Landmark-Free Facial Motion Transfer to VTubers

Shangzhe Di Sho Maeoki

1 Introduction

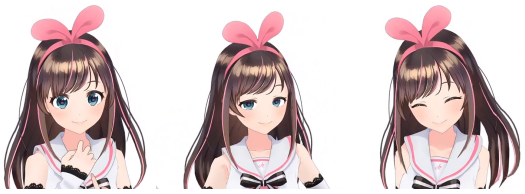

Figure 1: Some images of Kizuna AI, one of the most famous VTubers. As seen in the images, facial expressions as well as motion of hairs and arms are reflected in this application, while we solely focus on facial expressions.

Virtual YouTubers (or VTubers) are YouTubers represented by digital avatars generated by computer graphics (as shown in Fig. 1) [1]. They are created and managed by humans who use expensive motion capture devices and RGBD cameras, so the avatars mirror their real-time movements. Video cameras, however, are more widely available on PCs and mobile devices than those costly devices. Thus, we attempt to use a single video camera to achieve automatically facial tracking and animation so that VTubers can be developed in consumer-level applications. Our goal is to animate digital avatars based on RGB face videos. We approached this project by estimating pose and 3D expression coefficients from RGB video frames. The obtained parameters are then used to drive a 3D avatar via Unity.

2 Related Work

Cao *et al.* [2] demonstrates impressive real-time facial tracking and animation results and can robustly handle fast motions, large head rotations, and exaggerated expressions. They first detect face landmarks from input frames and regress pose and expression coefficients based on the landmarks; obtained parameters can be directly transferred to a digital avatar to drive its face. Thanks to research developments in the domain of computer graphics, there are a series of works which utilize convolutional neural networks (CNN). Zhu *et al.* [3] trained a CNN to regress parameters for the 3D morphable model (3DMM) [4], including pose, shape, and expression, taking face landmarks as input. However, Zhu *et al.* still rely on face landmarks, detection of which can be a bottleneck in the pipeline due to scale and occlusions. Meanwhile, there are some studies which use CNN to directly estimate 3DMM coefficients from an image (e.g., estimating $6DoF$ pose coefficients [5], and $29D$ expression coefficients [6].) These CNN-based methods do not rely on facial landmarks and can estimate coefficients for occluded faces appearing in

unprecedented in-the-wild viewing conditions. Therefore, we adopt these two approaches and use the predicted pose and expression coefficients to drive our VTuber model.

3 Methodology

We use the standard linear 3DMM representation to describe a human face (here, ignoring the pose parameters):

$$S' = \hat{s} + S\alpha + E\eta \tag{1}$$

where 3D expression deformations are provided as an linear combination of expression coefficients $\eta \in R^{29}$ and expression components $E \in R^{3n \times m}$. The VTuber model we used is based on blendshapes, which can be described as:

$$S = Bw \tag{2}$$

where B is a matrix containing 62 individual blendshapes, and w is blendshape parameters. Both of them use a linear combination to describe a face model. Thus, it is possible to transfer the expression coefficients of a human face to the blendshape parameters of a VTuber.

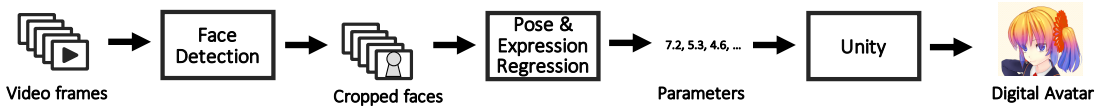

Figure 2: The pipeline of the approach.

The pipeline of this project is illustrated in Fig. 2, which can be divided into three parts. First, video frames are input to the Face Detection module to crop face regions. Second, the cropped face images are then fed into the Pose & Expression Regression module for estimating $6DoF$ pose and $29D$ expression coefficients, which is described by 3DDFA [7]. Finally, the obtained parameters are utilized to drive the 3D avatar via Unity. Note that the common used blendshape system for avatars does not share the same expression representation with the $29D$ coefficients, so we manually find the correspondence between them and use the converted blendshape parameters to drive the VTuber model via Unity. We used OpenCV for Face Detection module, and codes offered by Chang *et al.* [6, 5] for the Pose & Expression Regression module.

4 Results and Discussion

The expression transfer from a human face to a VTuber is a tricky problem because VTubers usually have abstract and extravagant faces. We have to manually find the correspondences between the 3DDFA expression bases and our avatar's blendshapes, which can cause wrong motions and not-so-correspondent expressions on the avatar. Fig. 3 shows two expression transfer results. To a certain extent, the result on the

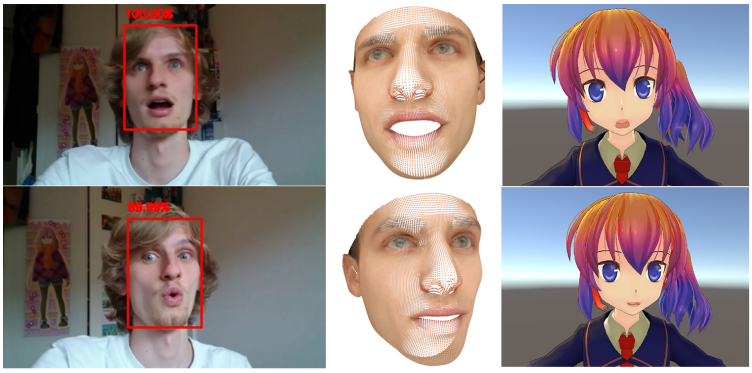

Figure 3: *Two expression transfer results*, including the input frame (left), the reconstructed 3D face (middle), and the 3D VTuber's face (right). The good matching result is on the top row, whereas the bad matching result is on the bottom.

top row has a good match between the expression of the VTuber and the one on the human face. The bottom row shows a mismatch situation. In this situation, the face reconstruction result does not match with the human face. Thus, the VTuber can not present the proper expression. This mismatch shows a weakness of the ExpNet that to some degree, it can not fit well on exaggerated and complicated expressions.

5 Conclusion

We have tackled a challenging Virtual YouTuber project based on CNN. In this project, we developed a program which utilizes input RGB videos to drive a 3D avatar based on pose and expression coefficient parameters. The work performs fairly well while there is room for improvement.

References

[1] Kazuaki Nagata. Japan's latest big thing: 'virtual YouTubers'. In The Japan Times, 2018. `https://www.japantimes.co.jp/news/2018/07/17/national/japans-latest-big-thing-virtual-youtubers/#.XPjLS9P7SCQ`.

[2] Chen Cao, Qiming Hou, and Kun Zhou. Displaced Dynamic Expression Regression for Real-time Facial Tracking and Animation. *ACM Transactions on graphics (TOG)*, 33(4):43, 2014.

[3] High-Fidelity Pose and Expression Normalization for Face Recognition in the Wild, author=Zhu, Xiangyu and Lei, Zhen and Yan, Junjie and Yi, Dong and Li, Stan Z. In *CVPR*, 2015.

[4] Volker Blanz, Thomas Vetter, et al. A Morphable Model for the Synthesis of 3D Faces. In *SIGGRAPH*, 1999.

[5] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. FacePoseNet: Making a Case for Landmark-Free Face Alignment. In *ICCV Workshop*, 2017.

[6] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. ExpNet: Landmark-Free, Deep, 3D Facial Expressions. In *FG*, 2018.

[7] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face Alignment Across Large Poses: A 3D Solution. In *CVPR*, 2016.