spacing

To decide an optimal number and location of evaluation points in the texture estimation, we measure the average processing time and the average texture estimation error when we generate individual textures from 30 individuals' face image with sampling interval $k = 2^0, 2^1, \cdots, 2^4$. The average texture estimation error can be calculated by

$$E = \frac{1}{3N} \sum_{i=1}^{N} ||\mathbf{T}_i^{tgt} - \mathbf{T}_i^{est}|| \tag{6}$$



k	2^0	2^1	2^2	2^3	2^4
Speed Ratio	1	x3.7	x9.9	x10.4	x10.9

Fig. 7. The speed ratio with various numbers of sampling interval k used for the texture estimation.

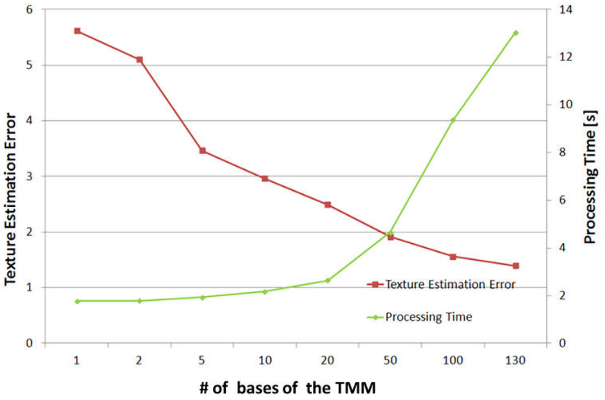


Fig. 8. The average texture estimation error and the processing time.

where, $\mathbf{T}_i^{tgt}$ and $\mathbf{T}_i^{est}$ are RGB color vector at pixel i of the target and the input texture respectively, and N means the number of evaluation points. Moreover, we calculated the speed ratio k_n/k_0 and visually confirmed generated textures and the error maps. Our experimental environment was Intel Core i7 3.4 GHz, 8 GB RAM. The result of this experiment is shown in Fig.7. In the error map, the hue value represents the error strength (Blue: 0 Red: more than 10). From this results, we found that $k = 2^3$ is the best sampling interval for the texture estimation. At this time, we can accelerate the texture estimation approximately ten times faster than the estimation without decimation of evaluation points while still preserving the quality of the estimated texture.

B. Decision of optimal number of basis for synthesis texture

To determine an optimal number of the bases of the TMM, we conduct the texture estimation experiment for 30 individuals' frontal face textures. Varying the number of bases of the TMM to $1, 2, 5, 10, 20, 50, 100, 130$, we measured the average estimation error and the average processing time. The experimental environment is the same as in section VI-A. In this experiment, we set the grid interval to $k = 2^3$ which is the best result as mentioned above section. The result of this experiment is shown in Fig.8. Also, we show input frontal face images and generated textures which are synthesized by using $1, 10, 20, 50,$ and 130 texture bases respectively in Fig.9. In Fig.8, two vertical axes represent the average estimation error and the average processing time, and the horizontal axis means the number of texture bases.

From Fig.8 and 9, we found that the texture estimation error could be reduced if we use more basis vectors of the TMM in the texture estimation. However, the texture quality tends to be low when we use more bases because high order bases contain high frequency components similar to noise that influences the texture estimation. Furthermore, the computational time is increased exponentially due to the cost of the linear combination. Therefore, we determined the optimal number of TMM's bases is 20 from the computational time and the quality view points. With this condition, we can generate an individual face texture in an average of 2.6 seconds after inputting a face image. Note that, we have not yet optimized the actual codes, and therefore further acceleration of processing speed is expected.

Using these optimal parameters, some examples of original input images, generated textures and the final output images on Mask-bot are shown in Fig.10 and supplementary video. As you can see from this figure, even though the proposed technique is less complex than the 3D Morphable Model originally proposed by Blanz et al. [8], the texture quality estimated by this technique is quite sufficient for such retro-projected facial animation. Moreover, we preliminary tested the 3D Morphable Model-based implementation [11] under the same experimental environment as in the section VI-A and we found out that our texture estimation is approximately 10 times faster. We therefore conclude that the proposed technique is effective for making a humanoid robotic face with a face-shaped display.

VII. Conclusion

In this paper, we have proposed an automatic face replacement method for a retro-projected 3D face-shaped display developed for humanoid robot faces.

We apply a new person's face image to a standard 3D face model which has already been calibrated to the display to avoid building a new 3D face model and its calibration which is normally required to add the new person. To add a new person's face to the system, an individual face texture which has same geometry as the standard 3D face model for the 3D face-shaped display is automatically generated by using the proposed method from a single snapshot. From the experimental results, we found that our method can generate an individual texture in approximately 3 seconds, which can be further decreased by optimization of our code. Our method makes retro-projected humanoid robot faces such as the Mask-bot more suitable and widely applicable to video conferencing and cognitive experiments.

