Mehta, Sridhar, Sotnychenko, Rhodin, Shafiei, Seidel, Xu, Casas, Theobalt

VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera

1 Summary

VNect regresses a single 3D human pose for the t^{th} frame of a sequence of RGB images $\{..., \mathbf{I}_{t-1}, \mathbf{I}_t\}$ by

1. predicting a 2D heatmap and root-relative X, Y, Z location maps for each joint via a fully conv. network
2. using positions of heatmap maximums as the 2D predictions
3. temporally filtering these 2D predictions using the one Euro filter
4. using the 2D predictions as indices into the X, Y, Z location maps to obtain root-relative 3D predictions
5. temporally filtering these 3D predictions using the one Euro filter
6. retargeting 3D bone lengths, preserving joint angles
7. running post-optimization (minimize energy function) to obtain **full** 3D pose in camera frame
8. temporally filtering the 3D pose, again with the one Euro filter

The result is a temporally coherent 3D pose sequence for the given input frames.

2 Issues with Previous Approaches to 3D Pose Estimation

- reconstruct 3D pose individually per image (temporally jittery)
- do not enforce constant bone lengths
- estimate local 3D pose relative to bounding box, not full global pose
- predict a vector of 3D locations directly, creating a propensity toward fully connected formulations
  - and thereby restricting inputs to tight crops and/or fixed resolutions

3 Post-Optimization

Energy function:

$$E_{\text{total}}(\theta, \mathbf{d}) = E_{\text{IK}}(\theta, \mathbf{d}) + E_{\text{proj}}(\theta, \mathbf{d}) + E_{\text{smooth}}(\theta, \mathbf{d}) + E_{\text{depth}}(\theta, \mathbf{d})$$

for

- $E_{\text{IK}}(\theta, \mathbf{d})$ the 3D inverse kinematics term: L2 distance between $\theta, \mathbf{d}$ and 3D root-relative output
- $E_{\text{proj}}(\theta, \mathbf{d})$ the projection term: L2 distance between reprojected 2D and detected 2D keypoints
- $E_{\text{smooth}}(\theta, \mathbf{d})$ the smoothness term: acceleration of pose across frames
- $E_{\text{depth}}(\theta, \mathbf{d})$ the depth term: variation in depth (z-component) across frames

The final pose is parameterized by the joint angles θ and the root position $\mathbf{d}$ in camera coordinates.

4 Auxiliary Prediction

In the middle of the network, we predict/compute (1) *parent-relative* location maps for each joint and (2) bone length maps, concatenating these with the intermediate features to help guide the network.

5 Other Details

- Each location map stores the root-relative x, y, or z-coordinate of the thing at each pixel.
- The CNN predicts *height-normalized* 3D positions (due to the scale ambiguity of monocular estimation).
- The network is ResNet50 through level 4, plus a custom output module.

References

[1] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, Christian Theobalt. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *arXiv preprint arXiv:1705.01583 (2017).*