Face Replacement and Image Animation System in Cultural Experience

Jiayue Li1*, Qing Wang1, Shiji Li1, Qian Zhou1 and Qiang Zhong1

1Department of Computer Science and Technology, China Agricultural University, Beijing 100083, China
*S20193081370@cau.edu.cn

Abstract. As a unique representative of Chinese culture and civilization, Chinese traditional portrait painting has a certain cultural influence all over the world. However, what kind of human-computer interaction system can attract more people's interest and attention, is still a field worthy of exploration. This paper focuses on the instantiation expression of Chinese traditional portrait, improves the regional feature point matching based on the research of Photo to Chinese Portrait method (P-CP), compares the effects of different color correction approaches on the experimental results, and introduces the form of real-time image driving animation. We hope that we can provide new inspiration for the dissemination of traditional culture and the application of modern technology.

1. Introduction
Inspired by the cardboard cutout for group photo in amusement parks and museums, this paper takes Chinese traditional portrait painting as an example, and also focuses on face replacement and image animation application in culture experience scenes.

In our previous research, we have combined traditional Chinese portrait with fast neural style transfer, completed the face replacement from photo to portrait painting, which provides gratifying output application. In this paper, we have done more in-depth research on face replacement and added a new form of interaction. The main contributions of this paper are summarized as follows.

a) We introduce Delaunay Triangulation to make the effect of facial structure replacement more natural and reasonable. The new method has the characteristics of local adaptability in technology, and the output result has more aesthetic value.

b) We compare the output effect and generation time of the combination of different feature point alignment approaches and color correction approaches, and put forward suggestions for application scenarios based on effect first and generation efficiency first.

c) Based on the static face replacement portrait, we further study the implementation of dynamic effect. By introducing First Order Motion Model, the effect of driving the target picture in real time is achieved, and the existing forms of ancient cultural experience are enriched.

2. Methods

2.1 Neural Style Transfer and Face Replacement
In order to make use of neural style transfer[1-3] to extract image style, we have redesigned it based on Johnson network[4], introduced instance normalization[5], fine tuned it, and proposed P-CP method. Our former P-CP architecture involves the training phase and the testing phase. The purpose
of training phase is to generate the Chinese portrait style transfer model. The purpose of the testing phase is to generate Chinese portrait style picture with the pretrained model, and also the Immersive Chinese Portrait Picture through the face replacement functional structure.

Face replacement refers to replacing the target identity with the content of the source identity, so that the target identity has partial characteristics of the source identity. Face replacement, along with reenactment, editing and synthesis, emerge in endlessly on human face visual research field[6].

The difficulty of face replacement lies in the differences of human face structure, skin color presented in pictures, skin texture and relative posture between face and camera. In this paper, we introduce triangular transformation to deal with the first problem, color correction to deal with the second, and style transfer to deal with the third problem. The last question is the direction to be studied in the future.

2.2 Delaunay triangle transform and color correction

To replace one face with another(see Figure 1(a)), we first use dlib[7] to detect 68 corresponding facial feature points automatically for each facial image. With two sets of 68 points (shown in blue in Figure 1(b)), and 8 points on the boundary of the output image (shown in green) — one set per image, we can calculate the average of corresponding points in the two sets and obtain a single set points.

On this set of average points, we calculate the convex hull and perform Delaunay Triangulation. The result of Delaunay triangulation is a list of triangles represented by the indices of points in the points array. The Delaunay triangulation used in this paper is the Bowyer-Watson algorithm in OpenCV. The triangles in the two images capture approximately similar regions. Therefore, we can use these triangles regions for Face Morphing.

Given two images $I$ and $J$, an in-between image $M$ is created by blending images $I$ and $J$, which is Face Morphing. The images $I$ and $J$ have pixel $(x_i, y_i)$ and $(x_j, y_j)$ correspondingly, and the blending of them is controlled by a parameter $\alpha$ that is between 0 and 1 ($0 \leq \alpha \leq 1$). The final output image is composed of the average of the pixel intensities from the target image and the source image.

2.3 Poisson image editing and histogram match

Here we compare two color correction processes. One is Poisson Image Editing[8] from OpenCV’s seamless clone, the other is histogram match.

By solving a Poisson equation, the central insight of Poisson Image Editing is that working with image gradients instead of image intensities can produce much more realistic results. The gradient of the result image in the masked region is about the same as the gradient of the source region in the masked region. Additionally, the intensity of the result image at the boundary of the masked region is the same as the intensity of the destination image.

Local histogram matching matches the histogram of the image region to another image, and changes the RGB color distribution, so as to adjust the color and illumination information to the specified style.
2.4 First Order Motion Model

Image animation is the task of automatically synthesizing video by combining the appearance extracted from the source image with the motion mode extracted from the driving video. This application is a recently popular form of human-computer interaction. Taking expression migration as an example, the traditional methods of image generation tasks, such as VAE and GAN, need a large number of face images and the annotation information of these face images (key points, facial motion units, three-dimensional models, etc.). The model proposed in [9, 10] can extract the motion pose information of the object without the prior information and label information of the source image.

The first order motion model(see Figure 2) is a relatively mature open source image animation model. Without using prior information such as annotations or calibration points, the model can obtain a motion model similar to any object through video training in the same category. The model uses self-supervised learning method to decouple the appearance surface and motion information, and uses multiple learned key points and local radial transformation to generate the target motion model. Finally, it can achieve the effect of driving similar static pictures with video.

3. Results & Discussion

Our experiments are carried out on Windows 10, CUDA 10.1.243 and cuDNN 7.6.5.32, which are conducted on a PC Intel Core i5 processor and NVIDIA GeForce GTX 1070.

