GLPose: Global-Local Attention Network with Feature Interpolation Regularization for Head Pose Estimation of People Wearing Facial Masks – Supplementary Material

Hsueh-Wei Chen[1]
r09922050@ntu.edu.tw
Yi Chen[1]
d10922004@csie.ntu.edu.tw
Pei-Yung Hsiao[2]
pyhsiao@nuk.edu.tw
Li-Chen Fu[1]
lichen@ntu.edu.tw
ZI-RONG DING[3]
rong@artc.org.tw

[1] Department of Computer Science and Information Engineering
National Taiwan University
Taiwan, ROC

[2] Department of Electrical Engineering
National University of Kaohsiung
Taiwan, ROC

[3] Automotive Research & Testing Center (ARTC)
Taiwan, ROC

This document provides supplementary material for the paper "GLPose: Global-Local Attention Network with Feature Interpolation Regularization for Head Pose Estimation of People Wearing Facial Masks". Section 1 briefly describes the rotation representation used in the proposed method. The generation process of supervised rotation labels used in the feature interpolation regularization module is introduced in Section 2. In Section 3, we show the qualitative results of the proposed method in comparison to other state-of-the-art methods.

1 Rotation representation used in GLPose

Rotation matrix is used for our proposed GLPose. Instead of predicting nine-dimensional representation directly, followed by [4, 6], we utilize the six-dimensional representation as the output of our proposed method. The following steps show how the rotation matrix is transformed from the six-dimensional representation. Suppose the prediction output of the model is $P \in \mathbb{R}^{3\times 2}$, we define the mapping function f_{GS} to transform P to the rotation matrix $M \in SO(3)$[1]:

$$f_{GS}(P) = f_{GS}\left(\begin{bmatrix} | & | \\ p_1 & p_2 \\ | & | \end{bmatrix}\right) = \begin{bmatrix} | & | & | \\ m_1 & m_2 & m_3 \\ | & | & | \end{bmatrix} = M \tag{1}$$

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

[1] SO(3) denotes the special orthogonal group for the three-dimensional rotation group

where $\{p_1, p_2, m_1, m_2, m_3\} \subset \mathbb{R}^3$. The rotation matrix $M \in \mathbb{R}^{3\times3}$ is obtained by the following Gram-Schmidt-like process and the cross product operation:

$$\begin{aligned} m_1 &= \frac{p_1}{||p_1||} \\ m_2 &= \frac{u_2}{||u_2||} \\ m_3 &= m_1 \times m_2 \end{aligned} \tag{2}$$

where $u_2 = p_2 - (m_1 \cdot p_2)\, m_1$. Consequently, an orthogonality constraint is imposed on our model to determine the six-dimensional representation that is translated into a 3×3 rotation matrix in a subsequent transformation.

2 Supervised rotation label used in Feature Interpolation Regularization Module

In the feature interpolation regularization module, we use the interpolation operation to generate intermediate feature embeddings to optimize the feature embedding space. In order to optimize intermediate feature embeddings, the corresponding rotation labels are needed. The generation processes are demonstrated. Suppose the intermediate feature embedding $\hat{f}_3$ is generated by interpolation operation of two feature embeddings $\hat{f}_1$ and $\hat{f}_2$ from the head pose estimation module. The rotation label R_3 for $\hat{f}_3$ is calculated from the two corresponding rotation labels R_1 and R_2 for $\hat{f}_1$ and $\hat{f}_2$. The first step to produce R_3 is to remove the last column vector of R_1 and R_2 by the function g_{GS}:

$$\begin{aligned} g_{GS}(R_1) &= g_{GS}\left(\begin{bmatrix} | & | & | \\ r_1^1 & r_2^1 & r_3^1 \\ | & | & | \end{bmatrix}\right) = \begin{bmatrix} | & | \\ r_1^1 & r_2^1 \\ | & | \end{bmatrix} = R_1^t \\ g_{GS}(R_2) &= g_{GS}\left(\begin{bmatrix} | & | & | \\ r_1^2 & r_2^2 & r_3^2 \\ | & | & | \end{bmatrix}\right) = \begin{bmatrix} | & | \\ r_1^2 & r_2^2 \\ | & | \end{bmatrix} = R_2^t \end{aligned} \tag{3}$$

where $\{r_1^1, r_2^1, r_3^1, r_1^2, r_2^2, r_3^2\} \subset \mathbb{R}^3$ and $\{R_1^t, R_2^t\} \subset \mathbb{R}^{3\times2}$. Afterward, R_1^t and R_2^t are summed to produce the new six-dimensional representation R_3^t:

$$R_3^t = R_1^t + R_2^t \tag{4}$$

Therefore, followed by transformation steps shown in Section 1, the new rotation label $R_3 \in \mathbb{R}^{3\times3}$ is produced. This rotation label R_3 is used in the feature interpolation regularization module to supervise the feature embedding $\hat{f}_3$ by the geodesic distance loss function.

