Eurographics Workshop on 3D Object Retrieval (2019)
S. Biasotti, G. Lavoué, B. Falcidieno, and I. Pratikakis (Editors)

POP: Full Parametric model Estimation for Occluded People

Riccardo Marin[1], Simone Melzi[1], Niloy J. Mitra[2], Umberto Castellani[1]

[1]University of Verona [2]University College London

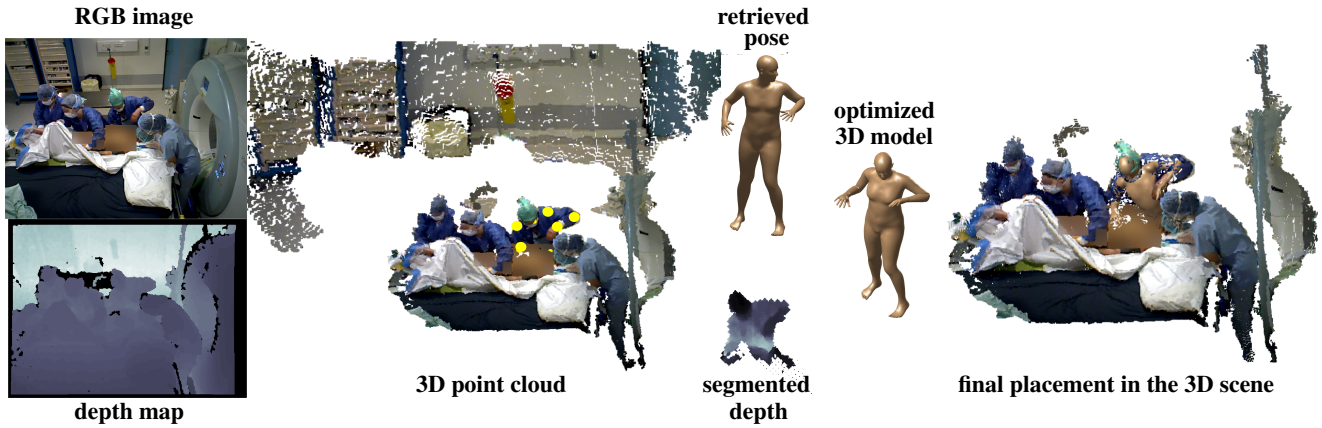

Figure 1: *Our estimation pipeline tested on a challenging example from the MVOR dataset [SIK*18]. From left to right: RGBD input, 2D image (top) and depth map (bottom); point cloud generated from the input and the camera parameters (top) and 3D joints of the estimated skeleton are depicted as yellow disks on the point cloud (bottom); data-driven pose initialization (top), and estimated segmentation of the depth map (bottom); model optimized on the input data; and the final model placed in the 3D space. The result is compelling for the quality of the estimation and the placement of the 3D shape, even in presence of several challenging properties of the input.*

Abstract

In the last decades, we have witnessed advances in both hardware and associated algorithms resulting in unprecedented access to volumes of 2D and, more recently, 3D data capturing human movement. We are no longer satisfied with recovering human pose as an image-space 2D skeleton, but seek to obtain a full 3D human body representation. The main challenges in acquiring 3D human shape from such raw measurements are identifying which parts of the data relate to body measurements and recovering from partial observations, often arising out of severe occlusion. For example, a person occluded by a piece of furniture, or being self-occluded in a profile view. In this paper, we propose POP, a novel and efficient paradigm for estimation and completion *of human shape to produce a full parametric 3D model directly from single RGBD images, even under severe occlusion. At the heart of our method is a novel human body pose retrieval formulation that explicitly models and handles occlusion. The retrieved result is then refined by a robust optimization to yield a full representation of the human shape. We demonstrate our method on a range of challenging real world scenarios and produce high-quality results not possible by competing alternatives. The method opens up exciting AR/VR application possibilities by working on 'in-the-wild' measurements of human motion.*

CCS Concepts

• ***Computing methodologies*** → ***Shape modeling;*** *Shape analysis;*

1. Introduction

Analysis and modeling of human shape from images and video is an important topic that is widely studied across several research domains including robotics for human-robot interaction [AC07, SLAL18], in pattern recognition for video surveillance and action recognition [KF18], in biometry for person (re-)identification and gait recognition [ZZS*17, BC07], and in computer graphics for authoring digital content creation [LMR*15, TWYF17, VCR*18].

In early efforts of human motion analysis, the overall aim was to accurately estimate 2D and, to a limited extent, 3D skeleton joint-

DOI: 10.2312/3dor.20191055

locations as a proxy for recovering *human pose* (i.e., human skeleton) [SBIK16]. A particularly challenging scenario consists of estimating *both* human pose and shape 'in-the-wild,' i.e., when one or more people move in a very generic environment and are oblivious of the acquisition goals [SBIK16]. In this scenario, since the subjects move uninhibited, occlusions are commonly arising due to the presence of other objects or from self-occlusion (see Figure 1).

We investigate the above problem relying on RGBD sensors for input snapshots. The available depth information, albeit noisy, effectively avoids the scale-ambiguity problem encountered using single RGB images instead [ZSG*18]. Further, depth helps to determine relative position between human body and occluding objects (e.g., furniture). With this motivation, we investigate the following problem: *Given a single RGBD image of human(s) in a natural environment, obtain a full parametric 3D estimation of human shape(s), even under occlusion*.

