A Generative Model for Depth-based Robust 3D Facial Pose Tracking (Supplementary Material)

Lu Sheng[1] Jianfei Cai[2] Tat-Jen Cham[2] Vladimir Pavlovic[3] King Ngi Ngan[1]
[1]The Chinese University of Hong Kong [2]Nanyang Technological University [3]Rutgers University
{lsheng, knngan}@ee.cuhk.edu.hk, {asjfcai, astjcham}@ntu.edu.sg, vladimir@cs.rutgers.edu

A. Probabilistic Face Parameterization

A.1. Multilinear Representation

In this paper, we collect the $N_{\text{id}} \times N_{\text{exp}} = 150 \times 47$ blendshapes to form the multilinear face representation, as suggested by Cao *et al*. [1]. Provided the shape dimension $N_{\mathcal{M}} \times 3 = 11510 \times 3$, the dataset forms a tensor $\mathcal{D} \in \mathbb{R}^{3N_{\mathcal{M}} \times N_{\text{id}} \times N_{\text{exp}}}$. By subtracting $\mathcal{D}$ with the mean face mesh $\bar{\mathbf{f}}$, the applied data tensor is $\mathcal{T} = \mathcal{D} - \bar{\mathbf{f}}$.

The multilinear face representation requires the unary matrices $\mathbf{U}_{\text{id}}$ and $\mathbf{U}_{\text{exp}}$, which can be derived from High-order SVD (HOSVD) with respect to the second and third dimensions for identity and expression, respectively. The core tensor is thus $\mathcal{C} = \mathcal{T} \times_2 \mathbf{U}_{\text{id}} \times_3 \mathbf{U}_{\text{exp}}$, and inversely the data tensor can be derived as $\mathcal{T} = \mathcal{C} \times_2 \mathbf{U}_{\text{id}}^\top \times_3 \mathbf{U}_{\text{exp}}^\top$. Therefore, one face can be represented by linear combination of the data tensor with respect to the identity $\mathbf{x}_{\text{id}} \in \mathbb{R}^{N_{\text{id}}}$ and expressions $\mathbf{x}_{\text{id}} \in \mathbb{R}^{N_{\text{exp}}}$, together with the mean face model

$$\begin{aligned}\mathbf{f} &= \bar{\mathbf{f}} + \mathcal{T} \times_2 \mathbf{x}_{\text{id}}^\top \times_3 \mathbf{x}_{\text{exp}}^\top \\ &= \bar{\mathbf{f}} + \mathcal{C} \times_2 (\mathbf{U}_{\text{id}}^\top \mathbf{x}_{\text{id}})^\top \times_3 (\mathbf{U}_{\text{exp}}^\top \mathbf{x}_{\text{exp}})^\top \\ &= \bar{\mathbf{f}} + \mathcal{C} \times_2 \mathbf{w}_{\text{id}}^\top \times_3 \mathbf{w}_{\text{exp}}^\top.\end{aligned}$$

$\mathbf{w}_{\text{id}}$ and $\mathbf{w}_{\text{exp}}$ are the parameters employed in this paper.

A.2. Learning the Identity and Expression Priors

Each face mesh in the training dataset is assigned a one-hot label vector for identity $\mathbf{x}_{\text{id}}$ and one for expression $\mathbf{x}_{\text{exp}}$. For example, the face mesh with the neutral expression from the first subject has $\mathbf{x}_{\text{id}} = [1, 0, \ldots, 0]^\top \in \mathbb{R}^{N_{\text{id}}}$ and $\mathbf{x}_{\text{exp}}^\top = [1, 0, \ldots, 0]^\top \in \mathbb{R}^{N_{\text{exp}}}$.

Therefore, the mean face model $\bar{\mathbf{f}}$ has $\bar{\mathbf{x}}_{\text{id}} = \frac{1}{N_{\text{id}}}\mathbf{1}$ and $\bar{\mathbf{x}}_{\text{exp}} = \frac{1}{N_{\text{exp}}}\mathbf{1}$. The identity has $\text{Var}(\mathbf{x}_{\text{id}}) = \frac{1}{N_{\text{id}}}\mathbf{I} - \frac{1}{N_{\text{id}}^2}\mathbb{I} \simeq \frac{1}{N_{\text{id}}}\mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $\mathbb{I}$ is the matrix filled by one. Similarly, $\text{Var}(\mathbf{x}_{\text{exp}}) \simeq \frac{1}{N_{\text{exp}}}\mathbf{I}$. The mean and variance of $\mathbf{w}_{\text{id}}$ and $\mathbf{w}_{\text{exp}}$ are thus

