HoloPose: Real Time Holistic 3D Human Reconstruction In-The-Wild

Extended Abstract for Third Workshop on Computer Vision for AR/VR, CVPR 2019

Rıza Alp Güler George Papandreou
Dan Stoddart Stefanos Zafeiriou Iasonas Kokkinos

Ariel AI

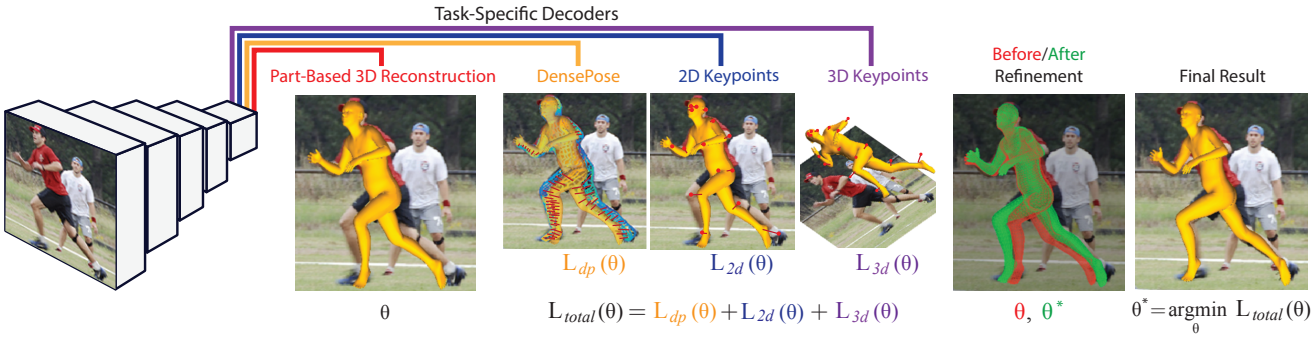

Figure 1: We introduce HoloPose, a method for holistic monocular 3D body reconstruction in-the-wild. We start with an accurate, part-based estimate of 3D model parameters θ, and decoupled, FCN-based estimates of DensePose, 2D and 3D joints. We then efficiently optimize a misalignment loss $L_{total}(\theta)$ between the top-down 3D model predictions to the bottom-up pose estimates, thereby largely improving alignment. The 3D model estimation and iterative fitting steps are efficiently implemented as network layers, facilitating multi-person 3D pose estimation in-the-wild at more than 10 frames per second.

1. Introduction

We introduce HoloPose, a method for holistic monocular 3D human body reconstruction. We introduce a part-based model for 3D model parameter regression that allows our method to operate in-the-wild, handling severe occlusions and large pose variation. We further train a multi-task network comprising 2D, 3D and Dense Pose estimation to drive the 3D reconstruction task. For this we introduce an iterative refinement method that aligns the model-based 3D estimates of 2D/3D joint positions and DensePose with their image-based counterparts delivered by CNNs, achieving both model-based, global consistency and high spatial accuracy thanks to the bottom-up CNN processing. We validate our contributions on challenging benchmarks, showing that our method allows us to get both accurate joint and 3D surface estimates, while operating at more than 10fps in-the-wild.

This work is originally published in CVPR 2019 and will also be presented as a real-time live demonstration.

2. Methods

2.1. Shape Prior for 3D Human Reconstruction

Mixture-of-Experts Rotation Prior: We parameterize the human body using the Skinned Multi-Person Linear (SMPL) model [7]. Apart from defining a prior on the shape given the model parameters, we propose here to enforce a prior on the model parameters themselves.

We argue that a simple and tight prior can be constructed by explicitly forcing the prediction to lie on the manifold of plausible shapes. We propose a simple 'mixture-of-experts' angle regression layer that has a simple and effective prior on angles baked into its expression. We start by using the data collected by [1] of joint angle recordings as humans stretch. These are expected to cover sufficiently well the space of possible joint angles. For each body joint we represent rotations as Euler angles, θ and compute K rotation clusters $\theta_1, \ldots, \theta_K$ via K-Means. These clusters provide us with a set of representative angle values. We allow our system to predict any rotation value within the convex hull of these clusters by using a softmax-weighted combination of the clusters. In particular, the Euler rotation Θ^i for the i'th body joint is computed as:

$$\theta^i = \frac{\sum_{k=1}^{K} \exp(w_k)\theta_K}{\sum_{k=1}^{K} \exp(w_k)} \quad (1)$$

where w_k are real-valued inputs to this layer.

