Learning Body Shape and Pose from Dense Correspondences

Yusuke Yoshiyasu[1,2] and Lucas Gamez[1]

[1]CNRS-AIST JRL (Joint Robotics Laboratory) UMI3218/RL
[2] National Institute of Advanced Industrial Science and Technology (AIST), Japan

Abstract

In this paper, we address the problem of learning 3D human pose and body shape from 2D image dataset, without having to use 3D supervisions (body shape and pose) which are in practice difficult to obtain. The idea is to use dense correspondences between image points and a body surface, which can be annotated on in-the-wild 2D images, to extract, aggregate and learn 3D information such as body shape and pose from them. To do so, we propose a training strategy called "deform-and-learn" where we alternate deformable surface registration and training of deep convolutional neural networks (ConvNets). Experimental results showed that our method is comparable to previous semi-supervised techniques that use 3D supervision.

1. Introduction

With the progress of deep learning, estimating 3D human body shape and pose from a single image is now possible by regressing the parameters of statistical body models. The main challenge in this task is the lack of a large-scale 3D dataset that contains a wide variety of people and background. Some dataset is captured in an experimental room using a Motion Capture (MoCap) system and RGB video cameras, which provides pairs of image and 3D pose, but are limited to a small number of subjects. Extending these 3D dataset to in-the-wild settings with a wide variety of body types seems not straightforward.

"Can we learn 3D human body shape and pose directly from 2D images?" In this paper, we tackle this challenging problem to bypass the 3D dataset scarcity problem by extracting and aggregating 3D information from dense correspondences annotated on 2D images. We propose a strategy called "deform-and-learn" where we alternate deformable surface registration, which fits a 3D model to 2D images, and training of deep neural network, which predicts 3D body shape/pose from a single image. These processes are iterated to improve accuracy. Experiments showed that our method is comparable to semi-supervised techniques that use 3D supervision.

2. Related Work

A common way to predict 3D human body shape and pose from an image is to employ pre-built statistical human models. The first method to do this using deep ConvNets was proposed in SMPLify [BKL*16] where the human statistical model called SMPL was fitted to the 2D key points estimated from an image using ConvNets by an optimization technique. Lassner et al. [LRK*17] proposed a method to construct a 3D human body shape and pose dataset by fitting a SMPL model to images. Compared to them, our approach does not require human interventions to validate shape and pose fits. Tan et al. [TBC17] proposed an indirect approach to learn body shape and pose by minimizing the estimated and real silhouettes. Tung et al. [TWYF17] proposed a self-supervised learning motion capture technique that optimizes SMPL body parameters and Kanazawa et al. [KBJM18] proposed an end-to-end learning system of human body and shape based on generative adversarial networks (GANs). More recently, silhouettes [PZZD18, VCR*18] and part segmentations [OLPM*18] are incorporated to improve prediction accuracy. In DensePose [RNI18] uv parametrizations of the segment parts are further provided and annotated on images to establish image to surface dense correspondences.

Concurrently, HoloPose [GK19] is proposed to learn to estimate human body shape and pose from dense correspondences. Our method differs from them in that our approach further leverages dense correspondences to recover 3D human shape from them to supervise ConvNets, requiring no explicit 3D supervision such as 3D joint positions from MoCap. Kolotouros et al. [KPBD19] proposes a method that alternates body shape/pose learning and SMPLify fitting [BKL*16] which uses 3D pose and shape priors. Compare to them, "deform-and-learn" does not rely on these priors and learns them from dense correspondences and 2D key points.

3. Method

The goal of this work is to learn a neural network model that predicts 3D body shape and pose from a single image. To that end, we use dense correspondence annotations [RNI18] between image points and a body surface, which can be annotated on 2D images in-the-wild and provides rich information about body shape and pose. Compared to silhouettes and part segmentations, dense correspondence annotations are less noisy around boundaries and can be obtained with some more additional human efforts where its annotation time is almost the same as that of part segmentation [RNI18].