Now, we attempt to implement the proposed method to the new version of the Mask-bot hardware [12]. Herewith, an user can also easily replace the 3D face screen with other masks, including either individualized, generalized,

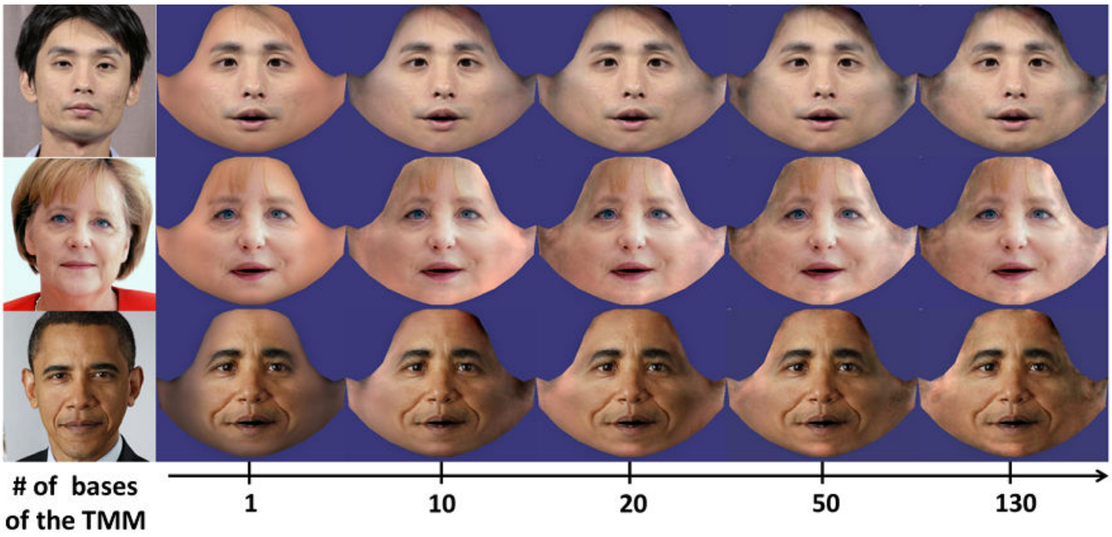


Fig. 9. The comparison of final image quality using different number of principal components.

or even caricatured face-shape screens as compared to the current version [5]. For further development of the proposed algorithm, we still have difficulties to generate a facial texture with fine details such as wrinkles, freckles, and birthmarks, because our texture morpahble model is a linear combination model of collected face textures without correspondences between fine features. Therefore, we need to solve this problem to generate a more realistic texture to provide more promising appearances. Moreover, in this paper, we decided the evaluation points by sampling from a uniform grid with arbitrary interval on the input image. However, the evaluation points are a little bit redundant for our texture estimation. To make the texture estimation more efficient, we would like to apply a kind of key point selection technique proposed by Mayer et al. to decide effective evaluation points[13].

We also started subjective evaluation - gender identification of the mask with different face textures as a beginning [14] and we plan to perform further experiments such as the face identification with various 3D face masks, and examine how synthesized texture and 3D geometry will contribute to personal identification for human users. We believe that such evaluations will guide us to develop robotic heads with optimal shape and animation qualities for various applications.

ACKNOWLEDGMENT

This work was supported by DFG cluster of excellence 'Cognition for Technical systems CoTeSys' of Germany and the ERASMUS MUNDUS BEAM Scholarship: L031000059.

REFERENCES

[1] H. Ishiguro, "Understanding humans by building androids," *SIGDIAL Conference*, pp. 175–175, 2010.

[2] D. Hanson, "Exploring the aesthetic range for humanoid robots," *CogSci-2006 Workshop: Toward Social Mechanisms of Android Science*, 2006.

[3] C. Kroos, D. Herath, and Stelarc, "The articulated head pays attention," *In proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI'10)*, pp. 357–358, 2010.

[4] K. Hayashi, Y. Onishi, K. Itoh, H. Miwa, and A. Takanishi, "Development and evaluation of face robot to express various face shape," *In proceedings of the 2006 IEEE International Conference on Robotics and Automation (ICRA 2006)*, pp. 481–486, 2006.

[5] T. Kuratate, Y. Matsusaka, B. Pierce, and G. Cheng, "Mask-bot: a life-size robot head using talking head animation for human-robot communication," *In proceedings of the 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2011)*, pp. 99–14, 2011.

[6] F. Delaunay, J. de Greeff, and T. Belpaeme, "Lighthead robotic face," *In proceedings of the 6th International Conference on Human-robot interaction (HRI'11)*, p. 101, 2011.

[7] S. A. Moubayed, S. Alexandersson, J. Beskow, and B. Granström, "A robotic head using projected animated faces," *In proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP 2011)*, p. 69, 2011.

[8] 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 (SIGGRAPH '99)*, pp. 187–194, 1999.

[9] L. Zhang, S. T. H. AI, , and S. Lao, "A fast and robust automatic face alignment system," *In IEEE International Conference on Computer Vision (ICCV 2005), Demo program*, 2005.