Cartesian surface parametrization We have found it advantageous to reparametrize the body surface with a locally cartesian coordinate system. This allows us to replace this

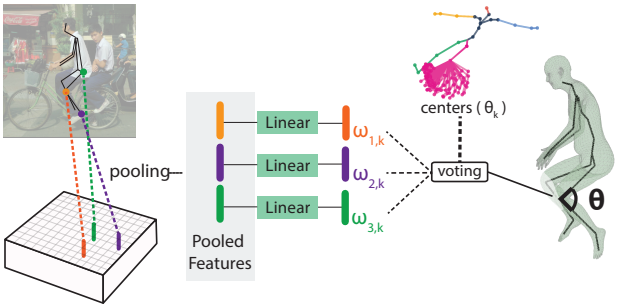

Figure 2: **Part-Based 3D Reconstruction.** We pool convolutional features around each keypoint, deriving a representation of local image structure that is invariant to global image deformations. Each keypoint affects a subset of kinematically associated body model parameters, casting its own 'vote' for the putative joint angles. These votes are fused through a mixture-of-experts architecture that delivers a part-based estimate of body joint angles.

tedious process with bilinear interpolation and use a Spatial Transformer Layer [4] to efficiently handle large numbers of points. In particular, we use a 32×32 grid within each of the 24 body parts used in [2] which means that rather than the 6890 3D vertices of SMPL we now have 24 tensors of size $32 \times 32 \times 3$. We also sample the model eigenshapes on the same grid and express the shape synthesis equations in terms of the resulting tensors. We further identify UV-part combinations that do not correspond to any mesh vertex and ignore UV points that map there.

2.2. Part-Based 3D Body Reconstruction

We extract localized features around human joints. The position where we extract features co-varies with joint position. The features are therefore invariant to translation by design and can focus on local patterns that better reveal the underlying 3D geometry. As shown in Fig. 2, we obtain features as a result of a deconvolution network and pool features at visible joint locations via bilinear interpolation.

We use a part-based variant of Eq. 1, where we pool information from $\mathcal{N}(i)$, the neighborhood of joint i corresponding to the angle θ^i:

$$\theta^i = \frac{\sum_{k=1}^{K} \sum_{j \in \mathcal{N}(i)} \exp(w_{k,j}^i)\theta_K}{\sum_{k=1}^{K} \sum_{j \in \mathcal{N}(i)} \exp(w_{k,j}^i)}. \quad (2)$$

As in Eq. 1 we perform an arg-soft-max operation over angle clusters, but fuse information from multiple 2D joints: $w_{k,j}^i$ indicates the score that 2D joint j assigns to cluster k for the i-th model parameter, θ^i. The neighborhood of i is constructed offline, by inspecting which model parameters directly influence human 2D joints, based on kinematic tree dependencies. Joints are found in the image by taking the maximum of a 2D joint detection module.

2.3. Holistic 3D Body Reconstruction

The network described so far delivers a 'bottom-up' estimate of the body's 3D surface in a single-shot, i.e. through a forward pass in the network. In the same feedforward manner we obtain 2D keypoints, 3D joint [8], or DensePose [2] estimates through fully-convolutional networks (FCNs). These provide complementary pieces of information about the human pose in the scene, with complementary merits. We introduce a refinement process that forces the model-based 3D geometry to agree with an FCN's predictions through an iterative scheme. This is effective also at test-time, where the FCN-based pose estimates drive the alignment of the model-based predictions to the image evidence through a minimization procedure.

We construct losses that penalize deviations between the 3D model-based predictions and the pose information provided by complementary cues. For example, Dense Pose associates an image position $\mathbf{x} = (x_1, x_2)$ with an intrinsic surface coordinate, $\mathbf{u} = (u_1, u_2)$. Given a set of model parameters $\phi = (\theta, \beta)$ we can associate every $\mathbf{u}$ vector with a 3D position $\mathbf{X}(\phi) = M(\phi, \mathbf{u})$, where M denotes the parametric model for the 3D body shape, e.g. [7]. This point in turn projects to a 2D position $\hat{\mathbf{x}}(\phi) = (\hat{x}_1, \hat{x}_2)$, which can be compared to $\mathbf{x}$ - ideally closing a cycle. Since this will not be the case in general, we penalize a geometric distance between $\hat{\mathbf{x}}(\phi)$ and $\mathbf{x} = (x_1, x_2)$, requiring that (ϕ) yields a shape that projects correctly in 2D. Summarizing, we have the following process and loss:

$$\mathbf{x} \overset{\text{DensePose}}{\rightarrow} \mathbf{u} \overset{M(\phi)}{\rightarrow} \mathbf{X} \overset{\Pi}{\rightarrow} \hat{\mathbf{x}} \quad (3)$$

$$\mathcal{L}_{\text{DensePose}}(\phi) = \sum_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2, \quad (4)$$

where $\hat{\mathbf{x}} = \Pi(M_\phi(\text{DensePose}(\mathbf{x})))$ is the model-based estimate of where $\mathbf{x}$ should be, Π is an orthographic projection matrix and i ranges over the image positions that become associated with a surface coordinate.

