# *Face Swapping under Large Pose Variations: a 3D Model Based Approach*

*Yuan Lin, Shengjin Wang*
Department of Electronic Engineering
Tsinghua University
Beijing, P.R. China
{linyuan, wsj}@ocrserv.ee.tsinghua.edu.cn

*Qian Lin, Feng Tang*
Multimedia Interaction and Understanding Lab
USA Hewlett-Packard Labs
Palo Alto, CA, USA
{qian.lin, feng.tang}@hp.com

***Abstract*—*Traditional face swapping technologies require the faces of source images and target images have similar pose and appearance (usually frontal). This limits its applications. This paper presents a method for face swapping based on personalized 3D head models. This framework builds a personalized 3D head model from a frontal face and can be rendered at any pose to match the characters in the image we want to swap. The 3D head model is constructed by a user uploaded frontal view face image. This construction process goes through face alignment and feature point matching. The final personalized 3D head is built by deforming a standard 3D head model using radial basis function to match the specific person. To make the synthesized face seamlessly blended into the image, color transfer and multi-resolution spline technique are applied. We use the proposed technique to create personalized storybook where the characters are replaced with a user's face and promising results are obtained. The system can be used in face de-identification as well.*

***Keywords- 3D Head Reconstruction, Face Swapping, Color Transfer, Multi-resolution Spline***

## I. INTRODUCTION

Face swapping is one of the hottest technologies in multimedia and image processing. Generally speaking, it replaces a face/faces in the target image with the face in the source image. It can be used in many applications, such as children's role plays in story books or face replacement of the stunt man's face with the user's. Actually, the present 2D images based methods always have strict limit of the similarity between the source face and the target face for example similar pose and appearance. This limits its applications.

Dmitri Bitouk et al. proposed a face swapping method [1] based on a large collection of face images. Face images with similar pose and appearance are first selected as candidate source images. Then, the pose, lighting and color of the candidate face images are adjusted to match the appearance of those in the target image. At last, the face swapping results with different candidate images are scored and presented to the user. This method generated great results for some images, but they are restricted to poses within ±25° in yaw and ±15° in pitch. This method also highly relies on the size and variety of the face images. Another face swapping method proposed by Jianke Zhu uses unsupervised face alignment technology [2]. It can generate impressive results when the pose and appearance of source image and target are very similar. When the pose of the source face image and the one of the target face image have noticeable difference, the algorithm often fail. To solve this problem, we propose to use a 3D head model to generate the face views in arbitrary viewpoints. Thus, the face swapping is not subject to the pose similarity restrictions.

There are several methods of building 3D head models from images. One category is based on 3D morphable models (3DMM) [3], another category is based on one standard 3D head and certain interpolation method, and another is based on stereo matching method. Since the 3DMM has been taken as an optimization method, it is easily trapped into local minimum values. And 3DMM is quite time consuming, which on the average takes 4.5 minutes on a 2GHz Pentium 4 processor. Both 3DMM and the stereo matching methods can only generate partial face, which lack reality. On the other hand, deforming a 3D standard head model to match the specific person is very flexible. It has analytical solutions, and it is a complete model when the 3D standard model is a head model instead of a face model. Generating one 3D head model for certain subjects with one frontal image view is attracting a lot of attention. Some commercial softwares are developed to build the 3D avatar for the user. It can generate realistic 3D head, but it requires significant user interactions.

There is some works deal with video face replacement as well, such as [4] and [5]. Yi-Ting Cheng's method [4] goes through 3D face reconstruction, expression matching, relighting and blending. Kevin Dale's method [5] goes through 3D face reconstruction, 3D multilinear model registration, face performance tracking, retiming and blending. The main difference with our approach is the speed. Their complicated reconstruction method takes 30-60seconds, while ours only takes 1.8 seconds on 3.2GHz Intel (R) Core (TM) i5 processor.

