Dabral, Mundhada, Kusupati, Afaque, Sharma, Jain
Learning 3D Human Pose from Structure and Motion

1 Summary

An extension of Zhou et al., *Towards 3D Human Pose Estimation in the Wild: a Weakly-Supervised Approach* (2017; the geometric loss paper). 3D pose estimation takes place over three stages:

- **SAP-Net** (Structure-Aware PoseNet): utilize the [stacked hourglass 2D] → [depth regression 3D] of Zhou et al., in addition to two new structural loss functions (illegal angle loss, symmetry loss) to predict 3D pose for a single RGB image
- **TP-Net**: (Temporal PoseNet): take in the current predicted 3D pose and the previous $(n-1)$ predicted 3D poses and output a temporally refined version of the current predicted 3D pose
- **Skeleton fitting**: if a specific skeleton is a available, fit the 3D pose to that skeleton (just map the predicted joint angles/bone directions to it)

The contributions of the paper are (1) the additional structural loss functions and (2) the TP-Net architecture (learning from structure, learning from motion).

2 The Title

- **from Structure:** SAP-Net, which includes these additional structure-based losses
- **from Motion:** TP-Net, which refines a pose based on a window of poses up to the current pose

3 Method

3.1 SAP-Net

Takes a single frame, outputs a single 3D pose.

- Built upon framework of Zhou et al., where stacked hourglass network is supervised to predict x, y and depth regression module (four residual/pooling modules plus fully connected layer) takes stacked hourglass feature maps and is (sometimes weakly) supervised to predict z.
  - Uses ground truth z for supervision when available, otherwise geometric loss.
- The SAP-Net improves upon the geometric loss, which is a comparatively paltry supervision. Namely, it adds two anatomical losses of its own:
  - **Illegal angle loss:** impose increasing loss when knee/angle joints are increasingly bent past 180°. For right elbow joint, e.g., define normal as collar bone × upper arm bone. Then dot product of lower arm bone $\mathbf{l}$ with normal $\mathbf{n}$ should be positive. (Define $E_e^r = \min(\mathbf{n} \cdot \mathbf{l}, 0)$, and right elbow loss as $-E_e^r e^{-E_e^r}$.) Exponentiate to penalize large deviations from legality. Note that left elbow/right knee are reversed (want opposite direction as normal), as shown by right-hand rule.

$$\mathcal{L}_a = -E_e^r e^{-E_e^r} + E_e^l e^{E_e^l} + E_k^r e^{E_k^r} - E_k^l e^{-E_k^l}$$

  - **Symmetry loss:** difference between lengths of corresponding left/right bones.

$\mathcal{L}_s$ = sum of L2 distances between *length of left bone* and *length of corresponding right bone*

Overall, the *weak supervision* loss is

$$\lambda_a \mathcal{L}_a(\tilde{P}^z, \hat{P}^{xy}) + \lambda_s \mathcal{L}_s(\tilde{P}^z, \hat{P}^{xy}) + \lambda_g \mathcal{L}_g(\tilde{P}^z, \hat{P}^{xy})$$

for λ the loss weights, $\mathcal{L}_a$ the illegal angle loss, $\mathcal{L}_s$ the symmetry loss, and $\mathcal{L}_g$ the geometric loss. $\tilde{P}^z$ is predicted depth, $\hat{P}^{xy}$ is ground truth x, y. (Use $\hat{P}^{xy}$ with weak supervision to simplify training.)

3.2 TP-Net

Takes a sequence of 3D poses for contiguous frames $\{..., \tilde{P}_{t-1}, \tilde{P}_t\}$, outputs temporally refined 3D pose $\tilde{P}_t$.

- Very simple architecture: "two layers, 4096 hidden neurons, fully connected with ReLUs."
- As input, the 3D poses are just flattened and concatenated.
- Trained with L_2 loss from ground truth current pose.

3.3 Training

Train in four stages:

1. Train 2D stacked hourglass network on MPII and H36M.
2. Train 3D depth module using only data with 3D annotations.
3. Train full SAP-Net with geometric and illegal angle losses ($\lambda_a = 0.03, \lambda_g = 0.03$).
4. Train full SAP-Net with all losses (add symmetry loss; $\lambda_a = 0.03, \lambda_g = 0.03, \lambda_s = 0.05$).

3.4 Analysis

The paper has some wonderful analysis visualizations (indeed, this paper was overall a very nice read).

For example, it shows loss surfaces for varying $x_{\text{left elbow}}, z_{\text{left elbow}}$ and fixed everything else, with just 2D projection loss, then + symmetry loss, and then + illegal angle loss (also with just full 3D L_2 loss).

- It shows a clear benefit at least in the given scenarios for adding the losses; shows where a good pose, a bad pose, and a worse pose appear on the loss surface.
- By the time the illegal angle loss is added, the good pose is clearly at a minimum region, while the worst pose is at the comparatively highest region.

For the TP-Net, they also identify through sensitivity analysis that the final predicted pose is not very sensitive to poses more than five time steps earlier.

- They use this to justify TP-Net's superior performance to RNNs, arguing that an extended context is unnecessary and difficult to utilize; a simple dense network with a *limited* context is more appropriate.

References

[1] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3D Human Pose from Structure and Motion. *arXiv preprint arXiv:1711.09250 (2018).*