We can use Eq. 4 to supervise network training, where DensePose stands for Dense Pose ground-truth and ϕ is obtained by the part voting expression in Eq. 2. Secondly, we can use Eq. 4 at test time to force the coupling of the FCN- and model- based estimates of human pose. We bring them in accord by forcing the model-based estimate of 3D structure to project correctly to the FCN-based DensePose/2D/3D joint predictions. For this we treat the CNN-based prediction as an initialization of an iterative fitting scheme driven by the sum of the geometric losses. Furthermore, to cope with implausible shapes we use the following simple loss to bound the magnitude of the predicted β values: $\mathcal{L}_{beta} = \sum_i \max(0, b - |\beta_i|)$, where $b = 2$ is used in all experiments. We use Conjugate Gradients (CG) to minimize a cost function formed by the sum of the above losses.

(a) (b) (c) (d) (e) (f)

Figure 3: **Qualitative results.** From left to right: (a) Input image, (b) HMR [6] results, (c) Our results, without refinement, (d) the visualization of the refinement, (e) our results, refined, (f) our results, refined, 3D rotated.

3. Experiments

We quantify performance in terms of two complementary problem aspects, namely mesh-based dense pose estimation and 3D object reconstruction. Our qualitative results demonstrate the performance of our system on challenging, "in-the-wild" images with heavy occlusion and clutter.

3.1. Surface Correspondence Performance

Dense correspondence measures 3D mesh alignment accuracy in challenging, "in the wild" scenarios. It complements 3D localization performance, because 3D localization can only be evaluated in constrained images where 3D pose information has been captured by appropriate setups, while dense correspondence can be established by manual annotators as in [2]. We measure the surface correspondence accuracy using the 'Ratio of Correct Points' (RCP) measure. The results show that the alignment accuracy of the proposed approach is clearly superior to the state-of-the-art approach of [6]. Also, the synergistic refinement is successfully using the information provided by the DensePose head to improve the alignment accuracy, but there is still space for improvement, e.g. by using a more expressive 3D shape model.

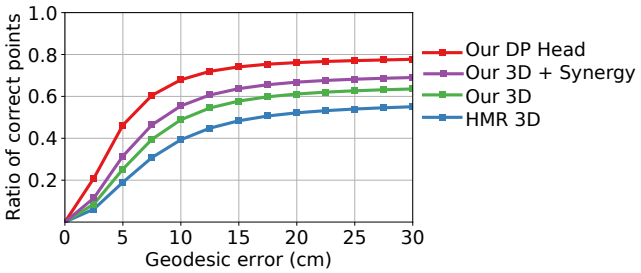

Figure 4: **Surface Correspondence Results:** Ratio of Correct points as a function of the the geodesic distance threshold. The proposed system uniformly outperforms the current state-of-the-art mesh-based results, while the refinement yields a further boost in surface alignment.

Method	*PA MPJPE*	*MPJPE*
HMR	56.8	87.97
Ours	**50.56**	**64.28**
Ours+ Synergy	**46.52**	**60.27**

Table 1: **Results on Human3.6M Dataset.** MPJPE in mm. PA MPJPE means the estimated keypoints were rigidly aligned to ground truth prior to evaluation.

3.2. 3D Keypoint Localization

We report the results of our system on the Human3.6M [3] benchmark. There are two commonly used evaluation protocols with different partitions of the dataset and different evaluation metrics; namely with and without rigid alignment (PA), both reported in Table 1.

For the sake of brevity, we only compare to HMR, the state-of-the-art system by Kanazawa *et al*. [5]. Our system improves over HMR by 5.2 mm (PA MPJPE) in *Protocol 1* and 23.6 mm in *Protocol 2* (MPJPE). Furthermore, we show that the synergistic refinement leads to a further improvement of 4mm in both protocols.

3.3. Qualitative Results

Qualitative results of our system are provided in Fig. 3. We observe that HMR is often distracted by clutter while delivering a pose estimate, whereas our part-based estimate is visibly more accurate; the refinement step further aligns the surface with the image, correcting in particular limb estimates.

References

[1] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[2] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 3

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*. 2

[5] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[6] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, pages 386–402, 2018. 3

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 2

[8] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral human pose regression. *arXiv preprint arXiv:1711.08229*, 2017. 2