In our system, the user uploads one frontal view, and then one 3D head model will be generated automatically. The major contribution of our work is to use a whole 3D head model in face swapping to solve the traditional restrictions about pose and appearance similarity between the source face and the target face. To make the swap seamless, color transfer and the multi-resolution spline technique are used to make the results look natural. While our algorithm is designed to be automatic, we also provide an interactive tool for the user to refine the feature localization to enhance the reconstruction accuracy. The only user interaction needed is to mask the target face region to be swapped. However, in many applications, for example personalized photobook, this only needs to be done once and used for different users. The

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.26

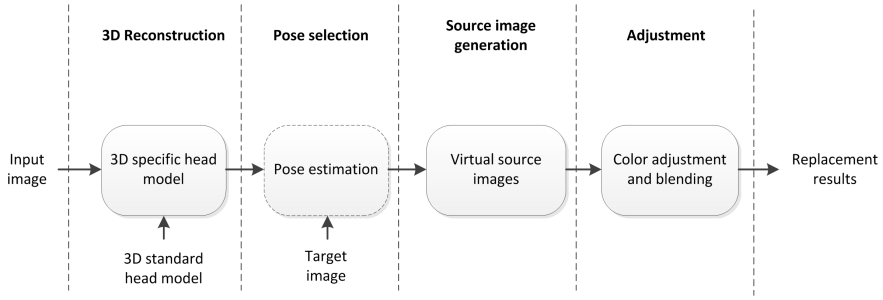



Figure 1. Schematic representation of the face swapping system

rest of the reconstruction and swapping process is fully automatic.

## II. The Face Swapping System

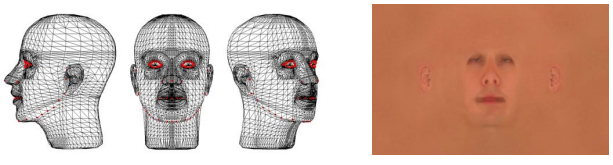
The whole swapping process contains 4 parts: 3D head reconstruction, pose estimation, source image generation and adjustment. The head reconstruction deforms the standard 3D head with the frontal face image based on face detection and face alignment. It is detailed in section III. Then, the pose of the target image is estimated. The reconstructed 3D head is rendered under the pose consistent with the estimated pose to generate the source face image (section IV). In section V, we discuss the color adjustment and multi-resolution spline techniques that are adopted to make the merge seamlessly. The whole process is shown in Fig. 1. The results and the conclusion are at the end of the paper.

## III. Personliazed 3D Head Reconstruction

In this paper, the 3D standard head model from Pighin's work [6] is adopted. The 3D model and a standard texture are show in Fig.2 (a) and (b) The reconstruction process is as follows. First, the key 2D and 3D feature points are defined (section III.A, III.B). Second, the facial alignment is applied to obtain the facial feature localization. Third, 2D and 3D feature points are matched, and the 3D standard head is deformed. Finally, a seamless texture is generated with the input image and the standard texture and mapped onto the 3D head model.

### A. 3D Feature Points Definition

The definition of the 3D head feature points and the extraction of their correspondences are crucial to the whole reconstruction process. They determine the realism and the accuracy of the generated results. We chose distinctive face points such as eyes contour, nose, mouth contour and face contour as feature points. These feature points are shown with red points in Fig. 2 (a).



(a) (b)

Figure 2. The adopted standard model. (a) 3D Standard head with feature points marked in red. (b) Standard texture

### B. Facial Alignment

Given the input frontal face image, we extract 2D feature points corresponding to 3D features for reconstruction and texture generation. For the facial alignment, an improved active shape model [7] is used. It is trained using a large number manually labeled dataset and is capable to accurately locate 88 facial feature points. An example result is shown in Fig. 3(a) and (b).

### C. Feature Matching and A Morphable Model

The facial alignment results may not naturally correspond to the previously defined 3D feature points. We employ a resampling method based on the face alignment results to match image feature points and the 3D feature points. The feature matching results are shown in Figure 3 (c) and (d).