The above problem is challenging due to three main reasons: (i) the raw input does not come with any object/human segmentation; (ii) information about which parts of human subjects are occluded and what objects cause the occlusion is unknown; and (iii) the raw RGBD scans are noisy and suffer from heterogeneous point cloud density based on camera location. We propose POP, a fully automatic pipeline that produces accurate human pose, shape and placement in the 3D space from single RGBD images, even in the presence of very significant occlusion.

Our main contributions are (i) proposing a first method explicitly designed for the analysis and modeling of human occlusion and self occlusion in single RGBD images; (ii) introducing a complete and fully automatic pipeline for 3D human pose and accurate full shape estimation that can deal with occlusions; (iii) developing an occlusion-aware shape retrieval strategy that recovers plausible information on the missing body parts, provides a reliable model parameter initialization for joints location and shape, and imposes a new constraint that avoids degenerate shape on the unseen part; (iv) segmenting the human subject(s) from the rest of the scene without requiring an explicit learning procedure or involving green screens; and (v) hallucinating the shape of the occluded part by exploiting the data-driven prior via a novel idea akin to *null-space* that constraints the optimization procedure to reliably estimations.

2. State of the Art

Human body modeling is a widely studied issue over the last two decades [SBIK16,IPOS14,CI11]. In the most of the proposed methods, the main objective is 3D pose estimation, i.e., location of 3D joints of the body according to a given skeleton [SBIK16]. Usually a two-steps procedure is employed: first, joints locations are estimated on the 2D image domain, and then, 3D joints are computed using a regression approach or a model-based re-projection strategy [BKL*16, LRK*17]. Recently, instead to rely over 2D joints estimation, direct methods have been proposed to esteem 3D pose directly from the entire image by exploiting additional information enclosed in the pixels [KBJM18].

An emerging trend is to estimate the 3D pose and the full body shape within the same framework, namely, *end-to-end modeling* methods [KBJM18,MCG*18,TWYF17]. The main idea consists of adopting a template-based approach estimating the shape and pose parameters of a given morphable models properly designed for human-shapes [LMR*15, JSS18]. Methods differ between those that use only 2D image and those that employ RGBD data [ZSG*18, BBLR15, IPOS14, CI11]. In the RGBD domain the main effort is devoted to 3D pose estimation in real-time [ZSG*18], by heavily harness the temporal constraint that can be introduced for video sequences [BBLR15, BRPMB17]. Other methods use multiple devices to enlarge the acquisition view and reduce the effect of occlusions (see survey [ZSG*18]). In contrast, we focus on the case of recovering full human body shape from a *single* RGBD scan with background clutter (i.e., without human body being pre-segmented) and in presence of medium-strong *occlusion*.

Methodologically, the estimation of shape and pose is usually obtained by formulating an optimization model [BKL*16, Lop14]. Recently, deep neural network methods is the widest used technique [KBJM18, TWYF17, VCR*18, DSO*17, SBIK16]. This has led to very impressive results even from single 2D image at the cost of a very accurate manual annotation of 2D and 3D joint positions, foreground-background segmentation, 2D silhouettes and so on [TWYF17, IPOS14, VRM*17]. However, modeling occlusion directly from RGBD inputs still remains a significant open challenge in this domain.

Dealing with occlusions. Although widely appreciated that human modeling can be drastically affected in the presence of occlusions and missing parts, very few works have treated this topic [SLAL18]. Some methods address implicitly this issue by imposing a pose-prior [AB15], by allowing only plausible poses. Similarly, learning-based approaches regularize the pose and shapes according to the examples observed during the training phase [HMH10, GYRF14]. These strategies can reduce the conditioning of occlusions, but they are not designed for this purpose. In [RGL15], a method for explicitly estimating the 3D pose of occluded parts from RGBD data was introduced. The position of the invisible joint is predicted through a classification of the semantic label of the occluded object. An alternative for human pose estimation from partially occluded RGBD data was proposed in [AD15], that relies on a probabilistic occupancy grid that is exploited to identify hidden body parts. Recently, the first systematic study [SLAL18] of various types of occlusions in 3D human pose estimation have also shown that employing data augmentation with new occluded scenes improves the overall pose estimation.

Our method. To the best of our knowledge, POP is the first method that proposes an explicit strategy to estimate the full body-shape, 3D pose and the 3D placement of the human body in the presence of strong occlusions and missing parts. These three estimations are provided consistently and at the same time. These are complete novels in literature.

Our method is focused on RGBD data trying to achieve the best results from both appearance (2D) and geometric (3D) data. We propose a two-steps procedure where 2D pose is estimated from RGB image while 3D pose and the full body-shape is estimated from the depth map. Our 2D pose estimation is used for the initialization procedure, and in the following optimization the estimated model is free to move avoiding the conditioning of a

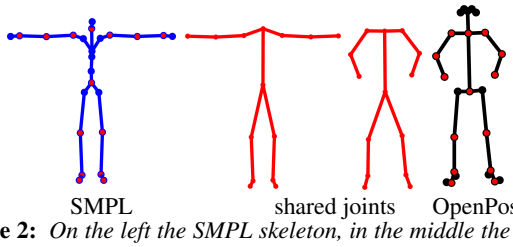

Figure 2: *On the left the SMPL skeleton, in the middle the shared joints and the OpenPose skeleton on the right.*

bad starting pose. Moreover, since we evaluate the confidence of the 2D estimate, only the most reliable joints are considered. Our method fosters an optimization approach with the use of a Convolutional Neural Network for only the 2D pose phase. We adopt a model based approach using the very popular SMPL morphable model [LMR*15]. Our strategy is data-driven since we rely on the assumption that alike occluded shape has been already observed on a dataset that is recovered through a 3D shape retrieval procedure. Similar idea was exploited in [BMB*11, MDO*15, IDY*18] for pose estimation only.