DOI: 10.2312/egs.20201012

Deform-and-learn iterative training strategy We propose a training strategy called "Deform-and-learn" that alternates deformable surface registration to fit a 3D model to 2D images and training of deep neural network that predicts 3D body shape/pose from a single image. As the first step of an iteration, we train a conditional generative adversarial networks (cGANs) similar to Kudo et al. [KOMO18] that predicts 3D joint positions from 2D joint positions, which guides the registration process. Given image-surface dense correspondences, the registration step fits a template model to images (Section 4). After registration, we obtain a collection of body parameters which is then used as supervisional signals to train deep ConvNets that predicts body parameters (Section 5). The body parameter estimates are used as initial solutions for surface registration in the next round.This training process is iterated for several times to get better results. Note that in the very beginning the initial pose of registration is in the T-pose.

Body shape and pose model Previous approaches [BKL*16, KBJM18] for predicting body shape and pose typically regress body parameters of a pre-built statistical model called SMPL. As opposed to previous approaches using shape blendshapes to model a body shape, we parametrize it by segment scales $\mathbf{s} \in \mathbb{R}^{24}$. This way, all we need is the mean shape, skeleton and skinning weight from SMPL but they could also be coming from any human rigged model. Optionally, pose blendshapes can be incorporated to model complex nonlinear deformation around joints. Since our model does not need the identity or shape parameters of SMPL, the resulting body shape is not confined in the space of statistical models which are constructed from young adult subjects.

The template mesh consists of n vertices, where n is 6980 in this paper. The vertex positions of the template, $\mathbf{v}_1 \dots \mathbf{v}_n$, are denoted by a $n \times 3$ vector, $\mathbf{v} = [\mathbf{v}_1 \dots \mathbf{v}_n]^{\mathrm{T}}$. The pose of the body is defined by a skeleton rig with 23 joints where the pose parameters $\mathbf{a} \in \mathbb{R}^{23\times 3}$ is defined by the axis angle representation of the relative rotation between segments. The body model is posed by joint parameters $\mathbf{a}$ via forward kinematics. In our skinning formulation we multiply a diagonal matrix containing a scaling for joint j, $\mathbf{S}^j = \mathrm{diag}([\mathbf{s}^j, \mathbf{s}^j, \mathbf{s}^j, 1])$, with a homogeneous bone transformation for $\mathbf{A}$ such that $\mathbf{T}^j = \mathbf{A}^j\mathbf{S}^j$. Then the body model is deformed by blending $\mathbf{T}^j$, where we define a deformation function, $\mathbf{v} = X(\mathbf{S}, \mathbf{a})$.

We use the weak-perspective camera model and solve for the global rotation $\mathbf{R} \in \mathbb{R}^{3\times 3}$, translation $\mathbf{t} \in \mathbb{R}^2$ and global scale $s \in \mathbb{R}$. Rather than using other rotational representation such as axis angle, we directly optimize for a rotation matrix with 9 parameters due to its property to represent orientations uniquely in 3D space. Since this approach makes a transformation deviating from a rotation matrix, we applied the Gram Schmidt normalization to orthonormalize the matrix. With the body parameters θ, deformation and projection of vertices into an image is achieved as:

$$\mathbf{x} = s\Pi(\mathbf{R}X(\mathbf{S}, \mathbf{a})) + \mathbf{t} \tag{1}$$

where Π is an orthogonal projection.

4. Image-surface deformable registration

We propose a deformable surface registration technique to fit a template mesh model to images to obtain 3D body shape and pose annotations for training deep ConvNets. Here deformable registration is formulated as a gradient-based method based on back propagation, which can be implemented with a deep learning framework and parallelized with GPUs. With the automatic differentiation mechanisms provided with a deep learning framework, adding and minimizing various kinds of losses have made easier. As a result, the proposed deformable registration technique thus incorporates kinematic, geometric and correspondence losses.