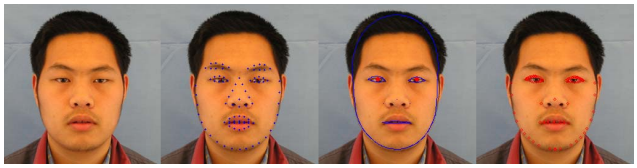
After the feature matching, the 3D standard head model is deformed with radial basis function:

$$\mathbf{vp}_j = \sum_{i=1}^{Q} \boldsymbol{\alpha}_i \phi\left(\left\|\mathbf{p}_j - \mathbf{vf}_i\right\|\right) + \mathbf{a} + \mathbf{b}x + \mathbf{c}y, \quad j \in [1, N] \qquad (1)$$

where $\mathbf{p}_j$ is the j-th vertex of the standard head, $\mathbf{vf}_i$ is the i-th feature point of the standard head, $\mathbf{vp}_j$ is the j-th vertex of the morphed head. x, y is from $\mathbf{p}_j$. Q is the number of the feature points, N is total vertices number. The radial basis function is $\phi(r) = e^{-||r||/k}$, k is related to the image size. In our experiment, the image size is 256*256, k is 64. Other parameters $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\boldsymbol{\alpha}_i$ can be obtained from the feature points matching result. After all the parameters are computed, the specific shape model can be obtained.

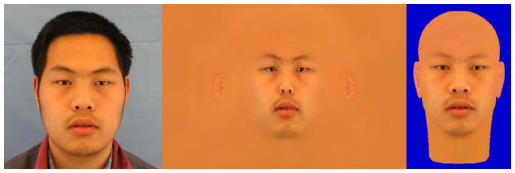
### D. Texture Generation and Texture Mapping

The frontal face area from the source image is used to generate the personalized texture map from the standard texture. To makes the texture seamlessly, color transfer [8]



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

Figure 3. The image feature points extraction. (a) The input frontal image. (b) Face alignment result. (c) Face feature points interpolation. (d) Feature matching result.



(a) (b) (c)

Figure 4. The texture generation results. (a) The input frontal image. (b) The synthesized texture. (c) The reconstructed 3D head

and multi-resolution spline technique [9] are used, the details are in section V. An example texture generation and mapping results are shown in Fig. 4.

## IV. Virtual Source Image Generation

The traditional 2D based face swapping method suffers from the pose and appearance similarity restriction of the target face. To solve this pose similarity requirement, we use the constructed 3D head model to generate projection of the face rendering from arbitrary view that is consistent with the target Traditional face swapping technologies can then be used for swapping.

To demonstrate our work, several cartoon images with known pose are drawn by an artist. After a user uploads one single frontal view, one 3D head model is generated using the aforementioned method. The rendering of face of the pose corresponding to the cartoon image is obtained naturally. We call it the virtual source image.

The target face region is manually segmented to specify the area we want to swap. Then three other feature points: two eye centers and the mouth center are extracted. Note that for some applications like personalized storybook, this process only needs to be done once for all users. With the aid of these three features, the source face image can be rotated and scaled. At the same time, the face region is located in the source image as well. The face region mask is predefined on the standard model. Once the pose is estimated, the 3D head is rotated to that pose, and the source mask is obtained, see Fig. 5. In Fig. 5, 3D face region projections with different yaws and pitches are shown. For the generated virtual source image, its mask image has the same size with it. And the projection area of the 3D face region is white and other part is black. In this way, two face region masks are obtained. If one of the eyes is missing, which means only half face is observed, the side face usually doesn't have much depth variations, which can be solved using an image based approach for example image warping. Our system handles the cases with both eyes visible to make the most appealing swapping.

