Pacific Graphics (2022)
Y. Yang, A. D. Parakkat, B. Deng, and S.T. Noh (Editors)

Work in Progress

DFGA: Digital Human Faces Generation and Animation from the RGB Video using Modern Deep Learning Technology

Diqiong Jiang [1], Lihua You[2], Jian Chang[2], Ruofeng Tong[1]

[1]Zhejiang University , China
[2]Bournemouth University, UK

Abstract

High-quality and personalized digital human faces have been widely used in media and entertainment, from film and game production to virtual reality. However, the existing technology of generating digital faces requires extremely intensive labor, which prevents the large-scale popularization of digital face technology. In order to tackle this problem, the proposed research will investigate deep learning-based facial modeling and animation technologies to 1) create personalized face geometry from a single image, including the recognizable neutral face shape and believable personalized blendshapes; (2) generate personalized production-level facial skin textures from a video or image sequence; (3) automatically drive and animate a 3D target avatar by an actor's 2D facial video or audio. Our innovation is to achieve these tasks both efficiently and precisely by using the end-to-end framework with modern deep learning technology (StyleGAN, Transformer, NeRF).

1. Instruction

In recent years, digital human body technology has been developed rapidly, and many technology companies have invested a lot of money and energy in researching digital human technology. However, creating a digital human is extremely labor-intensive, requiring comprehensive facial assets captured with complex hardware and hundreds of domain experts such as artists, programmers, and technical artists. To popularize digital humans to daily consumer users, the cost of generating digital humans must be reduced. Consumer-grade cameras and computing resources make it possible to achieve efficient and lost-cost modeling and animation of digital humans. To this aim, the system called DFGA will be developed from deep learning, which accepts the RGB video as input and reconstructs its 3D avatar as well as personalized blendshapes.

2. Related Work

Personalized dynamic texture. Synthesizing dynamic face texture is essentially image-to-image translation when using deep learning. In recent years, 2D face image generation has made significant progress, especially after the proposal of StyleGAN [KLA19]. Previous works capitalize on the power of 2D face generators to recover unseen parts of a human face. However, the pre-trained StyleGAN lacks expression richness because of its training set, making it difficult for StyleGAN to reconstruct facial textures with detailed expressions. NeRF [MST*20] has been explored to develop 3D-aware image synthesis techniques with powerful interpolation ability. However, its extrapolation ability is fragile and sufficient multi-view images are required. The proposed research will use StyleGAN, which contains many high-definition face priors, to augment data to generate more multi-view high-fidelity facial images and generate the face texture using NeRF. It can also efficiently deal with realistic hair and teeth and wrinkles with less data requirement.

Face animation. Abundant of works animate the 3D actor's 2D facial video or audio. FaceFormer [FLS*22a] encodes the long-term audio context and the history of face motions to autoregressively predict a sequence of animated 3D face meshes. Fan et al. [FLS*22b] proposed incorporating the contextual embeddings from transformer-based GPT-2 to understand the emotional context, to produce a more diverse range of facial expressions. The proposed research will use the transformer-based to understand the emotional context of the sequence and extract the meaning feature.

3. Research Methodology and approach

Digital neutral human face generation. We use existing methods with minor modifications for digital neutral human face generation. The StyleGAN can generate high-fidelity multi-view images, and the texture information from those images can be obtained by projecting the texture to a 3D face mesh. Specifically, a face image is first encoded to W space with a style encoder, and then the network is used to generate the image of the face rotated to a predetermined angle. Finally, 3DMM fitting and texture-stitching are used to synthesize a whole face from different angles.

Personalized human face generation: Modeling variational face expression is another important part of face geometry reconstruction. Recent works [BCLT21, LKZ*20] have shown how to

DOI: 10.2312/pg.20221249

Method	Ostec	Ours
Time(s)	984	3

Table 1: *Running Time of the methods.*

automatically build personalized blendshapes using a deep neural network. The proposed research will follow their method and customize the face rig by modifying the generic model.