Given image-surface dense correspondences annotated on images, the template mesh is fitted to images by optimizing body parameters $\theta = [\mathbf{a}, \mathbf{S}, \mathbf{R}, s, \mathbf{t}]$ subject to kinematic and geometric constraints. In total, the overall loss function for our registration is of the form:

$$\begin{aligned}\mathcal{L}_{\mathrm{regist}} = \omega_{\mathrm{dense}}\mathcal{L}_{\mathrm{dense}} + \omega_{\mathrm{KP}}\mathcal{L}_{\mathrm{KP}} \\ + \omega_{\mathrm{scale}}\mathcal{L}_{\mathrm{scale}} + \omega_{\mathrm{joint}}\mathcal{L}_{\mathrm{joint}} + \omega_{\mathrm{det}}\mathcal{L}_{\mathrm{det}}\end{aligned} \tag{2}$$

where $\mathcal{L}_{\mathrm{dense}}$ and $\mathcal{L}_{\mathrm{KP}}$ are the dense correspondence and key point losses that penalize the alignment inconsistency of the body model with images defined in terms of dense correspondences and key points. The losses $\mathcal{L}_{\mathrm{scale}}$ and $\mathcal{L}_{\mathrm{joint}}$ is the segment scaling smoothness and kinematic loss for regularization. The transformation determinant loss $\mathcal{L}_{\mathrm{det}}$ makes the determinant of the global transformation positive. In addition, ω_{dense}, ω_{KP}, ω_{scale}, ω_{joint} and ω_{det} are the respective weights for the above defined losses. The initialization of body parameters is provided from the predictions of deep ConvNets. For the very first iteration where the Convnet predictions are not available, segment scale $\mathbf{s}$ is set to 1 for all segments and pose $\mathbf{a}$ is set to 0 for all joints.

Dense correspondence loss Let us define a set of image-surface correspondences $\mathcal{C} = \{(\mathbf{p}_1, \mathbf{v}_{\mathrm{idx}(1)}) \dots (\mathbf{p}_N, \mathbf{v}_{\mathrm{idx}(N)})\}$, where $\mathbf{p}$ is the image points. In addition $\mathrm{idx}(i)$ is the index of the mesh vertex that is matched with image point i. Now we can define the dense correspondence loss as:

$$\mathcal{L}_{\mathrm{dense}} = \sum_{i\in\mathcal{C}} \|\mathbf{p}_i - \mathbf{x}_{\mathrm{idx}(i)}\|^2 \tag{3}$$

Here the mean squared error (MSE) between image point annotations $\mathbf{p}_i$ and the corresponding points on a surface projected to the 2D image $\mathbf{x}_{\mathrm{idx}(i)}$ is calculated.

Key point loss To produce 3D poses with statistically valid depths, the results from cGANs are used to guide deformable registration. Instead of attaching a discriminator to the registration framework, the depth values from cGANs and the ground truth 2D joint coordinates are provided as a soft constraint to constrain the position of the 3D joints based on the MSE loss:

$$\mathcal{L}_{\mathrm{KP}} = \sum_{i\in\mathcal{J}} \|x_i - \bar{x}_i\|^2 + \sum_{i\in\mathcal{J}} \|y_i - \bar{y}_i\|^2 + \sum_{i\in\mathcal{J}} \|z_i - z_i^{\mathrm{GAN}}\|^2 \tag{4}$$

where $\bar{x}_i$ and $\bar{y}_i$ are the ground truth of 2D key points. Also z_i^{GAN} is the depth at joint i predicted by cGANs. Other loss terms are explained in the supplemental material.