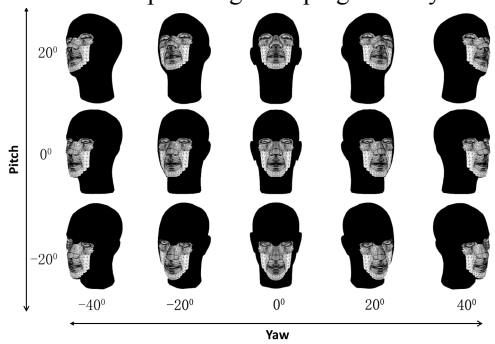



Figure 5. Demonstrations for facial region masks for different poses

## V. Color Adjustment And Image Blending

To adjust the color differences between source and target faces, color transfer [8] proposed by E Reinhard et al. is adopted. At first, the images are translated from RGB space to lαβ space. The mean and variance of the source face region are replaced with the mean and variance of the target face region with the aid of two face masks gotten in section IV. Then the result image is translated to RGB space again. The color transfer process is shown in Fig. 6.

To make the replacement seamless, multi-resolution spline technique [9] is used. Gaussian and Laplacian pyramids of the virtual source image and the target image are constructed. They are defined as $G_k$ and $L_k$. $G_0$ is the original image. Each Gaussian pyramid level is generated from its predecessor with a REDUCE operation. Then for 0<k<n:

$$G_k = \mathrm{REDUCE}\left[G_{k-1}\right] \tag{2}$$

by which we mean:

$$G_k(i,j) = \frac{1}{4}\sum_{a=0}^{1}\sum_{b=0}^{1} G_{k-1}(2i+a, 2j+b) \tag{3}$$

The Laplacian pyramids are constructed with Gaussian pyramid and a EXPAND operation. Then for 0<k<n:

$$L_k = G_k\text{-}\mathrm{EXPAND}\left[G_{k+1}\right] \tag{4}$$

Say $\mathrm{EXPAND}[G_{k+1}] = G_k^*$, $G_k^*$ is defined as follows.

$$\begin{aligned}
G_k^*(2i,2j) &= G_{k+1}(i,j)\\
G_k^*(2i+1,2j) &= \frac{1}{2}\left[G_{k+1}(i,j)+G_{k+1}(i+1,j)\right]\\
G_k^*(2i,2j+1) &= \frac{1}{2}\left[G_{k+1}(i,j)+G_{k+1}(i,j+1)\right]\\
G_k^*(2i+1,2j+1) &= \frac{1}{4}\begin{bmatrix} G_{k+1}(i+1,j)+G_{k+1}(i,j+1)+\\ G_{k+1}(i,j)+G_{k+1}(i+1,j+1)\end{bmatrix}
\end{aligned} \tag{5}$$

Based on (2) to (5), Gaussian and Laplacian pyramids for the virtual source image and the target image are constructed.

The merging mask is the union of the source mask and the target mask of the same level. And then it is filtered with Gaussian filter. Both Laplacian images from the same level are merged with the mask image. The formula is as (6).

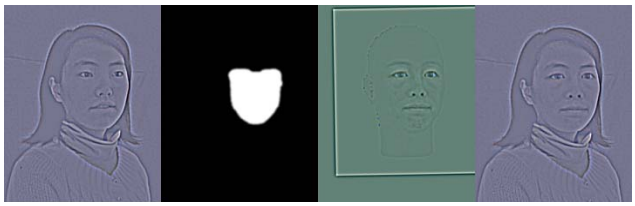
$$L_j = L_j^s * \mathrm{Mask}_j + (1-\mathrm{Mask}_j) * L_j^t \tag{6}$$

For level j, $L_j^s$ is the source image's Laplacian image, and $L_j^t$ is the target image's Laplacian image, $\mathrm{Mask}_j$ is the merging mask. $L_j$ is the result.