$$\begin{aligned}\boldsymbol{\mu}_{\text{id}} = \frac{1}{N_{\text{id}}}\mathbf{U}_{\text{id}}^\top\mathbf{1},\ \boldsymbol{\Sigma}_{\text{id}} \simeq \frac{1}{N_{\text{id}}}\mathbf{U}_{\text{id}}^\top\mathbf{U}_{\text{id}} = \frac{1}{N_{\text{id}}}\mathbf{I} \\ \boldsymbol{\mu}_{\text{exp}} = \frac{1}{N_{\text{id}}}\mathbf{U}_{\text{exp}}^\top\mathbf{1},\ \boldsymbol{\Sigma}_{\text{exp}} \simeq \frac{1}{N_{\text{id}}}\mathbf{U}_{\text{exp}}^\top\mathbf{U}_{\text{exp}} = \frac{1}{N_{\text{exp}}}\mathbf{I}.\end{aligned}$$

We also observe that the mean face $\bar{\mathbf{f}}$ corresponds to $\bar{\mathbf{x}}_{\text{id}}$ and $\bar{\mathbf{x}}_{\text{exp}}$, thus it results in

$$\begin{aligned}\|\mathcal{T} \times_2 \bar{\mathbf{x}}_{\text{id}}^\top \times_3 \bar{\mathbf{x}}_{\text{exp}}^\top\|_2 &= \|\mathcal{C} \times_2 \boldsymbol{\mu}_{\text{id}}^\top \times_3 \boldsymbol{\mu}_{\text{exp}}^\top\|_2 \\ &\simeq \|\bar{\mathbf{f}} - \bar{\mathbf{f}}\|_2 = 0 \\ \Longrightarrow \boldsymbol{\mu}_{\mathcal{M}} &\simeq \bar{\mathbf{f}},\end{aligned}$$

which means these priors will not bias the face model representations.

We need to mention that $\boldsymbol{\mu}_{\text{id}}$ and $\boldsymbol{\mu}_{\text{exp}}$ should not be zero vectors. If we assume that $\boldsymbol{\mu}_{\text{id}} = \mathbf{0}$, it will result in the independence of the face model with respect to the varying of the expression parameters, *i.e*., $\mathcal{C} \times_2 \mathbf{w}_{\text{id}}^\top \times_3 \mathbf{w}_{\text{exp}}^\top \simeq \mathbf{0}$ in which the $\mathbf{w}_{\text{id}}$ is usually near to the zero vector. The same problem happens when $\boldsymbol{\mu}_{\text{exp}} = \mathbf{0}$. What's worse, such priors lead to instable solutions, *i.e*., $\mathcal{C} \times_3 \mathbf{w}_{\text{exp}}^\top$ suffers such a small magnitude that the online adaptation system (derived from section 4.2) with respect to $\mathbf{w}_{\text{id}}$ will be unstable.

B. Probabilistic Facial Pose Tracking

B.1. Ray Visibility Score

The ray visibility score is the Kullback-Leibler divergence between $p_{\mathcal{P}}(\mathbf{y})$ and $p_{\mathcal{Q}\to\mathcal{P}}(\mathbf{y};\boldsymbol{\theta})$, written as:

$$\begin{aligned}\mathcal{S}(\mathcal{Q},\mathcal{P};\boldsymbol{\theta}) &= D_{KL}[p_{\mathcal{Q}\to\mathcal{P}}(\mathbf{y};\boldsymbol{\theta})||p_{\mathcal{P}}(\mathbf{y})] \\ &= \sum_{n=1}^{N_{\mathcal{M}}} D_{KL}[p_{\mathcal{Q}\to\mathcal{P}}(y_n;\boldsymbol{\theta})||p_{\mathcal{P}}(y_n)],\end{aligned}$$

where $D_{KL}[p_{\mathcal{Q}\to\mathcal{P}}(y_n;\boldsymbol{\theta})||p_{\mathcal{P}}(y_n)]$ is further derived as

$$\frac{\gamma_n}{2\sigma_o^2}\left\{\sigma_o^2+e^{2\alpha}\mathbf{n}_n^\top\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}\mathbf{n}_n+\Delta(\mathbf{T}(\boldsymbol{\theta})\circ\boldsymbol{\mu}_{\mathcal{M},[n]};\mathbf{p}_n)^2\right\}$$
$$+\frac{\gamma_n}{2}\left\{\log 2\pi+\log\sigma_o^2\right\}-(1-\gamma_n)\log U_{\mathcal{O}}$$
$$-\frac{1}{2}\left\{1+\log 2\pi+\log(\sigma_o^2+e^{2\alpha}\mathbf{n}_n^\top\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}\mathbf{n}_n)\right\}.$$

B.2. Optimization of the Ray Visibility Score

We apply the quasi-Newton update $\boldsymbol{\theta}^{(t)}=\boldsymbol{\theta}^{(t-1)}+\Delta\boldsymbol{\theta}$ using the trust region approach for $\mathcal{S}(\mathcal{Q},\mathcal{P};\boldsymbol{\theta}^{(t-1)})$ given the previous $\boldsymbol{\gamma}^{(t-1)}$, in which the key ingredient is the gradient and Hessian calculations with respect to $\boldsymbol{\theta}$. Straightforward calculation is not trivial, especially for rotation vector $\boldsymbol{\omega}$, but $\mathcal{S}(\mathcal{Q},\mathcal{P};\boldsymbol{\theta})$ can be written in an approximated form to reduce the complexity.