5. Estimating 3D body shape and pose from a single image

Using the results obtained by deformable registration as annotations for training deep ConvNets, we regress body shape and pose parameters with an image. We also add the dense correspondence

and key point losses for additional supervisions. In total, we minimize the loss function of the form:

$$\mathcal{L}_{\text{conv}} = \alpha\mathcal{L}_{\text{regress}} + \beta\mathcal{L}_{\text{dense}} + \gamma\mathcal{L}_{\text{KP}} \quad (5)$$

where $\mathcal{L}_{\text{regress}}$ is the regression loss for body parameters. α, β and γ are the respective weights. Let θ_i be the parameters for i-th sample, the regression loss is defined as:

$$\mathcal{L}_{\text{regress}} = \sum_i \text{smooth}_{L1}(\theta_i - \bar{\theta}_i) \quad (6)$$

where $\bar{\theta}$ is the annotation provided from the registration step. Here we use the smooth L1 loss because of its robustness to outliers. This choice was more effective than the L2 loss in contributing to decreasing the error during the iterative training strategy in the presence of potential outliers and noisy annotations.

6. Experimental results

Our method is implemented using Pytorch. We use the Adam optimizer for all the steps in our approach. We use ResNet50 pretrained on the ImageNet as the base network of our body regressor. Training takes 2-3 days using three NVIDIA Quadro P6000 GPUs. The body regressor is trained for 30 epochs with the batch size of 30 and the learning late of 0.0001. We set the parameters in the loss function to $\alpha = \gamma = 1$ and $\beta = 10$. For deformable surface registration, we use the learning rate of 0.1 and batch size of 10. We empirically determined the parameters to $\omega_{\text{dense}} = 1000$, $\omega_{\text{KP}} = 1$, $\omega_{\text{scale}} = 10$, $\omega_{\text{joint}} = 0.001$ and $\omega_{\text{det}} = 1$. For the first training iteration, we use $\omega_{\text{scale}} = 100$ and $\omega_{\text{joint}} = 1$ to make the body model stiff, which is a common strategy in deformable registration [ARV07] to recover a correct global orientation.

6.1. Dataset, protocol and metric

To train the model we use DensePose, [RNI18], DensePoseTrack [NTG*19] and Human 3.6M dataset dataset [IPOS14]. To obtain dense annotations on Human 3.6M, we use two approaches: projecting 3D models obtained with Mosh [LMB14] and predicting using the DensePose model [RNI18]. To obtain dense correspondences to fit a template 3D model to images, we find the closest points from image pixels to surface vertices in UV coordinates of every part. The nearest neighbor search is done in this direction because image pixels are usually coarser than surface vertices. We were able to obtain approximately 100k annotated training images.

We followed the same evaluation protocol (Protocol #1 [IPOS14]) in Human 3.6M dataset as was used in previous approaches [PZDD16, ZHS*17], where it uses 5 subjects (S1, S5, S6, S7, S8) for training and the rest 2 subjects (S9, S11) for testing. The error metric for evaluating 3D joint positions is called mean per joint position error (MPJPE) in *mm*. Following [ZHS*17] the output joint positions from ConvNets is scaled so that the sum of all 3D bone lengths is equal to that of a canonical average skeleton.

6.2. Results and comparisons

In Figs. 1, we show our results on body shape and pose estimation before and after refinement. As we can see from the figure, our technique can predict 3D body shape and pose from in-the-wild images.

We compared our method with state-of-the-art techniques (Table 1). Here we divide the methods into full 3D supervised approaches [KBJM18, GK19, SXLW17] which uses a large amount of 3D pose annotations (and shape when available) paired with images, semi-supervised techniques [RSF18, KBJM18] that uses a limited amount of 3D supervisions and the approaches with no 3D supervision ([KOMO18] and ours). For our techniques, we tested two models: the one trained with dense correspondences obtained by projecting 3D body shapes obtained using Mosh [LMB14] (Ours(Mosh)) and using DensePose predictions [RNI18] (Ours (DensePose)).

From Table 1 we can see that full 3D supervision approaches such as HoloPose [GK19] and Sun et al. [SXLW17] achieves the best results. Rhodin et al. [RSF18] use an auto-encoder to compress visual features and reconstruct 3D pose from it, which does not require a large amount of 3D human pose supervisions. HMR (unpaired) uses 3D pose and body shape dataset only for training GANs to provide 3D constraints without needing to have 3D poses paired with images. Kudo et al. [KOMO18] uses conditional GANs to predict depths from 2D joint coordinates, which learns a model from 2D information only as ours. Our method outperforms Rhodin et al. [RSF18] and Kudo et al. [KOMO18] in terms of MPJPE accuracy and is comparable to [KBJM18] in terms of MPJPE accuracy.