3 Qualitative results

We show the qualitative results on AFLW2000[7], BIWI[1] and MAFA[2] datasets in Figure 1 and 2. In the non-mask datasets like AFLW2000 and BIWI datasets, our proposed method has promising performance compared to HopeNet[4], FSA-Net[5] and 6DRepNet[3]. We

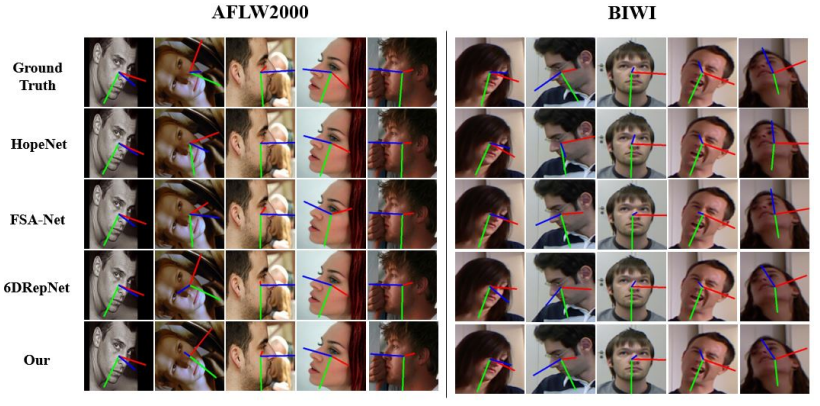

Figure 1: Qualitative visualization on the AFLW2000 and BIWI datasets. From top to bottom, they are ground truth, the predictions of HopeNet, the estimations of FSA-Net, the results of 6DRepNet and our results. The blue, green and red lines indicate the face direction of the front, downward and side respectively.

can observe that those works under some situations like large pose situations will cause performance degradation. On the other hand, the situation under facial masks is our objective. As shown in Figure 2 on the MAFA dataset, we only visualize the predictions because the dataset only provides the classification head pose labels, and our proposed method is more robust than other state-of-the-art methods.

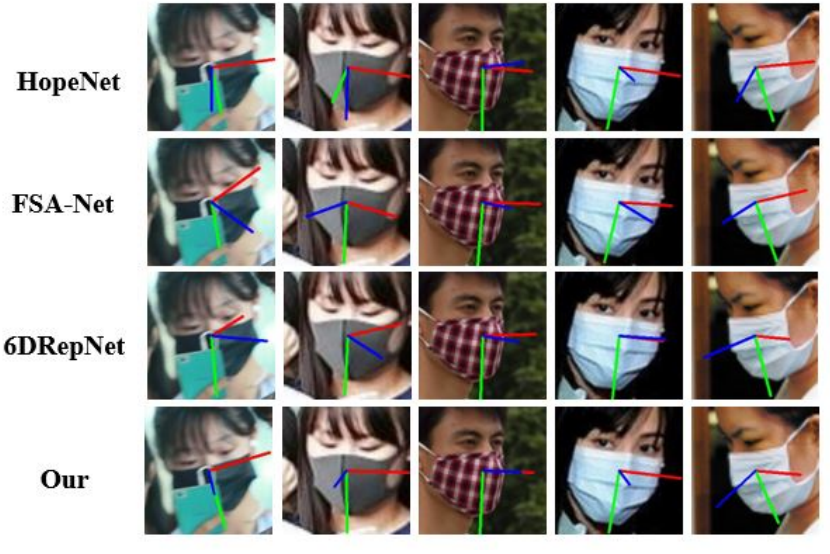

Figure 2: Qualitative results on the MAFA dataset. From top to bottom, they are the predictions of HopeNet, the estimations of FSA-Net, the results of 6DRepNet and our results.

References

[1] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101:437–458, 2012.

[2] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo. Detecting masked faces in the wild with lle-cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2682–2690, 2017.

[3] Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. *ArXiv*, abs/2202.12555, 2022.

[4] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2155–215509, 2018.

[5] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1087–1096, 2019.

[6] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, 2019.

[7] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016.