3. Overview

In our pipeline, we have carefully integrated some public available datasets and tools. A complete review is out of the scope of this paper and we refer to their references for details.

OpenPose is a fully automatic method for detecting the 2D pose of multiple people in an RGB image [WRKS16, SJMS17, CSWS17] wherein a non parametric pose representation, referred as *Part Affinity Fields* [CSWS17], has been proposed. This representation consists in a set of 2D vector fields, each of which encodes the orientation and the location of a limb in the image. A learning strategy is adopted on the whole image with high accuracy and real-time performance. For each of these joints, a confidence value is also provided. The final full body pose corresponds to a set of labelled 2D key points as ordered joints of a human skeleton.

The SMPL model [LMR*15] is a skinned vertex-based parametric model for the full human body. SMPL has few parameters but sufficient to generate a wide set of human bodies with different pose and shape. Pose and shape are controlled by two different sets of parameters: $\boldsymbol{\theta} \in \mathbb{R}^{72}$ are the pose ones defined as the relative rotation of each of 24 joints with respect to its parent in a hierarchical kinematic tree; $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are the shape parameters. SMPL provides a skeleton composed by the 24 joints. Of these, 15 joints can be matched with 15 joints in the OpenPose skeleton. Figure 2 shows the 24 joints from SMPL, the 25 joints from OpenPose, and the shared 15 joints directly used in in our optimization.

The SURREAL dataset [VRM*17] is a large-scale synthetically-generated dataset of more than 6 million frames. This dataset contains realistic scenes of people that are rendered using the SMPL model with real motion capture information. For each frame, a ground truth pose, a depth map, and a segmentation mask are provided.

OpenDR [LB14] is an approximate and differentiable renderer (DR) that explicitly connects the relationship between the SMPL parameters and the projection of the corresponding 3D shape to a 2D image. OpenDR is publicly-available and well suited to work with SMPL model and SURREAL dataset. Starting from a shape generated by SMPL in the 3D space, with OpenDR, we associate to this shape a 2D image and a 2D depth map representation of the scene. As already highlighted, the relation between the SMPL shape (i.e. its parameters) and this 2D representation is differentiable, and so can be used in an optimization pipeline.

4. Method

4.1. Pipeline in brief

The entire pipeline, depicted in Figure 3, can be outlined as follows:

INPUT: Single RGBD image with internal camera parameters.
STEP 1: From the input depth map D_{in} and camera parameters, we estimate the point cloud PC of the scene.
STEP 2: $\boldsymbol{J}_{2D}$ a standard skeleton on the 2D image is obtained using OpenPose [CSWS17].
STEP 3: A subset of the 2D OpenPose joints are then lifted on the 3D space obtaining $\boldsymbol{J}_{3D}$.
STEP 4: We retrieve the most similar 3D skeleton with respect to $\boldsymbol{J}_{3D}$ in a subset of the SURREAL dataset and select the correspondent SMPL pose parameters $\widetilde{\boldsymbol{\theta}}$.
STEP 5: The joints of SMPL are aligned to the $\boldsymbol{J}_{3D}$ optimizing for the scale of SMPL.
STEP 6: Based on the retrieval, we segment the human body input depth $\widetilde{D}_{in}$ and the human body point cloud $H \subset PC$.
STEP 7: We iteratively optimize the SMPL parameters in order to fit the J_{3D} and the nearest neighbor energy E_{NN} between the points in H and the SMPL surface.
STEP 8: We deform the SMPL minimizing the E_{depth}.
OUTPUT: The optimized 3D model placed in the 3D scene.

We now describe each step of our method. For each choice, we explicitly clarify the respective strategy for handling occlusions.

4.2. Input.

Our input is a single RGBD image with the internal camera parameters of the acquisition sensor. We use both the image representation and the 3D information in term of 3D cloud of points. We refer to D_{in} for the input depth map and PC for the point cloud. Although we now describe handling of a single human, the method can be easily iterated to deal with multi-person scenarios (see Section 5).

4.3. Coarse joints location and occlusion detection

2D skeleton. We apply the OpenPose framework to the input RGB image to obtain the 2D joints of the skeleton of a human body. We use the version 1.4 relying over BODY_25 skeleton model. An example of the skeleton provided by OpenPose is shown in Figure 2. OpenPose returns only visible joints, that in our case are at most 25. After a re-targeting procedure between the OpenPose and the SMPL skeletons we define $\boldsymbol{J}_{2D}$ as the subset of the 15 joints of SMPL that are shared with OpenPose and visible (see Figure 2

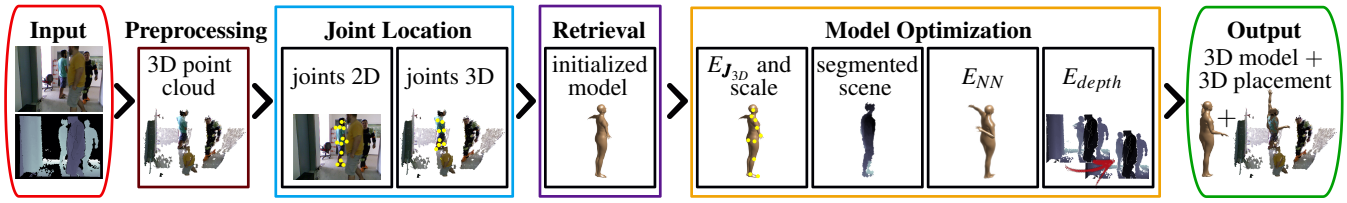

Figure 3: *POP pipeline. From left to right: Input (red), 3D point cloud construction (dark red), Coarse joints location and occlusion detection 4.3 (light blue), Retrieval-based model initialization 4.4 (purple), Model optimization 4.5 (yellow) and the Output (green).*

where overlapping joints are marked in red). The remaining joints are classified as occluded.