Fig. 7 shows the Laplacian images blending process at level 2 of a 3-level pyramid with (6). From left to right are the Laplacian image of the target image $L_1^t$, the merging

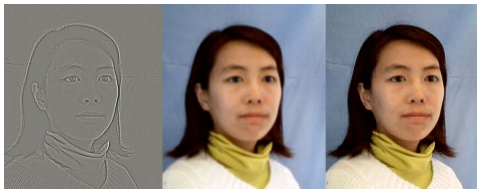
(a) (b) (c) (d)

Figure 6. Color transfer process. (a) The target image. (b) Face replacement mask. (c) Face replacement with original source image. (d) Face replacement with color adjustment source image



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

Figure 7. Laplacian images blending process. (a) The Laplacian image of the target image. (b) The merging mask. (c) The Laplacian image of the virtual source image. (d) The merging result of (a) and (c)



(a) (b) (c)

Figure 8. Image reconstruction process. (a) The blending result. (b) The expanded reconstruction image of the previous level. (c) The reconstruction result of the present level

mask $Mask_1$, the Laplacian image of the generated source image $L_1^s$, and the final blending result $L_1$.

Then all images $L_j$ ($0 \leqslant j \leqslant n$) are used to reconstruct the final image. The reconstruction is from the highest level n to the lowest level 0. The higher level reconstructed image $R_{j+1}$ is expanded to the same size of its previous level $L_j$ and adds with it. $R_n = L_n$. The final result is denoted as $R_0$. The reconstruction can be formulated as (7). The EXPAND operator is defined in (5).

$$R_j = L_j + \mathrm{EXPAND}(R_{j+1}) \tag{7}$$

The reconstruction process of (7) at level 2 of a 3-level pyramid is shown in Fig. 8. From left to right: the blending result $L_1$, the expanded previous reconstruction result $\mathrm{EXPAND}(R_2)$, the reconstruction result of present level $R_1$.

The multi-resolution spline technology based on (2) to (7) is shown in Fig. 9.

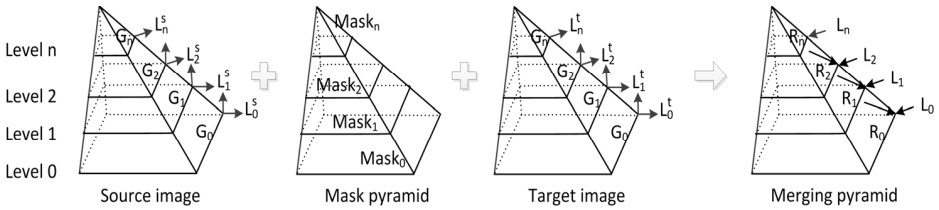
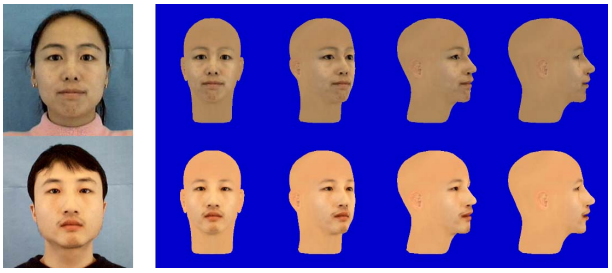



Figure 9. Multi-resolution spline technique



(a) (b)

Figure 10. The texture generation results. (a) The input frontal image. (b) The synthesized texture. (c) The reconstructed 3D head

## VI. EXPERIMENTAL RESULTS

### A. *Personalized 3D Head Models*

Some 3D head reconstruction results are shown in Fig.10. The one on the left is the input frontal image, the four images on the right are the generated virtual images with yaw of 0°, 30°, 60° and 90°.

### B. *Face Swapping*

In Fig. 11 from left to right are: the frontal view face image, the generated virtual source image, the target face image and a result generated by our system. It demonstrates our user can upload their images and do any kind of role play. The parents may generate one story book easily for the child using their favorite character. Remember, the traditional face swapping method cannot replace the faces with significant pose variations, such as the first pair in Fig. 11.

In Fig. 12, we give an example of one page of a storybook. It is from a famous Chinese novel: Romance of the Three Kingdoms.