6.3. Is the iterative training strategy effective?

To show the effectiveness of our iterative training strategy, we show a graph with the history of MPJPE errors for Ours(Mosh) in Fig. 2. Here, MPJPE values after deformable registration are calculated on training dataset. Our deform-and-learn strategy starts from image-surface registration using the T-pose as the initial pose. After the first registration phase, the train-set MPJPE for registration results is approx. 110 mm. Then, ConvNets is trained based on these registration results as supervisions. After 1 iteration, the test-set MPJPE of ConvNet predictions is 145 mm, which is slightly high. Next, deformable surface registration is performed again using the results of ConvNets as its initialization. These steps are iterated for several times. This strategy was shown to be effective in gradually decreasing the error. On the other hand, the long training of the regressor (200 epoch) wihout iteration improved MPJPE slightly but not as much as iterative deform-and-learn.

7. Conclusion

We presented a deep learning technique for estimating 3D human body shape and pose from a single color image. To that end, we propose an iterative training approach that alternates between deformable surface registration and training of deep ConvNets, which gradually improves accuracy of predictions by extracting and aggregating 3D information from dense correspondences provided on 2D images. This approach allows us to learn 3D body shapes and pose from 2D dataset only without having to use 3D annotations that are in general expensive to obtain. In future work, we would like to extend our segment scale model to incorporate anisotropic scales and use multi-view images in training to improve accuracy.

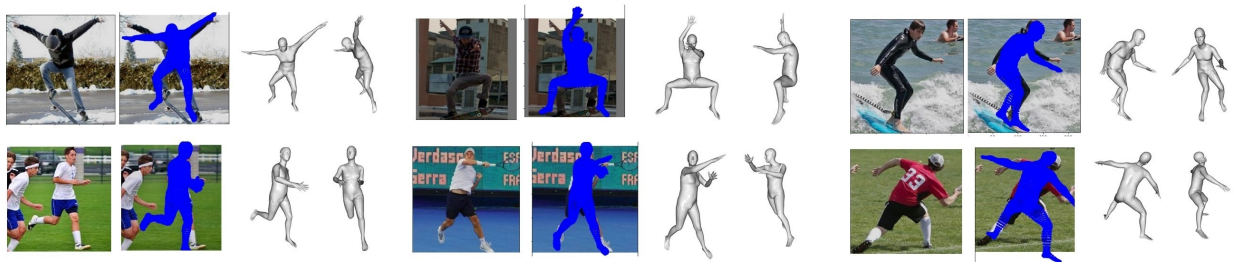

Figure 1: *Qualitative results before refinement. From left to right: original image, overlay, 3D reconstruction results viewing from the front and side. Our technique is able to recover body shape and pose from in-the wild images. Note that the viewing distance of the 3D reconstruction does not exactly match with that of an input image.*

Table 1: *Comparisons with state of the art. MPJPE [mm] is used for error metric.*

Full 3D supervision			Semi 3D supervision	Use external 3D data	No 3D training data		
HMR [KBJM18]	HoloPose [GK19]	Sun et al. [SXLW17]	Rhodin et al. [RSF18]	HMR (unpaired) [KBJM18]	Kudo et al. [KOMO18]	Ours (Mosh)	Ours (DensePose)
87.97	60.27	49.6	131.7	106.84	173.2	96.99	115.3

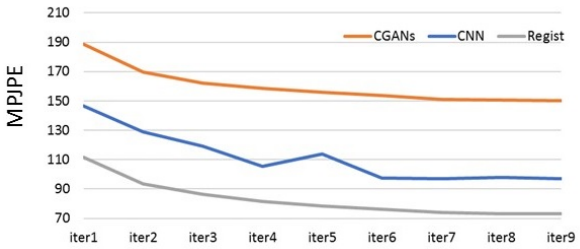

Figure 2: *History of MPJPE with respect to the number of iterations. Blue: MPJPE of ConvNet predictions on testing images; Orange: MPJPE of cGANs predictions for testing; Gray: MPJPE evaluations of registration results on training dataset.*