3D skeleton. Using the camera parameters we can project $\boldsymbol{J}_{2D}$ to the 3D space on the point cloud *PC*. However, these 3D points can be wrongly estimated due to noise and located in some inconsistent region far in the background. We compute a basic statistics to automatically detect and remove such unreliable points as outliers. Indeed we obtain the set of 3D joints $\boldsymbol{J}_{3D}$ after a position refinement to accomodate a consistent skeleton.

4.4. Retrieval-based model initialization.

From the SURREAL dataset, we select 1.6 milions frames from all the *run1* training set folder. For each of such frames, we apply the same steps explained above on the input RGBD data, providing a coherent representation for the input data and the frames from SURREAL. We explore all these frames to find the best match for which exist a transformation in the 3D space that minimize the average of the distance between all the joints $\boldsymbol{J}_{3D}$ of the input and the 3D joints estimated on the SURREAL frame. We consider only frames that have the same visible part and therefore the same occlusion.

For each considered instance i in the retrieval dataset we look for a global homogeneous transformation $\boldsymbol{T}$ composed by scale, rotation, reflection and translation given by the solution of:

$$\arg\min_i \left(\arg\min_{\boldsymbol{T}} (\|\boldsymbol{T}(\boldsymbol{J}_i) - \boldsymbol{J}_{3D}\|_F) \right), \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\boldsymbol{J}_i$ is the set of joints of the frame i. Note that restricting this search to the frames that share the same visible part $\boldsymbol{J}_i$) and $\boldsymbol{J}_{3D}$ are composed by the same joints thus Equation 1 is well defined. The solution is the index i of frame that best matches the $\boldsymbol{J}_{3D}$ skeleton. Every frame in the SURREAL dataset is associated with a SMPL set of parameter to generate the related body instance. We take those related to the solution frame retrieved by Equation 1 and use them to set SMPL pose parameters $\widetilde{\boldsymbol{\theta}}$.

Initialization of the SMPL parameters. From the Equation 1 we obtain the transformation $\boldsymbol{T}$. Applying the translation and scale components to the SMPL model, we have a good initialization in the 3D space placement. Note that the initialization $\widetilde{\boldsymbol{J}}$ obtained from the retrieval step also provides a good initialization for the occluded part. Thanks to this data-driven prior we both avoid an implausible initialization of SMPL (that direct parameters optimization can provide) and we improve efficiency starting closer to the correct pose.

4.5. Model Optimization.

We optimize the SMPL model in order to fit the input data. We refer to SMPL shape as $\mathcal{M}$ and to its vertices $\boldsymbol{V}_{\mathcal{M}} \in \mathbb{R}^{6890\times 3}$ represented as the collection of the 3D coordinates of its embedding.

Joints and scale optimization. Our SMPL model is initialized with the retrieved pose $\boldsymbol{\theta}$ and is placed coherently in the 3D space with respect to the $\boldsymbol{J}_{3D}$. The $\boldsymbol{J}_{3D}$ can also be involved in the optimization as a stability penalty; we force the joints of the SMPL that correspond to the joints in $\boldsymbol{J}_{3D}$ (denoted as $\widetilde{\boldsymbol{J}_{SMPL}} \subseteq \boldsymbol{J}_{SMPL}$) to remain near to $\boldsymbol{J}_{3D}$. This is expressed by the penalty term:

$$E_{\boldsymbol{J}_{3D}} = \|\boldsymbol{J}_{3D} - \widetilde{\boldsymbol{J}_{SMPL}}\|_F. \tag{2}$$

A first optimization is thus performed on the SMPL joints placement and on the scale of SMPL with respect to the energy $E_{\boldsymbol{J}_{3D}}$.

Constraints on the parameters. We start the optimization with strong constraints over $\boldsymbol{\theta}$ parameters because we would avoid extremely unreliable rotations. Subsequently we weaken them, increasing adherence with the seen joints.

Scene segmentation. Applying the OpenDR we obtain a synthetic depth map $D_{\boldsymbol{\beta},\boldsymbol{\theta}}$, which directly depends on the SMPL parameters. $D_{\boldsymbol{\beta},\boldsymbol{\theta}}$ and D_{in} differ for the presence in the D_{in} of all object outside our target; while in $D_{\boldsymbol{\beta},\boldsymbol{\theta}}$ all the points that do not belong to SMPL are on the far plane, in D_{in} other objects participate. $D_{\boldsymbol{\beta},\boldsymbol{\theta}}$ can be considered as a mask of the subject, and we can apply it to D_{in}, cutting out an approximated segment for the human. To improve the approximation of this segment we analyze the neighbor of the points that belong to the human segment. Let p be one such point. We consider a 2D neighbor defined on the 2D image B_p. For all points $q \in B_p$ we have two possibilities: q belongs to the human body segments or q belongs to the background. In the first case, we assign to q its value in D_{in}. In the second case, we classify q with respect to the inequality $|D_{in}(p) - D_{in}(q)| < \gamma$ for a fixed threshold $\gamma > 0$. If this inequality holds we assign to q the value $D_{in}(q)$, otherwise we set its value to the background. Through this procedure, we define a *clean* input depth map $\widetilde{D}_{in}$ that contains the values of the original D_{in} for all the points that are expected to belong to the human body, and the background value for the others. $\widetilde{D}_{in}$ is comparable to the artificial depth map $D_{\boldsymbol{\beta},\boldsymbol{\theta}}$ as they only describe the depth of the human body points in the scene. We refer to the human body segment in the point cloud as $H \subset PC$.