Dynamic texture generation: The second task is to generate personalized production-level dynamic facial skin textures. Given a single face input image in a neutral pose, [GDZ21, LCK*20] generated the high-quality face texture based on StyleGAN. [GTZN21] used Neural Radiance Fields to generate the realistic texture, ensuring multi-view consistency. However, they need more sufficient data. Because StyleGAN does not have the ability to generate expression details, the proposed research will use expressive three-dimensional face data to fine-tune the StyleGAN to generate the dynamic texture. The main composition of the framework is the conditional StyleGAN, which uses personalized blendshapes and natural albedo maps as the conditions to generate dynamic textures. The data required for training are face images of persons and their corresponding dynamic texture. These data are used to fine-tune the weights of StyleGAN so that StyleGAN has the ability to generate dynamic textures. NeRF will also be used to augment data.

Personalized blendshapes weight regression: While generating personalized blendshapes and personalized production-level dynamic facial skin textures as the face rig, the 3DMM parameters and facial landmarks will be extracted from a driving video or pre-defined facial action sequences. Fellow this information will be inputted to regress the weight of the face blendshapes of a target avatar. After that, a 3D target avatar will be automatically driven and animated by an actor's 2D facial video similar to [BCLT21, LKZ*20].

Detail animated map generation: To get expression-dependent details such as wrinkles, the proposed framework will learn an expression-conditioned detail model to infer facial details from a person-specific detail latent space and an expression space. Similar to [FFBB21], identity information will be added to generate animated displacements. However, different from them, the proposed method will use the transform structure. This transformer-based structure will be better to understand the emotional context of the sequence and extract the meaning feature.

4. Experiment

Our current work only completes the section of the static digital faces generation. The results of the generated static numbers are shown in Figure 1. Later, we will generate 3D face animation based on the static face. We use a neural network to regress the parameters of Flame [LBB*17] and reconstruct the static face geometry. To generate a static face texture, we use the method . Our method of reconstructing texture does not require any dataset. The Ostec [GDZ21] is similar to our method. But as shown in Table 1, our speed is significantly faster than theirs because our method does not require an iterative optimization process.

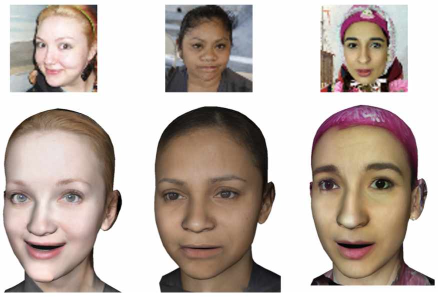

Figure 1: *The results of our the static digital faces generation. The first row is input images. The second row is reconstructed faces.*

5. Conclusion

The proposed research will investigate deep learning-based facial modeling and animation technologies to simplify and speed up these processes. The results show that our method is effective.

References

[BCLT21] Bai Z., Cui Z., Liu X., Tan P.: Riggable 3d face reconstruction via in-network optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 6216–6225. 1, 2

[FFBB21] Feng Y., Feng H., Black M. J., Bolkart T.: Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG) 40*, 4 (2021), 1–13. 2

[FLS*22a] Fan Y., Lin Z., Saito J., Wang W., Komura T.: Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18770–18780. 1

[FLS*22b] Fan Y., Lin Z., Saito J., Wang W., Komura T.: Joint audio-text model for expressive speech-driven 3d facial animation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques 5*, 1 (2022), 1–15. 1

[GDZ21] Gecer B., Deng J., Zafeiriou S.: Ostec: One-shot texture completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 7628–7638. 2

[GTZN21] Gafni G., Thies J., Zollhofer M., Niessner M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 8649–8658. 2

[KLA19] Karras T., Laine S., Aila T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 4401–4410. 1

[LBB*17] Li T., Bolkart T., Black M. J., Li H., Romero J.: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph. 36*, 6 (2017), 194–1. 2

[LCK*20] Lee M., Cho W., Kim M., Inouye D., Kwak N.: Styleuv: Diverse and high-fidelity uv map generative model. *arXiv preprint arXiv:2011.12893* (2020). 2

[LKZ*20] Li J., Kuang Z., Zhao Y., He M., Bladin K., Li H.: Dynamic facial asset and rig generation from a single scan. *ACM Trans. Graph. 39*, 6 (2020), 215–1. 1, 2

[MST*20] Mildenhall B., Srinivasan P. P., Tancik M., Barron J. T., Ramamoorthi R., Ng R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision* (2020), Springer, pp. 405–421. 1