3D Human Shape and Pose from a Single Depth Image with Deep Dense Correspondence Enabled Model Fitting

X. Wang[1], A. Boukhayma[3], S. Prevost[1], E. Desjardin[2], C. Loscos[1] and F. Multon[3] †

[1]LICIIS, [2]CReSTIC, University of Reims Champagne-Ardenne, France
[3]Inria, Univ. Rennes, CNRS, IRISA, M2S, France

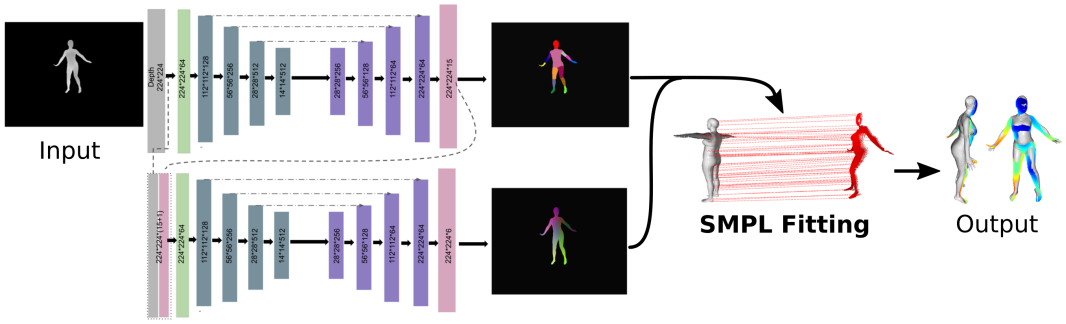

Figure 1: (Step 1) Correspondence between pixels of a depth input image and a template geometry with a double U-Net network to predict body part segmentation and to regress normalized canonical vertex coordinates. (Step 2) SMPL model fitting to the labelled point cloud.

Abstract

We propose a two-stage hybrid method, with no initialization, for 3D human shape and pose estimation from a single depth image, combining the benefits of deep learning and optimization. First, a convolutional neural network predicts pixel-wise dense semantic correspondences to a template geometry, in the form of body part segmentation labels and normalized canonical geometry vertex coordinates. Using these two outputs, pixel-to-vertex correspondences are computed in a six-dimensional embedding of the template geometry through nearest neighbor. Second, a parametric shape model (SMPL) is fitted to the depth data by minimizing vertex distances to the input. Extensive evaluation on both real and synthetic human shape in motion datasets shows that our method yields quantitatively and qualitatively satisfactory results and state-of-the-art reconstruction errors.

CCS Concepts

• ***Computing methodologies*** → ***Motion capture; Motion processing;***

1. Introduction

3D human shape and pose estimation are notorious problems in computer vision and graphics and have several applications like virtual and augmented reality avatars. Using deep learning (DL), some approaches can nowadays recover 3D models of humans from a single image. A first group of approaches relies on fitting the parametric human shape model SMPL [LMR*15] to monocular depth map observations. They generally consider aligning a dozen of joint positions estimated on the image and the ones of the parametric model (such as [JCZ19]). A second group use DL for computing the dense correspondence between a template body shape SMPL and depth image. However, such strategies can fail when the inputs are far from the training data distribution. We propose a hybrid method benefiting from the advantages of both groups of approaches.

2. Method

Given an input depth image of person in tight clothing, our method predicts a mesh representing the corresponding 3D human posed shape in the input camera coordinate frame. This is achieved through the two-stage method depicted in Fig. 1: 1) DL-based es-

† Founded by ANR-JPCH (ANR-17-JPCH-0004). Special thanks to the Centre Image at URCA for their computing resources.

DOI: 10.2312/egp.20221008

timation of the dense correspondence between 3D points and each vertex of the template, 2) alignment of the resulting labelled point cloud with the surface of a template configuration (shape and pose).

Step 1 - dense correspondence estimation. Given a depth image and a template geometry mesh (SMPL [LMR*15] model), we train a convolutional neural network to predict a dense mapping from the depth pixels to the vertices. In SMPL, the shape of a human body is a parametric deformable mesh $\mathcal{M}(\beta,\theta,\gamma)$, parameterized by a shape parameter $\beta \in \mathbb{R}^{10}$, a pose parameter $\theta \in \mathbb{R}^{72}$ and a translation $\gamma \in \mathbb{R}^3$. The model generates a 6890 vertices mesh $\mathcal{M}$. We obtain this mapping through the combination of a body part segmentation map and a pixel-to-vertex correspondence map.