Fitting to the visible part. We compute $\pi_{\text{NN}}(V_{\mathcal{M}})$, the list of the vertices $V_{\mathcal{M}}$ obtained as the ordered euclidean nearest neighbor

with respect to the points in H. Relying on $\pi_{\text{NN}}(V_{\mathcal{M}})$, we optimize first for the pose parameters $\boldsymbol{\theta}$, and then jointly for the pose and the shape ($\boldsymbol{\theta}$ and $\boldsymbol{\beta}$) minimizing $E_{\text{NN}} = \|H - \pi_{\text{NN}}(V_{\mathcal{M}}(\boldsymbol{\theta},\boldsymbol{\beta}))\|_F$.

Consistency with the depth map. To optimize directly the occluded body part in the closest plausible place, we define a *null-space*, where human body parts are *not allowed*. To do this we rely over the information from the depth map of D_{in} that is not represented in $\widetilde{D}_{in}$. It includes all objects in the environment that are possible causes of occlusions, thus it specifies all the places where the human body should not appear. We want to exploit these elements to hide unseen parts if this is a reliable solution. We generate the depth map $\hat{D}_{in}$ as: $\hat{D}_{in} = \text{far}, \text{if } D_{in}(u,v) \in \widetilde{D}_{in}$ otherwise $\hat{D}_{in} = D_{in}(u,v)$, where u and v are the image plane coordinates and *far* is the value assigned to the far plane. Then, we minimize $E_{depth} = \|min(D_{\boldsymbol{\beta},\boldsymbol{\theta}}, \hat{D}_{in}) - D_{in}\|_F$ to have $D_{\boldsymbol{\beta},\boldsymbol{\theta}}$ approximating D_{in} by hiding part behind objects present in the scene that are nearer to the camera or exploiting the body itself. Figure 10 shows an example where the left arm is moved to be self-occluded by the body, and the right one is hidden behind the other person in foreground.

5. Experiments and Results

We provide evaluations on different datasets and challenging cases highlighting the robustness to the occlusions. We omit comparison with other methods; it would be ambiguous because POP is the first method that provides at the same time an estimation of the shape, the pose and the 3D placement of the human body shape, it relies over depth information and also aim to solve occlusions.

Datasets. We evaluate our method on different datasets, that differ for conditions and challenges. F-BODY [SPT15], designed for human body occlusion (self-imposed or generated by people interactions). BIWI RGB-ID dataset [MBF*14] offers a variety of human shapes in similar pose and camera view. MVOR [SIK*18], a recent dataset with RGBD images in operating room. These scenes are heavily occluded and human elements are hidden from a variety of exacting factors. We select frames from other datasets to analyze different challenges: far views [CMA*17], different occlusions and poses [AD15] and body shapes [SX13]. Finally, we test our method on frames from SURREAL providing quantitative measures that permit future comparisons. We remark the huge variety of scenario from disparate environment and settings considered.

Quantitative evaluation on SURREAL. To provide a quantitative evaluation of our method we perform an extensive experiment on the SURREAL dataset. We select 18 frames with self occlusions from 18 different videos not used in the retrieval. For each frame, we evaluate the shape and pose parameters, and surface difference between the ground truth provided by SURREAL and the estimated one. The errors are computed as follows.

$$\textbf{Shape error (w.r.t. } \boldsymbol{\beta}\textbf{)} = \mathbf{err}_{\boldsymbol{\beta}} = \frac{\|\boldsymbol{\beta}_{gt} - \boldsymbol{\beta}\|_F}{\|\boldsymbol{\beta}_{gt}\|_F}. \tag{3}$$

$$\mathbf{err}_{\boldsymbol{J}_{SMPL}} = \sum_{j=1}^{23} \frac{\|\boldsymbol{J}^{SMPL}_{gt}(j) - \boldsymbol{J}^{SMPL}_{\boldsymbol{\beta},\boldsymbol{\theta}}(j)\|_F}{23}. \tag{4}$$

	mean	std
$\mathbf{err}_{\boldsymbol{\beta}}$	1.0008	0.013
$\mathbf{err}_{\boldsymbol{J}_{SMPL}}$	0.1157	0.0741
$\mathbf{err}_{\boldsymbol{pose}}$	0.1515	0.0950
$\mathbf{err}^{visible}_{\boldsymbol{pose}}$	0.1335	0.0788
$\mathbf{err}^{occluded}_{\boldsymbol{pose}}$	0.2724	0.1994
$\mathbf{err}_{\boldsymbol{p2p}}$	0.0228	0.0064

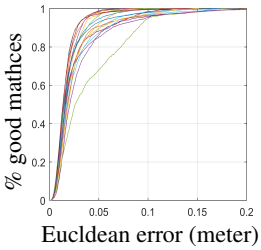

Figure 4: *Quantitative evaluation; mean and standard deviations on the left, and cumulative frequencies on the right*

$$\mathbf{err}_{\boldsymbol{pose}} = \sum_{j=1}^{14} \frac{\|\boldsymbol{J}^{3D}_{gt}(j) - \boldsymbol{J}^{3D}_{\boldsymbol{\beta},\boldsymbol{\theta}}(j)\|_F}{14}. \tag{5}$$

$$\mathbf{err}^{visible}_{\boldsymbol{pose}} = \sum_{j \in visible} \frac{\|\boldsymbol{J}^{visible}_{gt}(j) - \boldsymbol{J}^{visible}_{\boldsymbol{\beta},\boldsymbol{\theta}}(j)\|_F}{\sharp(visible)}. \tag{6}$$

$$\mathbf{err}^{occluded}_{\boldsymbol{pose}} = \sum_{j \in occluded} \frac{\|\boldsymbol{J}^{occluded}_{gt}(j) - \boldsymbol{J}^{occluded}_{\boldsymbol{\beta},\boldsymbol{\theta}}(j)\|_F}{\sharp(occluded)}. \tag{7}$$

