Everything's Talkin': Pareidolia Face Reenactment

Supplementary Material

Linsen Song[1,2*] Wayne Wu[3,4*] Chaoyou Fu[1,2] Chen Qian[3] Chen Change Loy[4] Ran He[1,2†]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]NLPR & CRIPAC, CASIA [3]SenseTime Research [4]Nanyang Technological University
songlinsen2018@ia.ac.cn, {wuwenyan,qianchen}@sensetime.com,
{chaoyou.fu,rhe}@nlpr.ia.ac.cn, ccloy@ntu.edu.sg

Abstract

This supplementary materials provide some details of our work that are not be clearly presented due to space limitation in our main paper: Sec. 1 describes the details about n-order composite Bézier curve fitting, how the motion decay along the curve scale and the architecture of the Autoencoder Network $\mathcal{G}$; Sec. 2 elaborates the details of our implementation; Sec. 3 presents the limitation and some failure cases of our method.

1. More Details of the Methodology

1.1. n-order Composite Bézier Curve Fitting

For a frame $\mathbf{H}$ from human portrait video, we assume that there are N_H branches in its boundaries $\mathcal{S}_{\mathbf{H}} = \{C_i\}_{i=1}^{N_H}$. We use n_i-order composite Bézier curve to fit the boundary branch C_i and denote the estimated composite Bézier curve as B_i. The overall optimization problem can be written as follows:

$$\min \sum_{i=1}^{N_H} ||C_i - B_i||^2, \tag{1}$$

where a composite Bézier curve B_i is composited by N_i vanilla Bézier curves and we denote these vanilla Bézier curves as B_{ij} $(i \leq N_S, j \leq N_i\})$. The composite Bézier curve B_i are splitted to N_i Bézier curves by $N_i - 1$ *joints*. There are also $N_i - 1$ joints on the boundary branch C_i that correspond to the joints on B_i. Thus, C_i is splitted as C_{ij} $(i \leq N_S, j \leq N_i\})$. Thus, the overall optimization problem is formed as:

$$\begin{gathered}\min \textstyle\sum_{i=1}^{N_H} ||C_i - B_i||^2 = \min \textstyle\sum_{i=1}^{N_H} \sum_{j=1}^{N_i} ||C_{ij} - B_{ij}||^2 \\ \Leftrightarrow \min ||C_{ij} - B_{ij}||^2, \ \forall i, j,\end{gathered} \tag{2}$$

where B_{ij} is a vanilla Bézier curve and in Eq. (2) the original optimization problem is splitted into $N_H \times N_i$ independent optimization problems. Thus, we consider the new optimization problems $\min ||C_{ij} - B_{ij}||^2, \ \forall i, j$. We omit the subscripts i, j for simplicity.

According to the definition of n-order Bézier curve, a Bézier curve B can be rewritten as follows,

$$B(\tau) = \sum_{k=0}^{n} \binom{n}{k} \tau^k (1-\tau)^{n-k} P_k, \tag{3}$$

where $\tau \in [0, 1]$ represents the relative position of point $B(\tau)$ on curve B and $\binom{n}{k}$ is the number of k-combinations. If we denote the components on x, y, z axis of B as B^x, B^y, B^z and the components on x, y, z axis of P_k as P_k^x, P_k^y, P_k^z respectively, *e.g.*, $B = (B^x, B^y, B^z)$, $P_k = (P_k^x, P_k^y, P_k^z)$. Then Eq. (3) can be rewritten as:

$$\left\{\begin{array}{ll} B^x(\tau) & = \sum_{k=0}^{n} \binom{n}{k} \tau^k (1-\tau)^{n-k} P_k^x \\ B^y(\tau) & = \sum_{k=0}^{n} \binom{n}{k} \tau^k (1-\tau)^{n-k} P_k^y \\ B^z(\tau) & = \sum_{k=0}^{n} \binom{n}{k} \tau^k (1-\tau)^{n-k} P_k^z \end{array}\right. , \tag{4}$$

similarly, we denote boundary $C = (C^x, C^y, C^z)$ where C^x, C^y, C^z are the boundary C's components on axis x, y, z. We denote a point on the skeleton C as $C(\tau) = (C^x(\tau), C^y(\tau), C^z(\tau))$ $(\tau \in [0, 1])$. The optimization problem in Eq. (2) as follows:

$$\min ||C - B||^2 \Leftrightarrow \left\{\begin{array}{l} \min ||C^x - B^x||^2 \\ \min ||C^y - B^y||^2 \\ \min ||C^z - B^z||^2 \end{array}\right. . \tag{5}$$

