Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video

Supplementary Material

Proof of Equation (14)

For simplicity, $\mathcal{L}(\theta;\boldsymbol{W})$ is denoted as

$$\begin{aligned}\mathcal{L}(\theta;\boldsymbol{W}) &= \frac{\nu}{2}\sum_{t=1}^{n}\left\|\boldsymbol{W}_t-\boldsymbol{R}_t\sum_{i=1}^{k}c_{it}\boldsymbol{B}_i-\boldsymbol{T}_t\boldsymbol{1}^T\right\|_F^2\\ &= \frac{\nu}{2}\left\|\boldsymbol{W}-\boldsymbol{Z}(\theta)\right\|_F^2, \end{aligned}\tag{1}$$

where $\boldsymbol{W}$ is the stack of all $\boldsymbol{W}_t$ and $\boldsymbol{Z}(\theta)$ is the stack of all $\boldsymbol{R}_t\sum_{i=1}^{k}c_{it}\boldsymbol{B}_i-\boldsymbol{T}_t\boldsymbol{1}^T$. Then

$$\begin{aligned}\int\mathcal{L}(\theta;\boldsymbol{W})\mathrm{Pr}(\boldsymbol{W}|\boldsymbol{I},\theta')d\boldsymbol{W} &= \frac{\nu}{2}\int\left\|\boldsymbol{W}-\boldsymbol{Z}(\theta)\right\|_F^2\ \mathrm{Pr}(\boldsymbol{W}|\boldsymbol{I},\theta')d\boldsymbol{W}\\ &= \frac{\nu}{2}\int\left\{\langle\boldsymbol{W},\boldsymbol{W}\rangle-\langle\boldsymbol{W},\boldsymbol{Z}(\theta)\rangle+\langle\boldsymbol{Z}(\theta),\boldsymbol{Z}(\theta)\rangle\right\}\ \mathrm{Pr}(\boldsymbol{W}|\boldsymbol{I},\theta')d\boldsymbol{W}\\ &= \frac{\nu}{2}\ \left\{\mathrm{const}-\int\langle\boldsymbol{W},\boldsymbol{Z}(\theta)\rangle\,\mathrm{Pr}(\boldsymbol{W}|\boldsymbol{I},\theta')d\boldsymbol{W}+\langle\boldsymbol{Z}(\theta),\boldsymbol{Z}(\theta)\rangle\right\}\\ &= \frac{\nu}{2}\ \left\{\mathrm{const}-\left\langle\int\boldsymbol{W}\mathrm{Pr}(\boldsymbol{W}|\boldsymbol{I},\theta')d\boldsymbol{W}\ ,\ \boldsymbol{Z}(\theta)\right\rangle+\langle\boldsymbol{Z}(\theta),\boldsymbol{Z}(\theta)\rangle\right\}\\ &= \frac{\nu}{2}\left\|\int\boldsymbol{W}\mathrm{Pr}(\boldsymbol{W}|\boldsymbol{I},\theta')d\boldsymbol{W}-\boldsymbol{Z}(\theta)\right\|_F^2+\mathrm{const}\\ &= \frac{\nu}{2}\left\|\mathrm{E}\left[\boldsymbol{W}|\boldsymbol{I},\theta'\right]-\boldsymbol{Z}(\theta)\right\|_F^2+\mathrm{const}\end{aligned}\tag{2}$$

Derivation of Equation (15)

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{W}|\boldsymbol{I},\theta'\right] &= \int\mathrm{Pr}(\boldsymbol{W}|\boldsymbol{I},\theta')\ \boldsymbol{W}\ d\boldsymbol{W}\\ &= \int\frac{\mathrm{Pr}(\boldsymbol{W},\boldsymbol{I},\theta')}{\mathrm{Pr}(\boldsymbol{I},\theta')}\ \boldsymbol{W}\ d\boldsymbol{W}\\ &= \int\frac{\mathrm{Pr}(\boldsymbol{I}|\boldsymbol{W})\mathrm{Pr}(\boldsymbol{W}|\theta')\mathrm{Pr}(\theta')}{\mathrm{Pr}(\boldsymbol{I}|\theta')\mathrm{Pr}(\theta')}\ \boldsymbol{W}\ d\boldsymbol{W}\\ &= \int\frac{\mathrm{Pr}(\boldsymbol{I}|\boldsymbol{W})\mathrm{Pr}(\boldsymbol{W}|\theta')}{Z}\ \boldsymbol{W}\ d\boldsymbol{W}\end{aligned}\tag{3}$$

Evaluation on HumanEva dataset

The evaluation results on the HumanEva I dataset [1] are presented. The evaluation protocol described in [2] was adopted. The walking and jogging sequences from camera C1 of all subjects were used for evaluation. The CNN joint detectors trained on the Human3.6M dataset were fine-tuned with the training sequences for each action separately. Each estimated 3D pose was aligned to the ground truth with the procrustes method. The mean 3D joint errors for the evaluation sequences were reported in Table 1.

	Walking			Jogging		
	S1	S2	S3	S1	S2	S3
Proposed	34.2	30.9	49.1	47.6	33.0	29.7
Simo-Serra et al. [2]	65.1	48.6	73.5	74.2	46.6	32.2

Table 1: Quantitative results on the HumanEva I dataset [1]. The numbers are the mean per joint errors in millimeters.

References

[1] L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4–27, 2010.

[2] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A Joint Model for 2D and 3D Pose Estimation from a Single Image. In *CVPR*, 2013.