$\mathbf{err}_{\boldsymbol{J}_{SMPL}}$ evaluates the difference between the 24 ground truth SMPL joints and the one obtained from our optimization. $\mathbf{err}_{\boldsymbol{pose}}$ is the same restricted to the 15 joints shared by SMPL and OpenPose. $\mathbf{err}^{visible}_{\boldsymbol{pose}}$ is limited to the joints (≤ 15) that are considered as visible by our pipeline. $\mathbf{err}^{occluded}_{\boldsymbol{pose}}$ consider the joints (≤ 15) that were not found by our pipeline. All these errors are computes excluding the root joint that only represents the placement in the 3D space. Together with these shape and pose measures we compute the normalized registration error:

$$\mathbf{err}_{\boldsymbol{p2p}} = \sum_{p \in H} \frac{\|H(p) - \pi_{\text{NN}}(V_{\mathcal{M}}(\boldsymbol{\theta},\boldsymbol{\beta}))(p)\|_F}{\sharp(H)}. \tag{8}$$

defined through the point-to-point distances between H and registered SMPL surface. The mean and the standard deviation of these errors are reported in the Table in Figure 4. Except for the $err_{\boldsymbol{\beta}}$ all the others errors are reported in meters. On the right of Figure 4, a quantitative evaluation of the point-to-point distance between our output and H is depicted. These curves represent cumulative frequencies of the above error for each of the considered frames. For the majority of subjects, our method stays for 90% under the threshold of 6cm of error. Although a fair comparison with other methods is not possible we can note that our error is coherent with the declared surface error for the state-of-the art method in [VCR*18] on the entire T1 Surreal middle frame, i.e., a less challenging scenario. In Figure 5, we visualize the error encoded by the heatmap; white is 0 while black represents large error saturated to $3cm$.

Qualitative pose estimation on the other Datasets. The retrieval step already provides good approximations of the $3D$ human pose, as shown in Figure 6, highlight the power of the proposed retrieval and the data driven approach. For all the examples in Figure 6 we

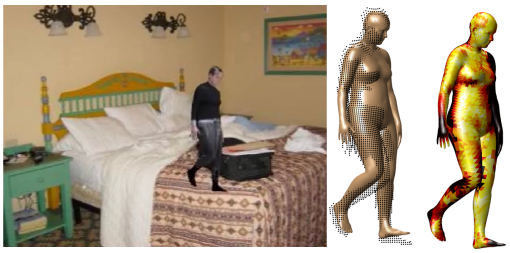

Figure 5: *The fitting error between H and our output, encoded by the heatmap, white is* 0 *error while black is* 3*cm or larger error.*

Figure 6: *Some 3D pose approximations obtained from the only retrieval step. These are the SMPL initializations in our pipeline.*

provide the final registration in Figures 7,8,9, showing how much the rest of the pipeline improves the quality of the results. Figure 10 shows the contribution of the consistency in the depth map.

Full pipeline results We show results in a large variety of cluttering, occlusions and noisy conditions. Results in Figure 8 are obtained on dataset [SPT15]. We would like to underline that the child in Figure 9 is an extreme case for the shape estimation. Finally, in Figure 7 we show that our method is robust also to the presence of many people and on the right of Figure 9 a case of a far and occluded subject.

Implemetation and Timing Both the SMPL model and the OpenDR tool are built upon a Python based autodifferentiation framework. For OpenPose, we use the free online version with the suggested parameter setting. The solution of 1 is solved using the *procustes* MATLAB function. Our pipeline needs around 5 minutes to produces the final 3D pose and shape estimation for a human body. We perform our experiments on an Intel 3.6 GHz Core i7-7700 cpu with 16GB RAM. To make our work fully reproducible we will release i) our code, ii) the identification of the tested scenes, and iii) the 2D joints estimated with Open Pose.

6. Conclusion and future work

We presented **POP**, a fully automatic pipeline for *end-to-end* modeling of human shape where RGBD data are exploited to estimate the pose and the accurate shape of a real person observed on a very generic scenarios (i.e., in the wild). We propose for the first time a *modeling from reality* method that is properly designed for handling occlusions. We have shown that ingredients and suggestions for modeling occlusions can be effectively employed in the proposed pipeline, from 2D joint estimation to model initialization and missing parts completion. Although the proposed method is based on the SMPL template our approach can be naturally extended on other parametric models.