[*] Equal Contribution

[†] Corresponding Author

For simplicity, we will optimize $\min||C^x - B^x||^2$ as example and the optimization problem can be expanded as follows:

$$\begin{gathered}\min||B^x - C^x||^2 = \int_0^1 ||B^x(\tau) - C^x(\tau)||^2 d\tau = \\ \lim_{\mathrm{card}(T)\to\infty} \sum_{\tau_i \in T} ||B^x(\tau_i) - C^x(\tau_i)||^2 = \\ \lim_{\mathrm{card}(T)\to\infty} \sum_{\tau_i \in T} || \sum_{k=0}^{n} \binom{n}{k} \tau_i^k (1-\tau_i)^{n-k} P_k^x - C^x(\tau_i)||^2\end{gathered}, \tag{6}$$

where $T = \{\tau_0, \tau_1, \cdots, \tau_m\}$ is a set of uniformly sampled points of $\tau \in [0, 1]$ and $\mathrm{card}(T) = m + 1$ is the cardinality of T. If we denote $a_{ik} = \binom{n}{k}\tau_i^k(1-\tau_i)^{n-k}$. Thus, we have:

$$\begin{gathered}\sum_{\tau_i \in T} || \sum_{k=1}^{n} a_{ik} P_k^x - C^x(\tau_i)||^2 = ||\mathbf{A}\mathbf{p} - \mathbf{b}||^2 = \\ \left\| \begin{bmatrix} a_{00} & \cdots & a_{0n} \\ \vdots & \ddots & \vdots \\ a_{m0} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} P_0^x \\ \vdots \\ P_n^x \end{bmatrix} - \begin{bmatrix} C^x(\tau_0) \\ \vdots \\ C^x(\tau_m) \end{bmatrix} \right\|^2\end{gathered}, \tag{7}$$

where $\mathbf{A} \in \mathbb{R}^{(m+1)\times(n+1)}$, $\mathbf{p} \in \mathbb{R}^{(n+1)\times 1}$, $\mathbf{b} \in \mathbb{R}^{(m+1)\times 1}$. We need to solve the $\mathbf{p}$ by minimizing $||\mathbf{A}\mathbf{p} - \mathbf{b}||^2$. Thus, the optimization problem in Eq. (1) is converted to solve the least square problem Eq. 7. The solution $\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} ||\mathbf{A}\mathbf{p} - \mathbf{b}||^2$ can be computed by Gauss–Newton algorithm or $\hat{\mathbf{p}} = \mathbf{A}^{\dagger}\mathbf{p}$ where $\mathbf{A}^{\dagger}$ is the pseudo inverse matrix of $\mathbf{A}$.

1.2. Motion Decay Along Curve Scale

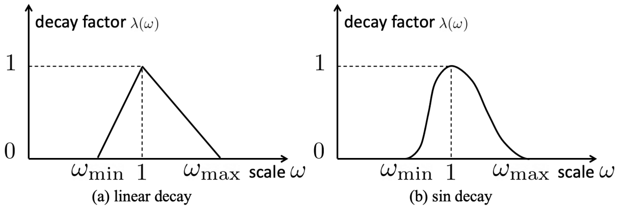
In the main paper, the motion at point $B_i(1, \tau_i)$ is denoted as $\mathcal{M}^e_{B_i(1,\tau_i)}$. The point $B_i(1, \tau_i)$ lies at a composite Bézier curve B_i that correspond to a motion seed. The motion $\mathcal{M}^e_{B_i(1,\tau_i)}$ will decay when it spreads from $B_i(1, \tau_i)$ to $B_i(\omega_i, \tau_i)$. We have the following motion decay function:

$$\mathcal{M}^e_{B_i(\omega_i,\tau_i)} = \lambda(\omega_i) \cdot \mathcal{M}^e_{B_i(1,\tau_i)}, \tag{8}$$

where $\lambda(\omega_i)$ is the motion decay factor that is determined by the curve scale factor ω_i. In practice, we design two different decay functions as Fig. 1 shows. In case that the motion seed of the mouth spreads to the eyes area, we use $\omega_{\min}$ and $\omega_{\max}$ to restrict the area to where a motion seed can spread. We find that performances of the *linear decay* and *sine decay* functions are similar. Thus, we use the more simplified linear decay function in our experiments. We leave the exploration of decay functions for the future work.

1.3. Architecture of the Autoencoder $\mathcal{G}$

The architecture of the Autoencoder network $\mathcal{G}$ is demonstrated in Tab. 1. In the table, the **Resolution** denotes the spatial resolution of the feature map. **EncBlock** denotes a 2D convolutional layer (stride is 2, padding is 1, kernel size is 4×4). **DecBlock** denotes a 2D convolutional layer (stride is 2, padding is 1, kernel size is 4×4), followed by a PixelShuffle [1] layer (upscale factor is 2).



(a) linear decay

(b) sin decay

Figure 1: **Motion Decay.** (a) linear decay: the motion linearly decay with the curve scale when it deviate 1. (b) sin decay: we use the sine function to describe how the motion decays smoothly.