The face swapping with real images based on our 3D reconstruction and a commercial software's reconstruction under different poses and different light conditions are shown in Fig. 13. There are four pairs swapping results. For the first three columns from left to right: the target face image, our generated virtual source image, and our face swapping result with the generated virtual source image. To compare with the two input views based face swapping, we use a commercial software to generate 3D virtual source image with one the frontal view image and one side view image, shown in the first row of Fig. 13. The fourth column of Fig. 13 shows the generated virtual source image by the commercial software. By using our face swapping method, the two input views based face swapping results can be obtained, shown in the fifth column in Fig. 13. Comparing the 3rd and 5th column of Fig. 13, it shows our results may

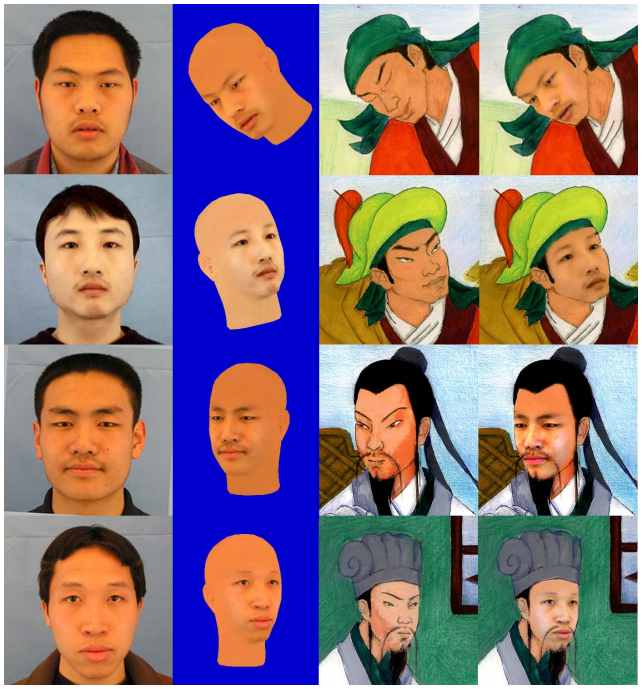


Figure 11. Role Play. From the left to right: the input frontal view face image, the rendered virtual source image, the target image and the role play results

Figure 12. An example of a storybook: Romance of the Three Kingdoms

lose some texture information only when the target face has a large yaw rotation. Otherwise, it seems our results are more realistic than the ones generated by the commercial software.

Another application here is face de-identification where faces are replaced to preserve privacy in online collections of photo. A face de-identification result comparison with [1] is done in Fig. 14. From the top to the bottom: The target image [1], the result from [1] and ours. Our result is quite good as well. The source images are from the database of [2].

In Fig. 15, it shows our method can deal with target images with large-view point differences which traditional 2D face swapping methods will fail. From left to right: the frontal view face image, the generated virtual source image,

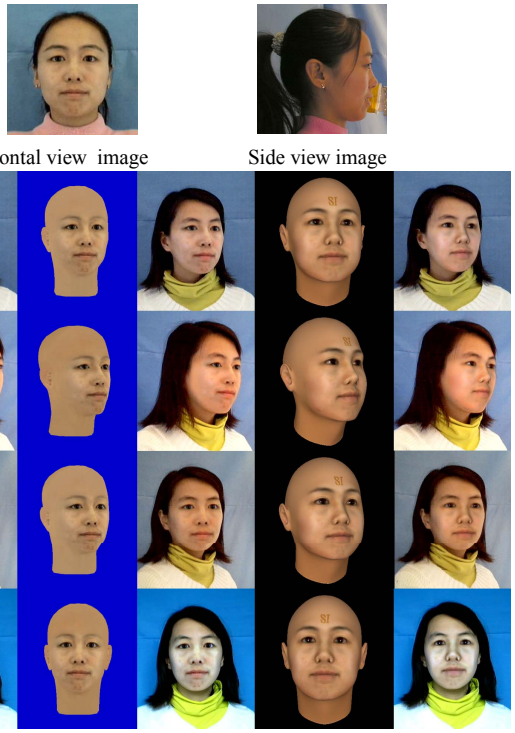



Figure 13. Face swapping. Four pairs of face swapping results comsparison. From left to right: the target image, our virtual source image generated with one frontal view image, our face swapping result, the virtual source image generated with two input views by a commerical software and the face swapping result with the software's generated virtual source face.