3.1 Morph degree experiment results

In general, Face Morph is divided into six steps: find point correspondences using facial feature detection, determine convex hull, perform Delaunay triangulation, calculate affine transforms, correct color and alpha blend warped images.

In the process of calculating region deformation and image fusion, the value of alpha directly affects the rendering effect of the output image. When the alpha is closer to 0 or 1, the facial features...
of the output picture will be closer to the original painting or user source picture (see Figure 3). When the alpha is closer to 0, the fusion result tends to the original painting, and the fusion brings more artifacts, but the overall view will not be affected when using the output with small resolution. When the alpha is closer to 1, the fusion result tends to the user source image with less artifacts, but the color mismatch problem is more prominent. Therefore, in the later experiments, we choose the relative compromise value alpha = 0.8.

3.2 Face replacement experiment results

| User Source | Painting Source of Different Size (pixels) | 2048<sup>2</sup> | 1024<sup>2</sup> | 512<sup>2</sup> | 256<sup>2</sup> |
|-------------|------------------------------------------|----------------|----------------|----------------|----------------|
| Face Morph  | Generation Time (s)                       | 9.34           | 4.49           | 3.88           | 2.83           |
| Face Alignment and Affine Transformation (FAAT) | | 35.99           | 11.79          | 6.51           | 5.31           |
| Delaunay Triangulation (DT) | | 20.03           | 10.62          | 7.97           | 7.23           |
| Color Correction | Poisson Image Editing (PIE) | 76.21           | 24.10          | 11.74          | 7.91           |
| | Histogram Matching (HM) | | | | |

User source has normal user source and style transfer user source two kinds of inputs. Face morph has face alignment & affine transformation(FAAT) and Delaunay Triangulation(DT) two strategies. Color correction has Poisson image editing(PIE) and histogram matching(HM) two strategies. Therefore, we have 8 treatments for each user, and their output results are listed in Table 1 and Figure 4. Limited to the article length, Figure 4 only lists results with style transfer user source.

In terms of generation time, the choice of user source has little effect on it. Meanwhile, when we choose painting sources of different sizes, it matters if it is larger than 512×512, and the generation time is proportional to its size. If the other conditions are consistent, DT takes about 2 times as long as FAAT, and HM takes about 3 times as long as PIE.

In terms of the quality of generated pictures, pictures that have used style migration can bring more traditional Chinese painting textures. The output of Delaunay triangle transformation is more in line with the user's face characteristics, and histogram matching can make the color distribution more in line with the color application of traditional Chinese painting.
In general, for better output effect, we should consider the combination of style transfer user source, Delaunay triangulation (DT) and histogram matching (HM). For higher output efficiency, we should consider using small size printing source input, normal face alignment & after transformation (FAAT) and position image editing.

3.3 Image animation experiment results

The picture on the left of Figure 5 show the previously generated target. The first row shows the driving videos. The bottom row contains the corresponding animated sequence with motion transferred from the driving video and object taken from the target image. The output result can convert the facial expression information of Driving Source to Target in real time, and the fluency is
within an acceptable range. However, this method also faces the limitation that the distance between the user and the camera needs to be within a certain range.

4. Conclusion
This paper improves our former Photo to Chinese Portrait model, introduces triangular transformation to deal with the differences of human facial structure, compares different color correction approaches to deal with the skin color differences presented by pictures, and adds a real-time animation driven model to enrich the form of human-computer interaction and enhance the immersion of cultural experience for users. Finally, the differences of skin texture and the relative posture between face and camera are still the focus of our next research.

Acknowledgments
Thanks to my mentor Qing Wang and my colleagues. Thanks for their valuable suggestions for my research and my article.

References
[1] Gatys, L. A., Ecker, A. S. and Bethge, M. (2015) Texture Synthesis Using Convolutional Neural Networks. arXiv preprint arXiv:1505.07376.
[2] Gatys, L. A., Ecker, A. S. and Bethge, M. (2015) A Neural Algorithm of Artistic Style. Journal of Vision.
[3] Gatys, L. A., Ecker, A. S. and Bethge, M. (2016) Image Style Transfer Using Convolutional Neural Networks. In: Conference on Computer Vision and Pattern Recognition (CVPR): IEEE.
[4] Johnson, J., Alahi, A. and Li, F., Perceptual Losses for Real-Time Style Transfer and Super-Resolution, in Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling^Editors. 2016. p. 694-711.
[5] Ulyanov, D., Vedaldi, A. and Lempitsky, V. (2016) Instance Normalization: The Missing Ingredient for Fast Stylization.
[6] Mirsky, Y. and Lee, W. (2021) The Creation and Detection of Deepfakes: A Survey. ACM computing surveys, 54(1): p. 1-41.
[7] Kazemi, V. and Sullivan, J. (2014) One Millisecond Face Alignment with an Ensemble of Regression Trees. In: IEEE Conference on Computer Vision and Pattern Recognition.
[8] Perez, P., Gangnet, M. and Blake, A. (2003) Poisson Image Editing. ACM TRANSACTIONS ON GRAPHICS, 22(3): p. 313-318.
[9] Siarohin, A., Lathuiliere, S., Tulyakov, S., and Ricci, E., et al., First Order Motion Model for Image Animation, in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, and F. D'Alche-Buc, et al., H. Wallach, H. Larochelle, A. Beygelzimer, and F. D'Alche-Buc, et al.^Editors. 2019.
[10] Shysheya, A., Zakharov, E., Aliev, K., and Bashirov, R., et al., Textured Neural Avatars, in IEEE Conference on Computer Vision and Pattern Recognition. 2019. p. 2382-2392.