UNIVERSITY OF CAMBRIDGE

Hierarchical Kinematic Probability Distributions for 3D Human Shape and Pose Estimation from Images in the Wild

Akash Sengupta Ignas Budvytis Roberto Cipolla

2021 ICCV OCTOBER 11-17 VIRTUAL

Problem

Monocular 3D human shape and pose estimation is **ill-posed.**

- Multiple 3D bodies may explain a given 2D image.

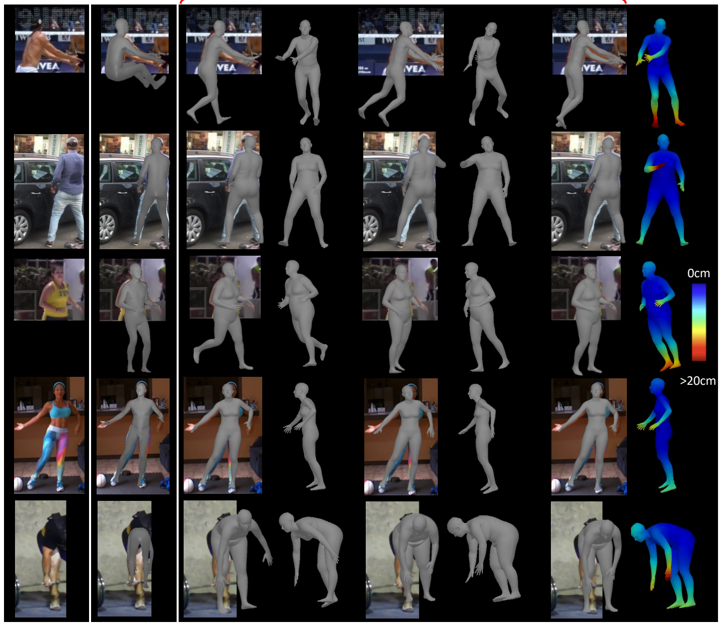


Image SPIN[1] **Ours: Multiple** 3D reconstruction samples Per-Vertex **Uncertainty**

Our approach: **predict probability distributions** over body shape and pose.

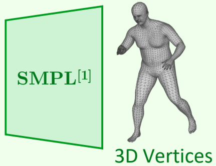
- Sample many plausible 3D reconstructions and estimate 3D uncertainty.
- Leverage the human body's hierarchical kinematic tree structure.
- Recognise that 3D joint rotations (i.e. pose) lie in *SO(3)* → non-linear manifold.

References

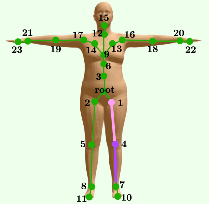
1. N. Kolotouros *et al.* Learning to reconstruct 3D human pose and shape via model-fitting in the loop. ICCV 2019.
2. M. Loper *et al.* SMPL: A Skinned Multi-Person Linear model. ACM SIGGRAPH Asia 2015.
3. T. D. Downs. Orientation statistics. Biometrika 59(3):665–676, December 1972.
4. D. Mohlin *et al.* Probabilistic orientation estimation with matrix fisher distributions. NeurIPS 2020.
5. B. Biggs *et al.* 3D Multibodies: Fitting sets of plausible 3D models to ambiguous image data. NeurIPS 2020.
6. A. Sengupta *et al.* Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. CVPR 2021.
7. N. Kolotouros *et al.* Probabilistic modeling for human mesh recovery. ICCV 2021.

Method

Preliminary: SMPL[2]



- Body model mapping shape and pose parameters to 3D mesh.
- Pose is parameterised by the 3D rotation ($\boldsymbol{R}_i$) of each joint relative to its parent in the kinematic tree.



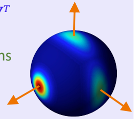
Preliminary: Matrix-Fisher Distribution[3]

- 3D joint rotations $\boldsymbol{R}_i$ lie in *SO(3)*.
- Matrix-Fisher distribution over *SO(3)*:

$$p(\boldsymbol{R}|\boldsymbol{F}) = \frac{1}{c(\boldsymbol{F})}\exp\left(\mathrm{tr}\left(\boldsymbol{F}^T\boldsymbol{R}\right)\right) = M(\boldsymbol{R};\boldsymbol{F})$$

- Parameter $\boldsymbol{F} \in \mathbb{R}^{3\times3}$ can be regressed by a neural network[4].

