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

Overview

Goal: 3D human shape and pose estimation

Interest: Several applications, notably for creating avatars in virtual and augmented reality applications.

Key challenges: **Reconstructing** both **shape** and **pose** of an actor using a **single** RGB or RGB-D **view**.

Our proposition: a **hybrid method** benefiting from the advantages of Deep Learning (DL) and optimization approaches.

1) DL network: estimation of the **dense correspondence between pixels in a depth image and each vertex of a human template**.

2) optimization framework: optimal template configuration (shape and pose) to **align the resulting labeled point cloud with the surface of the template**.

Related Work

Focus: monocular depth image input containing a single person with close-fitting clothing.

Two groups of DL based methods stand out:

1. **Fitting the parametric human shape model** SMPL [LMR∗15] to monocular depth images.
   ➔ Aligning the joint positions estimated on the image to the ones of the parametric model [JCZ19].

   Limitations: objective function criterion is based on very sparse information (dozen of joint center positions).

2. **Computing the dense correspondence between a template body shape SMPL and a point cloud** (computed using the depth image).
   ➔ Learned by:
   i) amassing training datasets with ground truth correspondence [ZKB20],
   ii) feature descriptors attached to RGB, depth or point cloud [HYVH20].

   Limitations: can fail when the inputs are far from of the training data distribution.

Methodology

Input: 1 depth image containing a close-fitting clothed person

Output: a mesh M (6 890 vertices) representing the corresponding 3D human posed shape in the input camera coordinate frame.

Human representation: SMPL [LMR∗15], parametric deformable mesh **M(β,θ,γ)**

$\beta \in \mathbb{R}^{10}$: human shape parameter

$\theta \in \mathbb{R}^{72}$: pose parameter

$\gamma \in \mathbb{R}^{3}$: translation

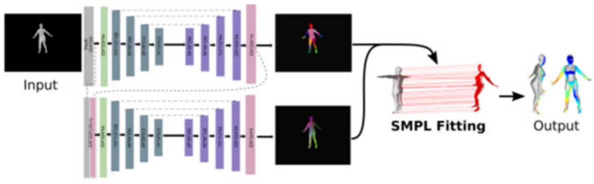


Figure 1: overview of the two-steps process

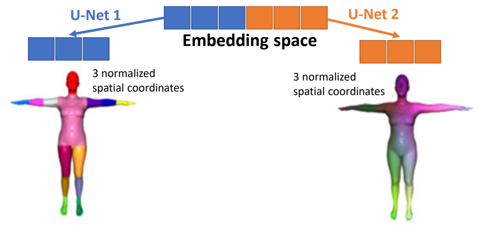
Step 1) Mapping between pixels of a depth input image and a template geometry with a double U-Net network to predict body part segmentation and to regress normalized canonical vertex coordinates.

Step 2) SMPL model fitting to the labelled point cloud.

Step 1 - Dense correspondance

Inputs: 1 depth image + 1 template geometry mesh (fig 2)

Goal: map pixels of the depth image to the template geometry embedding space (6D embedding). **2 U-Net [RFB15] networks**



- U-Net 1: depth input image ➔ **body part segmentation** (15 template classes). 1 class = 1 color.
  Training with the combination of cross-entropy loss.
- U-Net 2: (**regression** branch): body part segmentation + depth image ➔ 3-channel image.
  Training with an L2 loss on the output of the normalized color.

➔ **Estimate a Pixel-to-Vertex Correspondence:** vertex j matching pixel i (nearest template vertex in the embedding space)

Step2 - Model fitting

Using the template Geometry Embedding, fits the SMPL model to the 3D point cloud ➔ compute **human shape (β) and pose (θ) parameters**

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

- E_D: data term = L2 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 in the segmentation map.
- E_θ: body pose prior penalizes joints that bend unnaturally. $E_\theta(\theta)=\Sigma\exp(\theta_i)$
- E_β: shape prior = L2 regularization on the shape parameters $E_\beta(\beta) = \|\beta\|^2$.
- $\lambda_s,\lambda_\theta,\lambda_\beta$: trade-off weights between the objective function terms.

Training

- Batchsize of 12 using the RMSprop optimizer
- Learning rate of $6.14e^{-4}$.
- We run the optimization for 20 iterations.

Results

Datasets

- Standard datasets of **3D close-fitting clothed human shape in motion**: SURREAL [VRM∗17] (synthetic data), DFAUST [BRPMB17] (real data) and DanseDB (dancedb.eu) (synthetic human models fitted to real motion capture data).
- Datasets rendered to **simulate depth images of same resolutions but with different viewpoints**.
  50,000 training frames and **10,000 testing frames u**niformly sampled.

Qualitative results

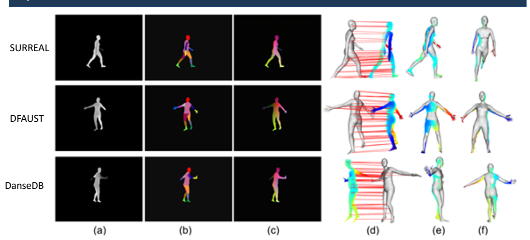


Figure 2: Results. (a) Input depth image; (b) Output human part segmentation; (c) Regressed template vertex color; (d) Correspondences between the depth point cloud and the fitted mesh; (e) & (f) Output fitted mesh visualized from 2 different viewpoints. The point cloud is colored according to the depth values.

Error Visualization

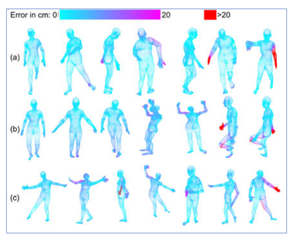


Figure 3: Spatial distribution of reconstruction errors on (a) SURREAL, (b) DFAUST and (c) DanseDB.

Evaluation Metric.
Reconstruction quality = Mean Average Vertex Error

Most errors are close to few millimeters (light blue)
Large errors (red) remain in challenging cases, like side views and self-occluded areas.

Computation time:

- NN inference stage: about 35ms
- Optimization stage: 6.43s on a NVIDIA 1080Ti GPU.

These results show the robustness of our method to changes in body poses, shapes, self-occlusions and viewpoints.
T**the accurate and dense mapping** between depth pixels and fitted 3D model topology provides more detailed information compared to optimization methods that use only the joint centers.

Acknowledgement

anr® ANR-JPCH (ANR-17-JPCH-0004).

Centre Image Maison de la Simulation de Champagne-Ardenne

Université de Reims Champagne-Ardenne

M2S Sport Sciences

Université Rennes 2

LICIIS DIGIT

EG'22 TheEvent

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.

[HYVH20] HUANG X., YANG H., VOUGA E., HUANG Q.: Dense correspondences between human bodies via learning transformation synchronization on graphs. Advances in Neural Information Processing Systems 33 (2020).

[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.

[KPD19] KOLOTOUROS N., PAVLAKOS G., DANIILIDIS K.: Convolutional mesh regression for single-image human shape reconstruction. In CVPR (2019).

[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.

[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.

[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.

[WHC∗16] WEI L., HUANG Q., CEYLAN D., VOUGA E., LI H.: Dense human body correspondences using convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016), pp. 1544–1553.

[WXZ∗20] WANG K., XIE J., ZHANG G., LIU L., YANG J.: Sequential 3d human pose and shape estimation from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 7275–7284.

[ZKB20] ZHU T. L., KARLSSON P., BREGLER C.: Simpose: Effectively learning densepose and surface normals of people from simulated data.In ECCV (2020).