References

[AB15] AKHTER I., BLACK M. J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015* (2015). 2

[AC07] A. G. M., C. S. A.: Human-robot interaction: A survey. *Found. Trends Hum.-Comput. Interact. 1*, 3 (2007), 203–275. 1

[AD15] ABDALLAH DIB F. C.: Pose estimation for a partially observable human body from rgb-d cameras. 2, 5, 8

[BBLR15] BOGO F., BLACK M. J., LOPER M., ROMERO J.: Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 2300–2308. 2

[BC07] BOULGOURIS N. V., CHI Z. X.: Human gait recognition based on matching of body components. *Pattern Recognition 40* (2007). 1

[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 *Computer Vision – ECCV 2016* (2016). 2

[BMB*11] BAAK A., MÜLLER M., BHARAJ G., SEIDEL H.-P., THEOBALT C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In *IEEE 13th International Conference on Computer Vision (ICCV), (IEEE 2011)* (11 2011), pp. 1092–1099. 3

[BRPMB17] BOGO F., ROMERO J., PONS-MOLL G., BLACK M. J.: Dynamic FAUST: Registering human bodies in motion. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017* (2017). 2

[CI11] CATALIN IONESCU FUXIN LI C. S.: Latent structured models for human pose estimation. In *International Conference on Computer Vision* (2011). 2

[CMA*17] CAMPLANI M., MADDALENA L., ALCOVER G. M., PETROSINO A., SALGADO L.: A benchmarking framework for background subtraction in rgbd videos. In *International Conference on Image Analysis and Processing* (2017), Springer, pp. 219–229. 5, 7

[CSWS17] CAO Z., SIMON T., WEI S.-E., SHEIKH Y.: Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR* (2017). 3

[DSO*17] DUSHYANT M., SRINATH S., OLEKSANDR S., HELGE R., MOHAMMAD S., HANS-PETER S., WEIPENG X., DAN C., CHRISTIAN T.: Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics 36*, 4 (2017). 2

[GYRF14] GHIASI G., YANG Y., RAMANAN D., FOWLKES C. C.: Parsing occluded people. In *IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 2401–2408. 2

[HMH10] HUANG J.-B. Y., MING-HSUAN: Estimating human pose from occluded images. In *Asian Conference on Computer Vision – ACCV* (2010), pp. 48–60. 2

Figure 7: *An example from SBM dataset [CMA*17]. Our method offers a good solution for reconstruct group of people without ambiguity.*

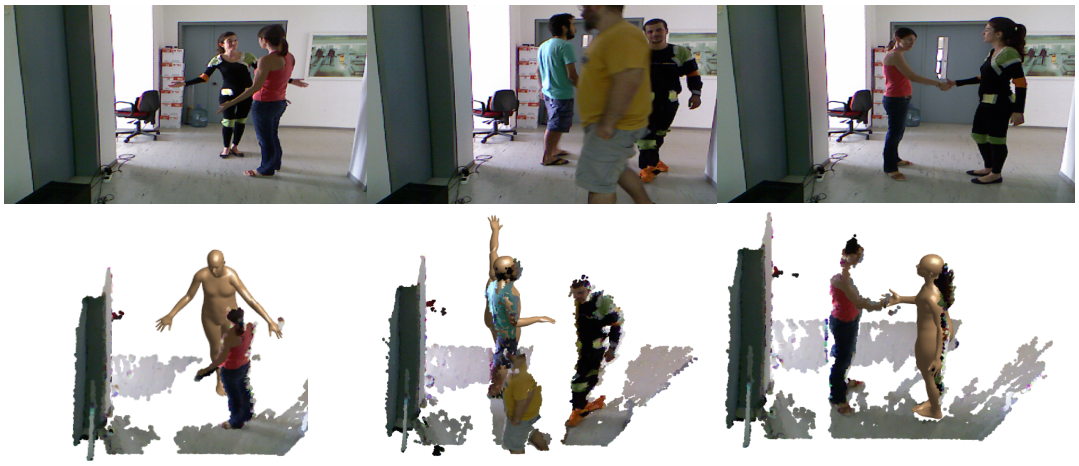

Figure 8: *Some experiments from [SPT15] dataset show different type of occlusions caused by external agents. Multi-person does not introduce confusion. In the middle, the arm has been placed to a different solution for the occluded part, but consistent with acquired view.*

[IDY*18] IQBAL U., DOERING A., YASIN H., KRÜGER B., WEBER A., GALL J.: A dual-source approach for 3d human pose estimation from single images. *Computer Vision and Image Understanding* (2018). 3

[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 Transactions on Pattern Analysis and Machine Intelligence* (2014). 2

[JSS18] JOO H., SIMON T., SHEIKH Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. *CoRR abs/1801.01615* (2018). 2

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

[KF18] KONG Y., FU Y.: Human action recognition and prediction: A survey. *CoRR abs/1806.11230* (2018). 1

[LB14] LOPER M. M., BLACK M. J.: Opendr: An approximate differentiable renderer. In *ECCV* (Cham, 2014), Springer International Publishing, pp. 154–169. 3

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. *ACM Transaction on Graphhics 34*, 6 (2015), 248:1–248:16. 1, 2, 3

[Lop14] LOPER M.: Chumpy autodifferentiation library, 2014. http://chumpy.org/. 2

[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 *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017* (Piscataway, NJ, USA, July 2017), IEEE. 2

[MBF*14] MUNARO M., BASSO A., FOSSATI A., VAN GOOL L., MENEGATTI E.: 3d reconstruction of freely moving persons for re-identification with a depth sensor. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on* (2014), IEEE, pp. 4512–4519. 5

[MCG*18] MOHAMED O., CHRISTOPH L., GERARD P.-M., V. G. P., BERNT S.: Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *International Conference on 3D Vision (3DV)* (Verona, Italy, 2018). 2

[MDO*15] MITZEL D., DIESEL J., OSEP A., RAFI U., LEIBE B.: A fixed-dimensional 3d shape representation for matching partially observed objects in street scenes. In *2015 IEEE International Conference on Robotics and Automation (ICRA)* (2015), pp. 1336–1343. 3

[RGL15] RAFI U., GALL J., LEIBE B.: A semantic occlusion model for human pose estimation from a single depth image. In *IEEE Confer-*