References

[ARV07] AMBERG B., ROMDHANI S., VETTER T.: Optimal Step Nonrigid ICP Algorithms for Surface Registration. In *CVPR* (2007). 3

[BKL*16] BOGO F., KANAZAWA A., LASSNER C., GEHLER P., ROMERO J., BLACK M. J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV* (2016), Springer, pp. 561–578. 1, 2

[GK19] GULER R. A., KOKKINOS I.: Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR* (2019). 1, 3, 4

[IPOS14] IONESCU C., PAPAVA D., OLARU V., SMINCHISESCU C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI 36*, 7 (jul 2014), 1325–1339. 3

[KBJM18] KANAZAWA A., BLACK M. J., JACOBS D. W., MALIK J.: End-to-end recovery of human shape and pose. In *CVPR* (2018). 1, 2, 3, 4

[KOMO18] KUDO Y., OGAKI K., MATSUI Y., ODAGIRI Y.: Unsupervised adversarial learning of 3d human pose from 2d joint locations, 2018. 2, 3, 4

[KPBD19] KOLOTOUROS N., PAVLAKOS G., BLACK M. J., DANIILIDIS K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV* (2019). 1

[LMB14] LOPER M. M., MAHMOOD N., BLACK M. J.: MoSh: Motion and shape capture from sparse markers. *ACM Trans. Graphics 33*, 6 (Nov. 2014), 220:1–220:13. 3

[LRK*17] LASSNER C., ROMERO J., KIEFEL M., BOGO F., BLACK M. J., GEHLER P. V.: Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR* (July 2017). 1

[NTG*19] NEVEROVA N., THEWLIS J., GÜLER R. A., KOKKINOS I., VEDALDI A.: Slim densepose: Thrifty learning from sparse annotations and motion cues. 3

[OLPM*18] OMRAN M., LASSNER C., PONS-MOLL G., GEHLER P. V., SCHIELE B.: Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *3DV* (Sept. 2018). 1

[PZDD16] PAVLAKOS G., ZHOU X., DERPANIS K. G., DANIILIDIS K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. *CoRR abs/1611.07828* (2016). 3

[PZZD18] PAVLAKOS G., ZHU L., ZHOU X., DANIILIDIS K.: Learning to estimate 3D human pose and shape from a single color image. In *CVPR* (2018). 1

[RNI18] RIZA G., NATALIA N., IASONAS K.: Densepose: Dense human pose estimation in the wild. *arXiv* (2018). 1, 3

[RSF18] RHODIN H., SALZMANN M., FUA P.: Unsupervised geometry-aware representation learning for 3d human pose estimation. In *ECCV* (2018). 3, 4

[SXLW17] SUN X., XIAO B., LIANG S., WEI Y.: Integral human pose regression. *arXiv preprint arXiv:1711.08229* (2017). 3, 4

[TBC17] TAN V., BUDVYTIS I., CIPOLLA R.: Indirect deep structured learning for 3d human body shape and pose prediction. In *BMVC* (2017). 1

[TWYF17] TUNG H., WEI H., YUMER E., FRAGKIADAKI K.: Self-supervised learning of motion capture. In *NIPS* (2017). 1

[VCR*18] VAROL G., CEYLAN D., RUSSELL B., YANG J., YUMER E., LAPTEV I., SCHMID C.: BodyNet: Volumetric inference of 3D human body shapes. In *ECCV* (2018). 1

[ZHS*17] ZHOU X., HUANG Q., SUN X., XUE X., WEI Y.: Weakly-supervised transfer for 3d human pose estimation in the wild. *arXiv preprint arXiv:1704.02447* (2017). 3