Assume we have the mean and variance for the current warped face model point $\bar{\mathbf{q}}_n=\mathbf{T}(\boldsymbol{\theta})\circ\boldsymbol{\mu}_{\mathcal{M},[n]}$ and $\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}$, we are interested in the incremental pose $\Delta\boldsymbol{\theta}$ in the ray visibility score. The incremental rotation matrix is approximated a skew-symmetric representation, *i.e.*, $\mathbf{R}(\Delta\boldsymbol{\omega})=(\mathbf{I}+[\Delta\boldsymbol{\omega}]_\times)$. Therefore, we have

$$\begin{aligned}&\Delta(\mathbf{T}(\boldsymbol{\theta}+\Delta\boldsymbol{\theta})\circ\boldsymbol{\mu}_{\mathcal{M},[n]};\mathbf{p}_n)\\&=\mathbf{n}_n^\top\left\{e^{\Delta\alpha}\mathbf{R}(\Delta\boldsymbol{\omega})\bar{\mathbf{q}}_n+\Delta\mathbf{t}-\mathbf{p}_n\right\}\\&=\mathbf{n}_n^\top\left\{(\mathbf{I}+[\Delta\boldsymbol{\omega}]_\times)e^{\Delta\alpha}\bar{\mathbf{q}}_n+\Delta\mathbf{t}-\mathbf{p}_n\right\}\\&=\mathbf{n}_n^\top\left\{-[e^{\Delta\alpha}\bar{\mathbf{q}}_n]_\times\Delta\boldsymbol{\omega}+\Delta\mathbf{t}+e^{\Delta\alpha}\bar{\mathbf{q}}_n-\mathbf{p}_n\right\},\end{aligned}$$

and the term about the warped variance is

$$\begin{aligned}&e^{2\Delta\alpha}\mathbf{n}_n^\top\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega}+\Delta\boldsymbol{\omega})}\mathbf{n}_n\\&=e^{2\Delta\alpha}\mathbf{n}_n^\top\mathbf{R}(\Delta\boldsymbol{\omega})\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}\mathbf{R}(\Delta\boldsymbol{\omega})^\top\mathbf{n}_n\\&=e^{2\Delta\alpha}\mathbf{n}_n^\top(\mathbf{I}+[\Delta\boldsymbol{\omega}]_\times)\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}(\mathbf{I}+[\Delta\boldsymbol{\omega}]_\times)^\top\mathbf{n}_n\\&=(\mathbf{n}_n+[\mathbf{n}_n]_\times\Delta\boldsymbol{\omega})^\top e^{2\Delta\alpha}\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}(\mathbf{n}_n+[\mathbf{n}_n]_\times\Delta\boldsymbol{\omega}).\end{aligned}$$

$\Delta(\mathbf{T}(\boldsymbol{\theta}+\Delta\boldsymbol{\theta})\circ\boldsymbol{\mu}_{\mathcal{M},[n]};\mathbf{p}_n)^2$ and $e^{2\Delta\alpha}\mathbf{n}_n^\top\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega}+\Delta\boldsymbol{\omega})}\mathbf{n}_n$ are approximated quadratic forms with respect to $\Delta\boldsymbol{\omega}$, thus the gradient and Hessian calculations about the rotation vector become much simpler. The related calculations with respect to the scale factor $\Delta\alpha$ and translation vector $\Delta\mathbf{t}$ involve the previous approximations and thus become simpler as well.

Moreover, to reduce the computational complexity even further, we can force the variance $\boldsymbol{\Sigma}_{\mathcal{M},[n]}=\sigma_{[n]}^2\mathbf{I}$, with $\sigma_{[n]}$ derived from the canonical face point representation. Thus $\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}=\boldsymbol{\Sigma}_{\mathcal{M},[n]}=\sigma_{[n]}^2\mathbf{I}$ and $\mathbf{n}_n^\top\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}\mathbf{n}_n=\sigma_{[n]}^2$ are constant with varying $\boldsymbol{\omega}$. In this case, only $\Delta(\mathbf{T}(\boldsymbol{\theta})\circ\boldsymbol{\mu}_{\mathcal{M},[n]};\mathbf{p}_n)^2$ makes contribution to the update of $\boldsymbol{\omega}$.