[10] J.-Y. Noh, D. Fidaleo, and U. Neumann, "Animated deformations with radial basis functions," *In proceedings of the ACM symposium on Virtual reality software and technology (VRST '00)*, pp. 166–174, 2000.

[11] "FaceGenModeler," *http://www.facegen.com*, (last accessed on July 8, 2012).

[12] B. Pierce, T. Kuratate, C. Vogl, and G. Cheng, "Mask-Bot 2i: An active customisable robotic head with interchangeable face," *In proceedings of the 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, 2012.

[13] M. Meyer and J. Anderson, "Key point subspace acceleration and soft caching," *ACM Trans. Graph.*, vol. 26, no. 3, 2007.

[14] T. Kuratate, M. Riley, B. Pierce, and G. Cheng, "Gender identification bias induced with texture images on a life size retro-projected face screen," *In proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN2012)*, pp. 43–48, 2012.

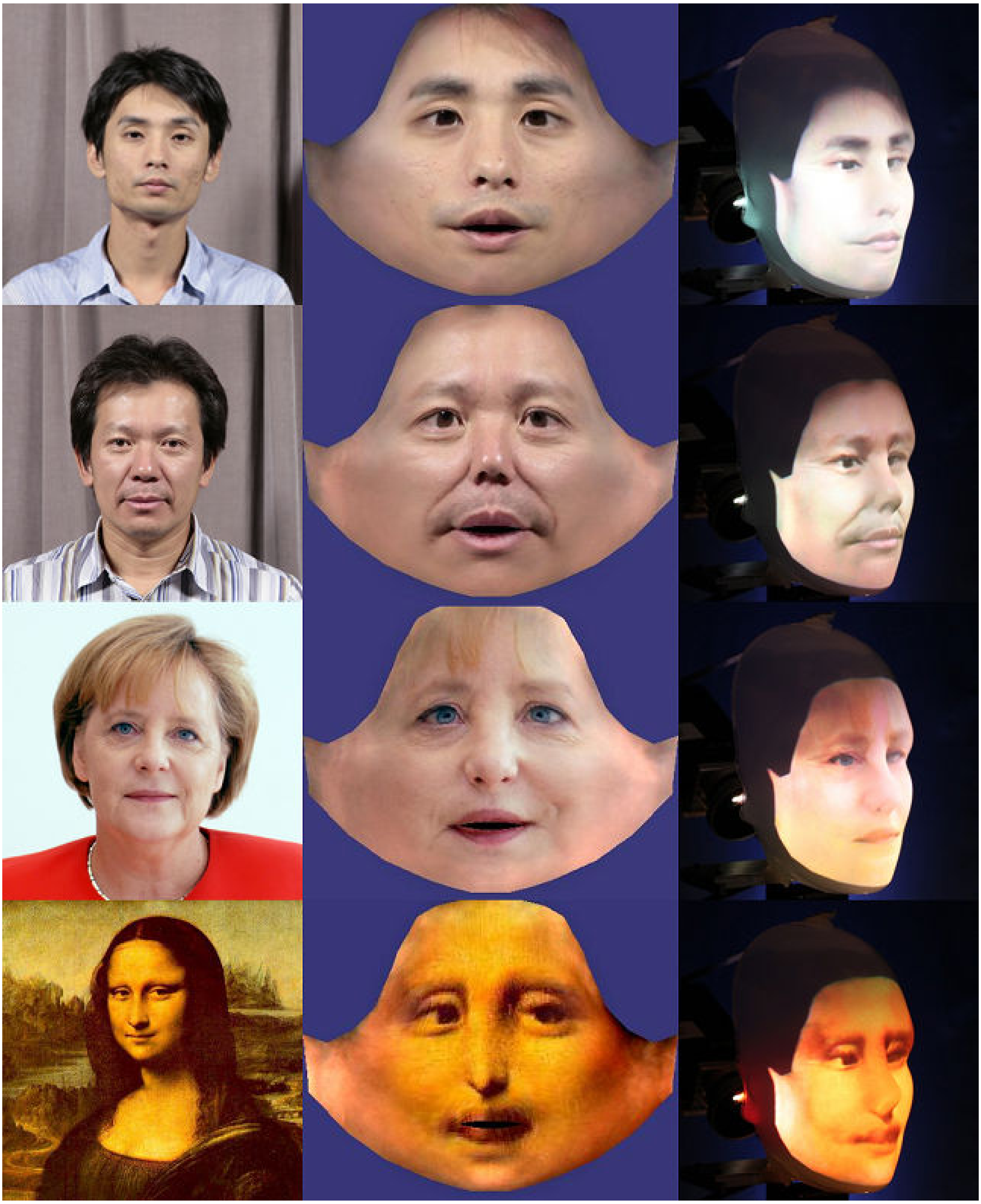


Fig. 10. The results of texture projection for the face-shaped display.