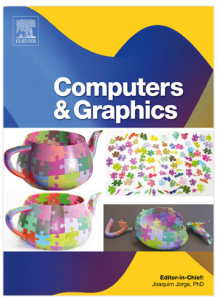
Evaluation of hybrid deep learning and optimization method for 3D human pose and shape reconstruction in simulated depth images

Xiaofang Wang, Stéphanie Prévost, Adnane Boukhayma, Eric Desjardin, Céline Loscos, Benoit Morisset, Franck Multon

Journal Pre-proof



Evaluation of hybrid deep learning and optimization method for 3D human pose and shape reconstruction in simulated depth images

Xiaofang Wang, Stéphanie Prévost, Adnane Boukhayma,
Eric Desjardin, Céline Loscos, Benoit Morisset, Franck Multon

PII: S0097-8493(23)00134-6
DOI: https://doi.org/10.1016/j.cag.2023.07.005
Reference: CAG 3744

To appear in: *Computers & Graphics*

Received date : 10 February 2023
Revised date : 22 June 2023
Accepted date : 4 July 2023

Please cite this article as: X. Wang, S. Prévost, A. Boukhayma et al., Evaluation of hybrid deep learning and optimization method for 3D human pose and shape reconstruction in simulated depth images. *Computers & Graphics* (2023), doi: https://doi.org/10.1016/j.cag.2023.07.005.

Computers & Graphics (2023)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Evaluation of hybrid deep learning and optimization method for 3D human pose and shape reconstruction in simulated depth images

Xiaofang Wang[a,b], Stéphanie Prévost[b], Adnane Boukhayma[c], Eric Desjardin[b], Céline Loscos[b], Benoit Morisset[a], Franck Multon[c,d]

[a]*AI Verse*
[b]*University of Reims Champagne Ardenne, Reims, France*
[c]*Inria, Univ. Rennes, CNRS, IRISA, Rennes, France*
[d]*Univ. Rennes, M2S, Rennes, France*

ARTICLE INFO

Article history:
Received June 22, 2023

Keywords: Human motion capture, shape reconstruction, deep learning, computer vision, depth sensor

ABSTRACT

In this paper, we address the problem of capturing both the shape and the pose of a human character using a single depth sensor. Some previous works proposed to fit a parametric generic human template into the depth image, while others developed deep learning (DL) approaches to find the correspondence between depth pixels and vertices of the template. We designed a hybrid approach, combining the advantages of both methods, and conducted extensive experiments on the SURREAL [1], DFAUST datasets [2] and a subset of AMASS [3]. Results show that this hybrid approach enables us to enhance pose and shape estimation compared to using DL or model fitting separately. We also evaluated the ability of the DL-based dense correspondence method to segment also the background - not only the body parts. We also evaluated 4 different methods to perform the model fitting based on a dense correspondence, where the number of available 3D points differs from the number of corresponding template vertices. These two results enabled us to better understand how to combine DL and model fitting, and the potential limits of this approach to deal with real-depth images. Future works could explore the potential of taking temporal information into account, which has proven to increase the accuracy of pose and shape reconstruction based on a unique depth or RGB image.

Acknowledgements

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

1. Introduction

Reconstructing the pose and shape of a human character using a single camera is of great interest in applications where the user is represented by a realtime avatar, such as in immersive social media or mixed reality videogames.

Several approaches proposed to reconstruct the 3D shape and pose of a human character using a single RGB image, by fitting parametric models (such as SMPL [4], SMPLify [5]) to the RGB information, or by directly learning parameters of parametric models [6, 7]. Previous works demonstrated that Deep Learning (DL) was promising to learn the correspondence

between the image domain and the SMPL parameter space [8]. However, this correspondence problem is complex due to high variation in human poses, shapes, and camera viewpoints [9, 9, 8, 10].

We make the assumption that adapting these RGB-based approaches to depth images is promising. Indeed, using depth instead of RGB images helps to resolve ambiguity from 2D to 3D. With the dissemination of low-cost depth sensors in the consumer market, such as the Microsoft Kinect, reconstructing the body shape and pose from depth images and point clouds has become a very active field of research. Most previously proposed methods focused on pose and shape reconstruction from human part [11], [12], skeleton joints [5] [13], or dense correspondence [14], but remain not enough precise. Indeed, depth images provide incomplete and noisy information, mainly due to occlusions and the inherent noise of cheap sensors, which makes it challenging to build a complete and accurate body surface [15, 16, 17, 18, 19]. In complex scenes with a ground, object, and background, this problem is more complex, as the character may not be correctly segmented.

In [20], we proposed to combine the advantages of DL-based dense correspondence estimation, with a parametric model fitting for the fine tuning of the shape and the pose. We assumed that it would enhance the accuracy of the pose and shape reconstruction, compared to using them separately. Hence, the two main hypotheses we validated in this paper are:

H1 Depth image background segmentation could be performed jointly with the dense correspondence. Hence, similar to [9, 10], as a first step, we establish dense correspondences via mapping 3D vertices to the color domain. We use a Double-Unet network [21] to obtain this color embedding for each depth pixel. A first U-Net aims at segmenting the depth images into 15 classes (body parts and background), which should help a second U-Net to regress color embedding for each pixel.

H2 Using dense correspondence as an input of the model fitting algorithm should improve the performance of pose and shape reconstruction. Most of the previous works, based on this model fitting, used joint position estimation as an input of the optimization, which makes the approach very sensitive to noise and inaccuracies. We assume that using thousands of pixel-to-vertex correspondences instead of 15 joint positions would increase the accuracy of the reconstruction.