B.3. Online Identity Adaptation

The online identity adaptation can be progressively personalized to the test subject, as visualized in Figure 2. With different poses are used to update the identity distribution, the face model is continuously adapted to the test subject.

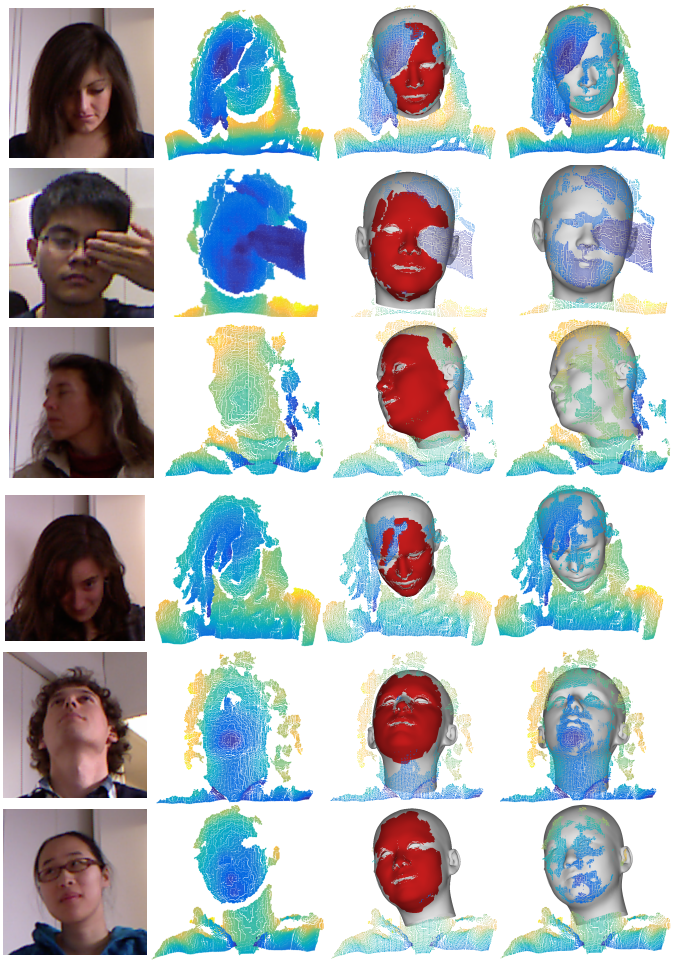

Figure 1. Examples of facial pose results with the visibility detection. The third column shows the visibility masks. The last column shows the personalized face models warped on the point clouds. Best viewed in color.

C. Facial Poses with Visibility Detection

In addition to the facial pose results visualized in the paper, we also illustrate some examples with the personalized face models, as shown in Figure 1. The proposed method can effectively check the visibility of a face model with respect to the input point cloud, and its pose estimation is robust to severe occlusions, *e.g.*, profiled faces, accessories and hands, *etc*.

References

[1] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3D facial expression database for visual computing.

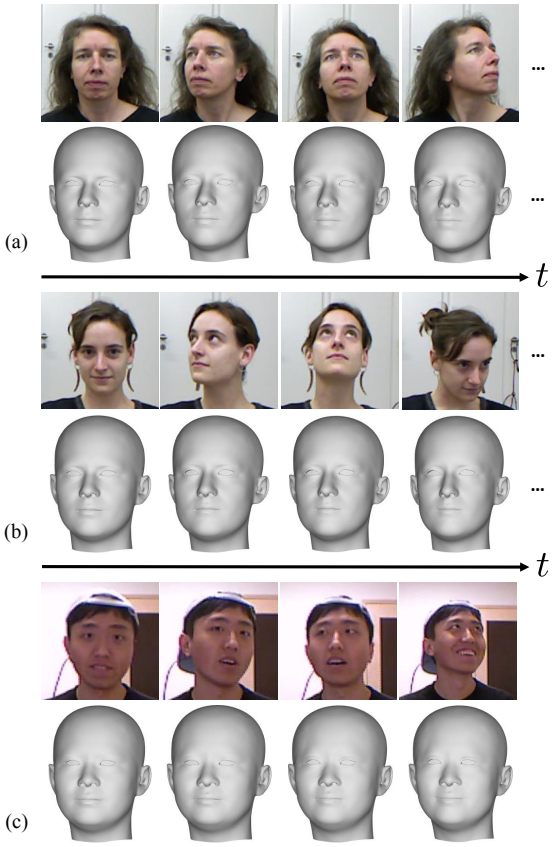

Figure 2. We continuously adapt the identities of the face model to different users. (a)-(c) are three examples showing that the face model can be gradually personalized when the facial depth data from different poses are captured during the tracking process. The face model is initialized with the generic face model.