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. Lосos\(^1\) and F. Multon\(^3\)

1. LCIIS, 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 creating 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 [ZKH20].

   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 [ZKH20].
   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 (LRM+15), parametric deformable mesh \( \beta \in \mathbb{R}^{10} \), human shape parameter \( \theta \in \mathbb{R}^{72} \), pose parameter \( y \in \mathbb{R}^2 \).

Step 1 - Dense correspondence

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 [RF15] networks

• U-Net 1: depth input image → body part segmentation (15 template classes) 3 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)

Dataset:

• Standard datasets of 3D close-fitting clothed human shape in motion: SURREAL [VRM+17] (synthetic data), DRAUST (BPRM17) (real data) and DanceDB (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 uniformly sampled.

Qualitative results

Figures 2: Results. (a) Input depth image. (b) Output human part segmentation. (c) Regressed template vertex colors. (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.

Figures 3: Spatial distribution of reconstruction errors on (a) SURREAL, (b) DRAUST and (c) DanceDB.

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.

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.

REFERENCES