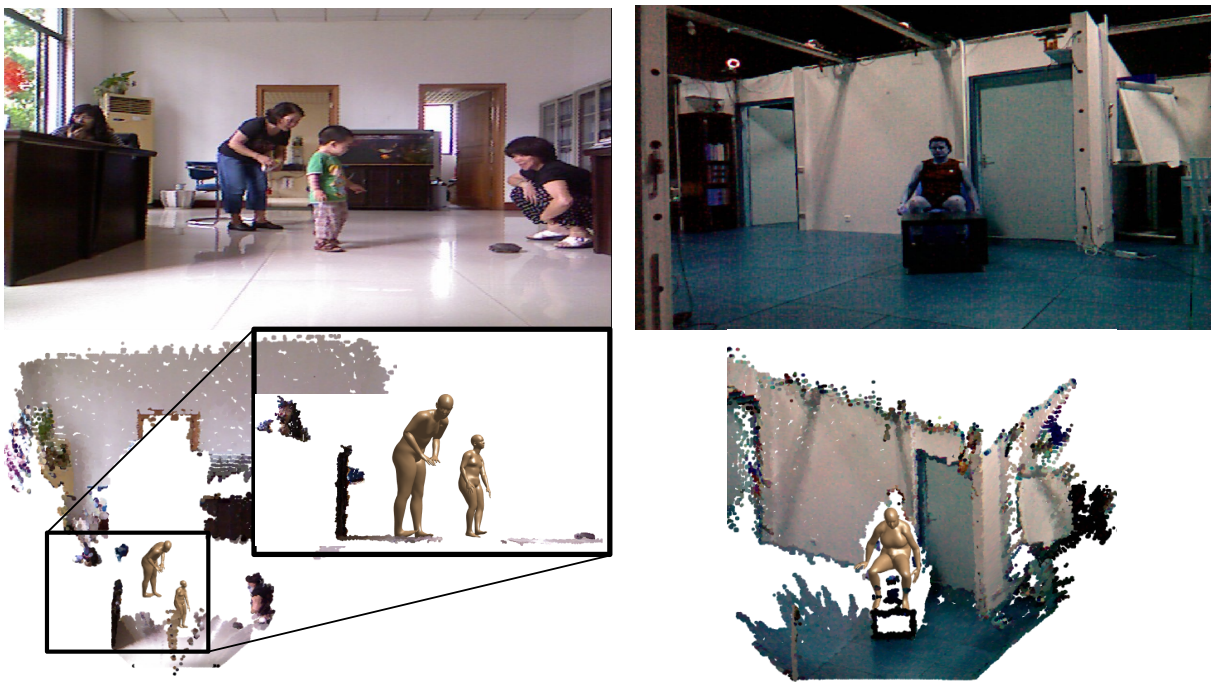

Figure 9: *Two results from [SX13] and [AD15] respectively. Child is an extreme case of human body shape due to his proportions. In spite this, we have a good approximation. On the right, a challenging case of a man sat far from cam and partially occluded by a table.*

ence on Computer Vision and Pattern Recognition Workshops (CVPRW) (2015), pp. 67–74. 2

[SBIK16] SARAFIANOS N., BOTEANU B., IONESCU B., KAKADIARIS I. A.: 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding 152* (2016), 1–20. 2

[SIK*18] SRIVASTAV V., ISSENHUTH T., KADKHODAMOHAMMADI A., DE MATHELIN M., GANGI A., PADOY N.: Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation. *arXiv preprint* (2018). 1, 5

[SJMS17] SIMON T., JOO H., MATTHEWS I., SHEIKH Y.: Hand keypoint detection in single images using multiview bootstrapping. In *CVPR* (2017). 3

[SLAL18] SÁRÁNDI I., LINDER T., ARRAS K. O., LEIBE B.: How robust is 3d human pose estimation to occlusion? In *IROS Workshop - Robotic Co-workers 4.0* (2018). 1, 2

[SPT15] SIGALAS M., PATERAKI M., TRAHANIAS P.: Full-body pose tracking - the top view reprojection approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence PP*, 99 (2015). 5, 6, 7

Figure 10: *Synthetic depth map, before and after its optimization.*

[SX13] SONG S., XIAO J.: Tracking revisited using rgbd camera: Unified benchmark and baselines. In *The IEEE International Conference on Computer Vision (ICCV)* (December 2013). 5, 8

[TWYF17] TUNG H., WEI H., YUMER E., FRAGKIADAKI K.: Self-supervised learning of motion capture. In *Neural Information Processing Systems (NIPS)* (2017). 1, 2

[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 *European Conference on Computer Vison (ECCV)* (2018). 1, 2, 5

[VRM*17] VAROL G., ROMERO J., MARTIN X., MAHMOOD N., BLACK M. J., LAPTEV I., SCHMID C.: Learning from synthetic humans. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017), pp. 4627–4635. 2, 3

[WRKS16] WEI S.-E., RAMAKRISHNA V., KANADE T., SHEIKH Y.: Convolutional pose machines. In *CVPR* (2016). 3

[ZSG*18] ZOLLHÖFER M., STOTKO P., GÖRLITZ A., THEOBALT C., NIESSNER M., KLEIN R., KOLB A.: State of the Art on 3D Reconstruction with RGB-D Cameras. *Computer Graphics Forum (Eurographics State of the Art Reports 2018) 37*, 2 (2018). 2

[ZZS*17] ZHENG L., ZHANG H., SUN S., CHANDRAKER M. K., TIAN Q.: Person re-identification in the wild. pp. 3346–3355. 1