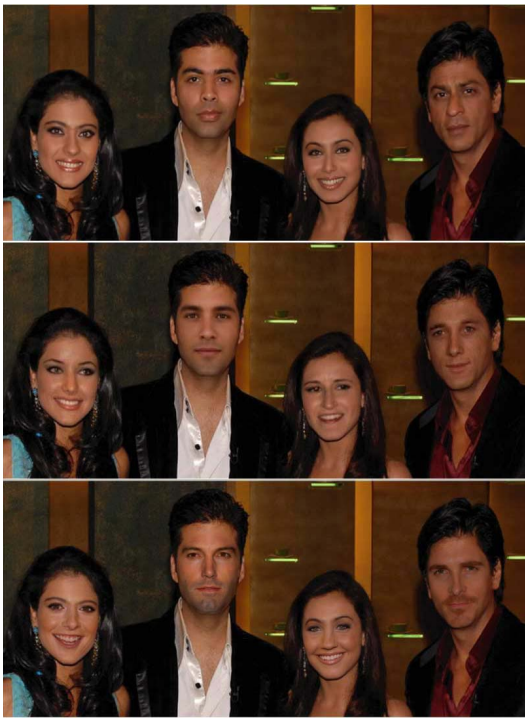


Figure 14. Face de-identification. From the top to the bottom: the target image from [1], the face de-identification result of [1] and our result. The source images of our result are from the database of [2]

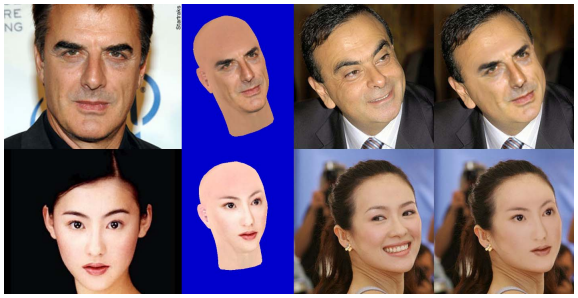


Figure 15. Face swapping between inputs with large pose variations. From the left to right: the input frontal image, the generated virtual source image, the target image, and the face replacement results of our experiment. Source and target images are from the database of [2].

the target image, and the face replacement results of our experiment. The source and target images are from the database of [2]. All the pose estimation from Fig. 11 to Fig. 15 is done with the method of [10].

Since our application is for face swapping, it is impossible to get ground-truth for the swapped face. The best way to evaluate is still subjective, whether the resulting image is visually appealing. For the fifteen face swapping results from Fig. 11 to Fig 15, fourteen testers are asked to give scores with 5, 4, 3, 2, 1. 5 points stands for the wonderful performance, and 1 point stands for the poor performance. The average score for every result is shown in Fig. 16. The total average score for all the results is 3.9 points, which shows our system is quite impressive. The scores for real images face swapping are quite high (4.5 points), while the scores for the cartoon images are relative low (3.3 points). The lowest score in Fig. 16 is for third replaced face in Fig 12. The reason is that the cartoon faces are usually quite different from real human faces. We only adopt 3 feature points to align the source face and the target face, which will generate not that satisfied result. More representing feature points will be selected in the future.

