MHR-Net: Multiple-Hypothesis Reconstruction of Non-Rigid Shapes from 2D Views - *Appendix*

Haitian Zeng[1,2], Xin Yu[1], Jiaxu Miao[3], and Yi Yang[3]

[1] University of Technology Sydney,
[2] Baidu Research,
[3] Zhejiang University,
haitian.zeng@student.uts.edu.au,xin.yu@uts.edu.au
jiaxu.miao@yahoo.com,yangyics@zju.edu.cn

1 Implementation details

Code of MHR-Net will be made available in the future. Here, we give details of network and hyper-parameters. As stated in the main paper, we use the backbone network from [3] as $\mathcal{H}$, which consists of 6 stacked residual blocks with (1024, 256, 1024) channels. The dimension of output feature is 1024.

We replace the invariance loss of PoseDict with the canonicalization loss [3]. A canonicalization network ϕ is adopted, which takes the randomly rotated estimated shape $\mathrm{R}_{\mathrm{rand}}\mathrm{S}_i$ as input and outputs another basis coefficients α_i'. Then α_i' is passed to the basis layer Ψ_B to obtain S_i'. The canonicalization loss forces S_i and S_i' to be the same:

$$\mathcal{L}_{\mathrm{cano}} = \left\|\mathrm{S}_i - \mathrm{S}_i'\right\|_{\mathrm{F}}. \tag{1}$$

The canonicalization network is built using a similar architecture to $\mathcal{H}$ except for the first layer and last layer, which are modified to fit the dimension of inputs and outputs.

For experiments on Human3.6M and 300-VW, we set the number of hypothesis $N_m = 3$, the weighting factors $\lambda_B = 0.8$, $\lambda_{\mathrm{res}} = 1.0$, the shape neighbourhood $\epsilon = 0.075$, and the dimension of basis, deformation and noise $K_b = 10$, $K_d = 64$, $dim_z = 4$. For SURREAL, as the number of points is large, we doubled the channels in $\mathcal{H}$ to (2048, 512, 2048) and set $\lambda_B = 0.95$, while other hyper-parameters remains the same as on Human3.6M and 300-VW.

2 Additional Results & Analysis

2.1 Visualization of Results on 300-VW

We show the reconstruction results of facial landmarks in Fig. 1. The results demonstrate the MHR-Net produces accurate reconstruction of face keypoints.

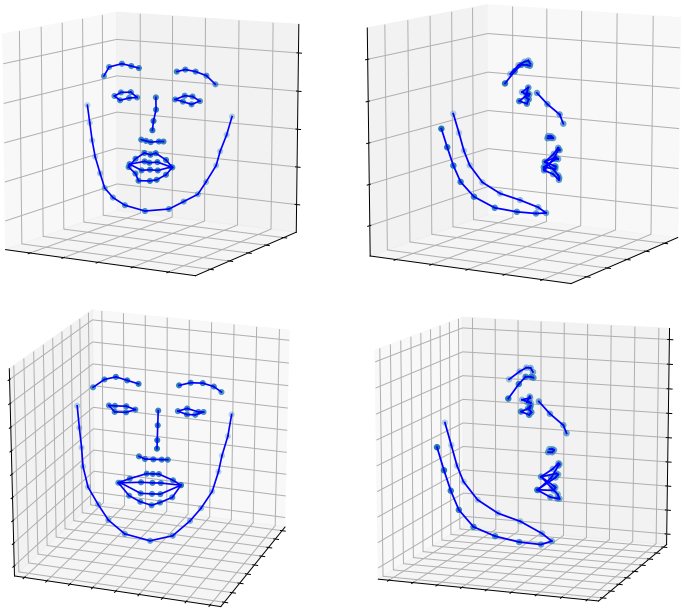

Fig. 1. Visualization of results on 300-VW.

2.2 Degenerated Case of Multiple-Hypothesis Reconstruction

We notice that there are cases where the variation of hypotheses is zero and multiple hypotheses degenerate to a single mode. This degeneration happens under two conditions: (1) the observations are noiseless; (2) N_p is large. (3) Most of points are nearly-rigid. The first condition remove the ambiguity produced by noisy observations. Under the second condition, we are likely to get a deterministic reconstruction because of a large number of observed points. That is, in the classic linear system of NRSfM [1]:

$$\begin{bmatrix} \mathrm{W}_1 \\ \vdots \\ \mathrm{W}_{N_f} \end{bmatrix} = \begin{bmatrix} \mathrm{M}_1 & & \\ & \ddots & \\ & & \mathrm{M}_{N_f} \end{bmatrix} \begin{bmatrix} \mathrm{c}_1^1 & \cdots & \mathrm{c}_{K_b}^1 \\ \vdots & \ddots & \vdots \\ \mathrm{c}_1^{N_f} & \cdots & \mathrm{c}_{K_b}^{N_f} \end{bmatrix} \begin{bmatrix} \mathrm{B}_1 \\ \vdots \\ \mathrm{B}_{K_b} \end{bmatrix}, \tag{2}$$

there is a total of $6N_f + N_f K_b + 3K_b N_p$ of unknowns, and the number of constraints from observations is $2N_f N_p$. Therefore, a large N_p brings extra constraints to the problem, leading to an over-determined system. The third condition is from [2]. When the deformation is simple and most points are rigid, the reconstruction will be unambiguous. On SURREAL dataset with 8690 points, we observe the phenomenon that the variation of deformations quickly goes down to zero in training.

References

1. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision Second Edition. Cambridge University Press (2004)
2. Ijaz, A., Yaser, S., Sohaib, K.: In defense of orthonormality constraints for nonrigid structure from motion. In: CVPR (2009)
3. Novotny, D., Ravi, N., Graham, B., Neverova, N., Vedaldi, A.: C3DPO: canonical 3d pose networks for non-rigid structure from motion. In: ICCV (2019)