- Proper SVD: $\boldsymbol{F} = \boldsymbol{USV}^T$
- Mode: $\boldsymbol{UV}^T$
- Principal Axes: columns of $\boldsymbol{U}$
- Dispersion: singular values $\boldsymbol{S}$



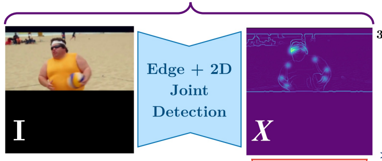
$\boldsymbol{F} = \mathrm{diag}[25, 5, 1]$

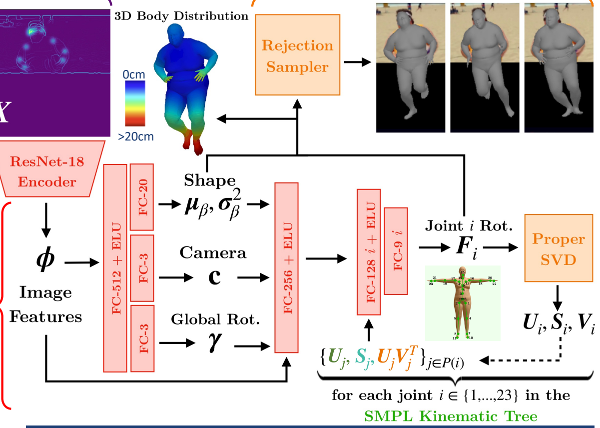
(i) Proxy representation computation

- Input image is converted into an edge + 2D joint heatmap representation.
- Bridges the gap between synthetic training data (with diverse shapes and poses) and real test data.

(iii) 3D Body Sampling and Projection

- SMPL meshes are sampled from from the predicted distribution using rejection sampling.
- Rejection sampling is made differentiable using the re-parameterisation trick → enables sample re-projection loss.

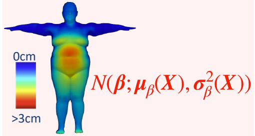




(ii) 3D Shape and Pose Distribution Prediction

Deep neural network outputs:

(1) Gaussian distribution over shape parameters



(2) Matrix-Fisher distribution over 3D joint rotations

$M(\boldsymbol{R}_i; \boldsymbol{F}_i)$ for $i \in \{1,\ldots,23\}$

Each joint's distribution parameter ($\boldsymbol{F}_i$) is conditioned on its parents' modes, principal axes and dispersions

Results

Method	3DPW								SSP-3D			
	MPJPE (mm)				MPJPE-PA (mm)				PVE-T-SC (mm)			
Number of Samples:	1	5	10	25	1	5	10	25	1	5	10	25
Biggs *et al.* [5]	93.8	82.2	79.4	75.8	59.9	57.1	56.6	55.6	-	-	-	-
Sengupta *et al.* [6]	97.1	95.8	93.1	89.7	61.1	59.4	58.2	56.5	15.2	14.8	13.6	11.9
ProHMR [7]	-	-	-	-	59.8	56.5	54.6	52.4	-	-	-	-
Ours (Independent)	88.3	85.0	82.6	78.5	56.6	54.5	52.8	50.2	13.9	12.9	12.0	10.3
Ours (Hierarchical)	**84.9**	**81.6**	**79.0**	**75.1**	**53.6**	**51.4**	**49.6**	**47.0**	**13.6**	**12.3**	**11.3**	**9.8**

Comparison with recent 3D human shape and pose distribution estimation methods

- 3D shape (PVE-T-SC) and pose (MPJPE/MPJPE-PA) metrics are computed using the minimum error sample[5] for each test image in the 3DPW and SSP-3D datasets.
  - Motivation: ground-truth 3D body only represents *one* plausible 3D solution out of many.
- 3D metrics improve with increasing number of samples → our predicted distribution is able to model the 3D ground-truth as a possible sample.