Table 1: **Architecture of the AutoEncoder Network $\mathcal{G}$**

Layer Name	Resolution	Layer Structure
Input	256×256	Input Image
$\mathcal{E}_1$	128×128	EncBlock 3 → 64 + LeakyReLU(0.2)
$\mathcal{E}_2$	64×64	EncBlock 64 → 128 + LeakyReLU(0.2)
$\mathcal{E}_3$	32×32	EncBlock 128 → 256 + LeakyReLU(0.2)
$\mathcal{D}_3$	64×64	DecBlock 256 → 128 + ReLU
$\mathcal{D}_2$	128×128	DecBlock 128 → 64 + ReLU
$\mathcal{D}_1$	256×256	DecBlock 64 → 3

2. Implementation Details

In the Parametric Shape Modeling, we find that it is hard to define nose, ears, eyebrows and jawline for pareidolia faces. Thus, we only animate the mouth and eyes of pareidolia faces. The mouth of a pareidolia face is animated by the inner lip boundary of the given human video. For the mouth or eyes of a human/pareidolia face, its boundary is splitted into two branches (upper and lower halves) and each branch is parameterized as a composite Bézier curve. In practice, we find that each branch of the mouth/eye can be precisely parameterized by a composite Bézier curve defined by 5-7 control points.

In the Expansionary Motion Transfer, for any pixel location $\mathbf{p}$, we compute ω_i, τ_i that correspond to the composite Bézier curve B_i (related to motion seed), which is prepared for our motion spread strategy. Both the motion field and inverse motion field are computed on the discrete pixel grid. We regard the motion field $\mathcal{M}^e$ as a function of the pixel grid and infer the inverse motion field $\overleftarrow{\mathcal{M}^e}$ as the inverse function of $\mathcal{M}^e$. We use first-order difference of $\mathcal{M}^e$ in $\frac{d}{d\mathbf{p}}\overleftarrow{\mathcal{M}^e}(\mathbf{p})$. In experiments, we find that the First-order Motion Approximation works well when $||\Delta\mathbf{p}|| = 1, 2$ and increasing $||\Delta\mathbf{p}||$ does not bring further improvement.

In the Unsupervised Texture Animator, the image resolution is 256×256 for all the input human/pareidolia faces and the resolution of the output pareidolia faces is 256×256, too. During training the Autoencoder network $\mathcal{G}$, we set the loss weights $\alpha_1 = \alpha_2 = 1$ empirically.

3. Limitation

large poses of human faces: The facial motion extracted from human faces strongly relies on the robustness of the facial 3D landmark alignment tool. For large poses of human faces, the 68 3D landmarks extracted by a face alignment tool might not be good enough, which makes the extracted facial motion of human faces inaccurate. Then, the subsequent Expansionary Motion Transfer and Texture Animator will also be influenced. We present some failed results in Fig. 2.

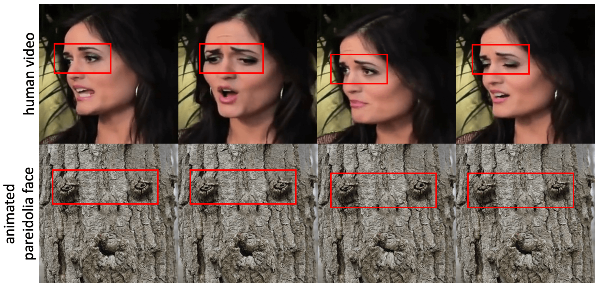


Figure 2: **A failure case caused by large poses of human faces.** Note that the pareidolia face does not well imitate the eyes movement of the human face.

automatic boundary extraction: Currently, our method requires us to label the facial boundaries for each input pareidolia face. Thus, our pareidolia face reenactment is not a fully automatic method and we leave the automatic boundary extraction for pareidolia as future exploration. In addition, as shown in Fig. 3, it is hard to label the facial boundaries for some pareidolia faces such as side faces.

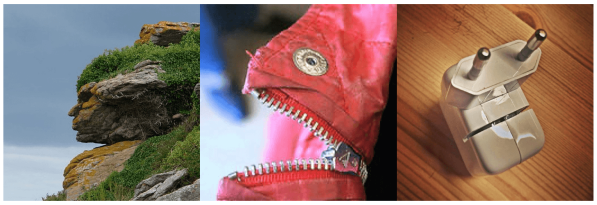


Figure 3: It is hard to label the facial boundaries for these pareidolia faces

failure cases: For pareidolia faces with very complex boundaries and textures of facial parts, our proposed pareidolia face reenactment method might not well. Note that we make the first attempt in animating a pareidolia face by the facial motion of a human face, we present some failure cases in Fig. 4.

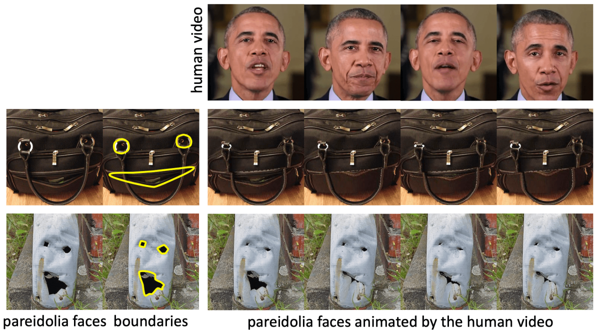


Figure 4: **Failure cases.** In the 2nd row, the texture of the bag's handle becomes incontinuity. In the 3rd row, the global structure of the mouth is broken.

References

[1] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 2