## VII. CONCLUSIONS

In this paper a 3D head model based face swapping method is proposed. It can eliminate the traditional pose and appearance similarity restriction. The results show our method is effective even when the target character has a non-frontal face. The whole 3D head model reconstruction is fully automatic and quite efficient, which takes 1.8 seconds on a 3.2GHz Intel (R) Core (TM) i5 processor. The only user interaction needed is to mask the target face region to be swapped, which can be automatized by present face detection and face segmentation method. The rest of the swapping is fully automatic. While using only one image is convenient for the user, the major limitation is that there is no information for the side face, so the reconstructed 3D on the side view may not look realistic. One solution is to use more images to incrementally refine the initial reconstructed 3D model, such as two views based 3D head reconstruction in section VI.

The main drawback of our method is that the method may fail when there are huge illumination differences between the source image and the target image. The adapted

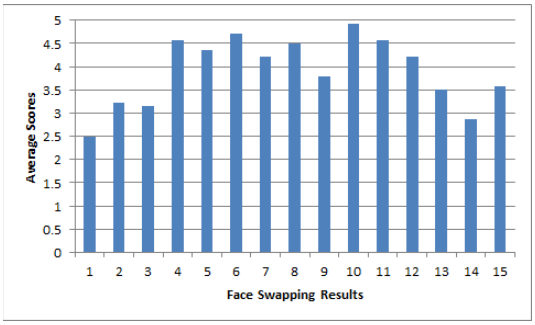



Figure 16. The average scores given by fourteen testers for fifteen face swapping results. The scores are from 5 to 1, descending with their favorated degree.

color transfer technique does not work well when the differences are too significant. In the future, more powerful lighting adjustment method will be introduced. And to make the cartoon face swapping more pleasant, more representing feature points will be adopted.

## ACKNOWLEDGMENT

This work was supported by USA Hewlett-Packard Development Company and the National High Technology Research and Development Program of China (863 program) under Grant No. 2009AA11Z214 and the National Natural Science Foundation of China under Grant Nos. 61071135 and the Doctoral Fund of Ministry of Education of China under Grant No. 20090002110077.

## REFERENCES

[1] D. Bitouk, N. Kumar, S. Dhillon, et al., "Face Swapping: Automatically Replacing Faces in Photographs", in Proc. of ACM SIGGRAPH 2008, Los Angeles, 2008, pp. 1-8.

[2] J. Zhu, L. V. Gool, S C.H. Hoi., "Unsupervised face alignment by robust nonrigid mapping", ICCV 2009, pp. 1265-1272.

[3] V. Blanz, T. Vetter, "A morphable model for the synthesis of 3D faces", in Proc. of ACM SIGGRAPH99, Los Angeles, 1999, pp. 187-194.

[4] Yi-Ting Cheng, Virginia Tzeng, Yu Liang, Chuan-Chang Wang, Bing-Yu Chen, Yung-Yu Chuang, and Ming Ouhyoung. 2009. 3D-model-based face replacement in video. In SIGGRAPH '09: Posters (SIGGRAPH '09). ACM, New York, NY, USA, , Article 29 , 1 pages.

[5] Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. 2011. Video face replacement. ACM Trans. Graph. 30, 6, Article 130 (December 2011), 10 pages.

[6] F. Pighin, J. Hecker, D. Lischinski, R. Szerisky, and D. Salesin,"Synthezing realistic facial expressions from photographs," in Proc. of ACM SIGGRAPH98, San Antonio, 1998, pp. 75–84.

[7] Liting Wang, Xiaoqing Ding, Chi Fang, "Generic face alignment using an improved Active Shape Model", ICALIP 2008, Shanghai 2008. pp. 317-321.

[8] E. Reinhard , M. Adhikhmin, B. Gooch, et al., "Color transfer between images", IEEE Computer Graphics and Applications, vol. 21, no. 5, pp. 34-41, 2001.

[9] P. J. Burt, E. H. Adelson, "A multiresolution spline with application to image mosaics", ACM Trans on Graphics, vol. 2, no. 4, pp. 217-236, 1983.

[10] Yali Li, Shengjin Wang, Xiaoqing Ding, "Person-Independent Head Pose Estimation Based On Random Forrest Regression", ICIP 2010, pp.1521-1524