We establish a mapping function $c : \Gamma \rightarrow \mathcal{T}$ putting pixels in the depth image domain $i \in \Gamma$ in correspondence with vertices in the template mesh $j \in \mathcal{T}$ using a deep neural network. To decrease the computation cost, we embed the template geometry in a low dimensional space, noted $E : \mathcal{T} \rightarrow [\![0,1]\!]^6$. The first 3 embedding components are defined as 3 normalized spatial coordinates of the template mesh in the canonical T-pose. Inspired by the image-to-image translation architecture, we stacked two U-Net networks [RFB15]. From the depth input, the first U-Net predicts body part segmentation into the 15 classes (body parts including background class). The second U-Net (regression branch) predicts the 3 last embedding components. The network was trained using 1) a cross-entropy loss on the output of the segmentation branch, and 2) an L_2 loss on the output of the normalized color regression branch.

Finally, to obtain correspondences c for a given depth image to the template geometry, we first map the image pixels to the low dimensional embedding using the double U-Net architecture. The vertex j matching pixel i is then defined as the nearest template vertex in the embedding space.

Step 2 - Model Fitting From the depth map and pixel-to-vertex correspondences, the following function fits the SMPL model to the observation, recovering human shape and pose parameters:

$$E(\theta,\beta,\gamma) = \lambda_D E_D(\theta,\beta,\gamma) + \lambda_\theta E_\theta(\theta) + \lambda_\beta E_\beta(\beta). \quad (1)$$

The data term E_D consists in a L_2 penalty between pixel i's 3D point p_i, obtained using the intrinsic matrix and the pixel's depth value, and the corresponding vertex $v_{c(i)}$, summed over all pixels that belong to the body region $\Omega \subset \Gamma$ in the segmentation map. We use a robust differential Geman-McClure penalty function ρ to deal with noisy estimates. E_θ represents the body pose prior $E_\theta(\theta) = \sum \exp(\theta_i)$ which penalizes knee and elbow joints that bend unnaturally. The shape prior E_β implements an L_2 regularization on the shape parameters $E_\beta(\beta) = ||\beta||^2$. Hyper parameters $\lambda_s, \lambda_\theta, \lambda_\beta$ are trade-off weights between the objective function terms.

3. Evaluation

We conduct experiments on standard datasets of 3D human shape in motion: SURREAL [VRM*17] (synthetic data), DFAUST [BRPMB17] (real data) and DanseDB (dancedb.eu) (composed of synthetic human models fitted to real motion capture data). Our final testing dataset is created by rendering 3D models from these three datasets to simulate depth images of same resolutions from different viewpoints. We then uniformly sampled approximately 50,000 training frames and 10,000 testing frames. We quantify the reconstruction quality with the Mean Average Vertex Error in millimeter (mm), averaged subsequently over all testing frames.

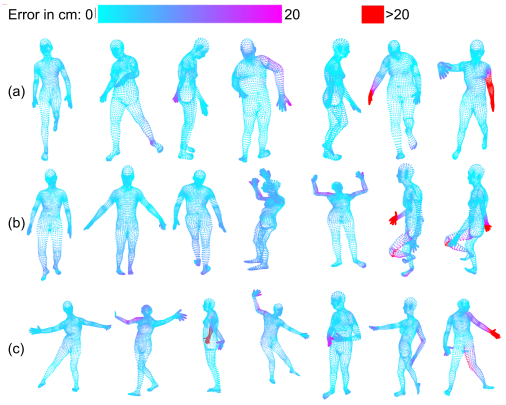

Figure 2: Spacial distribution of reconstruction errors on (a) SURREAL, (b) DFAUST and (c) DanseDB.

This error (see Figure 2) was 50mm and 53mm for SURREAL and DFAUST datasets respectively, compared to methods using model fitting only (140mm and 110mm for [LMR*15] for example). While most errors are close to few millimeters (light blue), large errors remain in challenging cases, like side views and self-occluded areas. More recent works [JCZ19] report 16mm and 8mm errors respectively on these datasets when using DL.

4. Conclusion

We demonstrate in this work that combining DL and model fitting significantly improves the quality of the results, compared to using model fitting only. It could be interesting to replace the double U-Net network used in this work by recent DL-based works to check if adding model fitting also improves the accuracy of these methods. Additional ablation studies would also estimate the actual impact of each of our contributions. We would also like to analyze further the geometrical distribution of the reconstruction error.

References

[BRPMB17] BOGO F., ROMERO J., PONS-MOLL G., BLACK M. J.: Dynamic faust: Registering human bodies in motion. In Proceedings of the IEEE conference on computer vision and pattern recognition (2017), pp. 6233–6242. 2

[JCZ19] JIANG H., CAI J., ZHENG J.: Skeleton-aware 3d human shape reconstruction from point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision (2019), pp. 5431–5441. 1, 2

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) 34, 6 (Oct. 2015), 248:1–248:16. 1, 2

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (2015), Springer, pp. 234–241. 2

[VRM*17] VAROL G., ROMERO J., MARTIN X., MAHMOOD N., BLACK M. J., LAPTEV I., SCHMID C.: Learning from synthetic humans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017), pp. 109–117. 2