For dense correspondence estimation, we trained a neural network to map depth pixels to a low-dimensional canonical template geometry representation (geometry embedding). This representation entails normalized spatial coordinates of the T-pose human SMPL template vertices, in addition to body part segmentation labels. Based on the success of previous works [8, 10], we regress this representation in an image-to-image manner. One of the key ideas is to associate a specific color encoding for the background, to jointly perform body parts and background segmentation. This pixel-to-vertex correspondence is next used to optimize the shape and pose parameters of SMPL, inspired by previous works on hands [22, 23]. However, the number of available 3D points differs from the number of template vertices in the SMPL model. Hence, we propose in this paper to test several strategies to select the best correspondence between the 3D points and the template vertices.

We compared our method to state-of-the-art competition that solves for both monocular RGB and depth inputs on standard human shape in motion datasets following the experimental setting of [18], using synthetic (SURREAL), pseudo-real (DanseDB), and real (DFAUST) data. We also provided an in-depth ablative analysis of the various components involved in our method. These first tests are applied to segmented images where the background is suppressed, and there is no occlusions with the environment. We then evaluated the ability of this hybrid approach to deal with more complex depth images, with background. More specifically, we evaluated if the dense correspondence network based on the geometry embedding can actually segment the background in a specific color.

In the following, we first review previous works most related to our approach in section 2. This two-step approach is presented in section 3. We then present a comparison to previ-

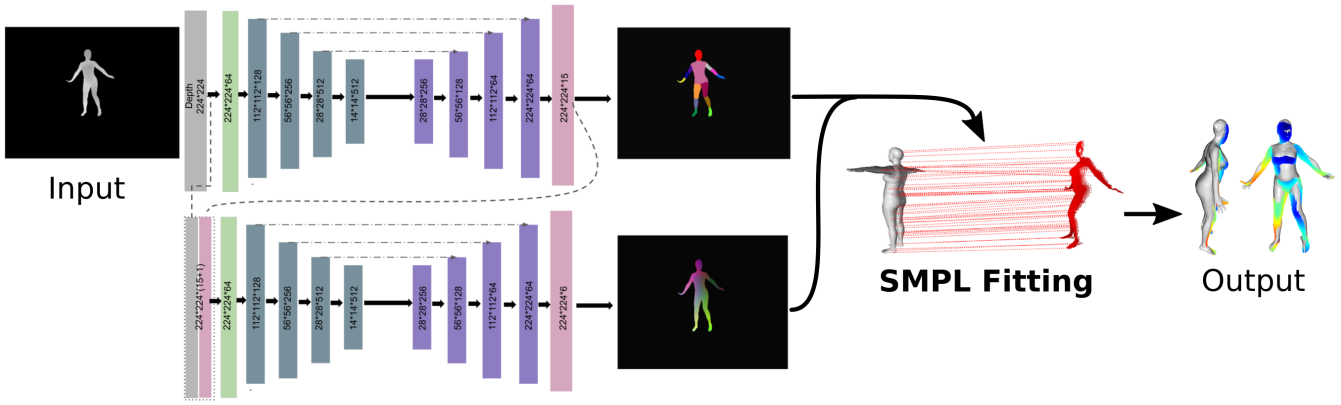


Fig. 1. Overview of the proposed framework. Our method can predict 3D human shape and pose from an input-depth image. A double U-Net network is applied to predict body part segmentation and to regress normalized canonical vertex coordinates. These outputs are used to compute dense correspondence between the input depth pixels and the template geometry via the nearest neighbor in a low-dimensional embedding. We then fit the SMPL model to the input depth by minimizing the distances between vertices and their corresponding depth pixels. The final output is shown on the right-hand side from two viewpoints with the overlaid input depth point cloud.

ous works in section 4, and perform an extensive evaluation of the approach with non-segmented depth images. Finally, we explore various strategies to take into account the dense correspondence in model fitting (section 5) , before concluding.

2. Related works

Human 3D shape reconstruction and pose estimation have generated vast literature. We refer the reader to [24, 25, 26] for more extensive overviews.

3D Human Shape and Pose from Depth Images. Previous 3D human body modeling from depth images can be roughly categorized into template-based, template-free capture and hybrid methods. The template-based methods utilize template priors for the 3D body model recovery, such as embedded skeletons [27, 12], template models [17, 28, 15, 29], or parametric models [16, 18]. With the evolution of depth sensing, range data acquired by commodity depth sensors such as the Microsoft Kinect, can be used as prior information. An improved SCAPE model can also be fitted to the range data [14, 30]. Researchers also proposed several different cues from a depth sensor to estimate pose and shape via a silhouette, depth or color data [30, 27, 29, 14]. Bashirov et al. [13] proposed a neural network that used the 3D joints position delivered by the Kinect API to infer SMPL pose parameters. The DoubleFusion approach [16] starts with a pre-constructed 3D template mesh and uses a template-free method (i.e., DynamicFusion [31]) to update the current mesh in combination with the SMPL parametric model to construct an inner human body. Although it shows very promising performances, the initial configuration and subject pre-scanning are not trivial inputs. Recently, DL-based methods have shown impressive performance improvement. Most of these learning methods rely on 3D human models, such as SMPL. Zhang et al. [32] trained a weakly supervised network from depth or point cloud to learn 3D joints from annotated 2D joints, but they did not recover human shape information. Wang et al. [18] proposed to regress 3D coordinates of mesh vertices at different resolutions from the latent features of point clouds. Jiang et al. [19] also proposed a deep network that takes 3D point cloud as input and learns to predict SMPL shape and pose parameters.

3D Human Shape and Pose from RGB Images. With the development of deep neural networks, capturing a 3D human shape and pose from a single color image has become possible through several diverse approaches [25]. A family of works leveraged 2D joint information in predicting 3D human pose [33] and shape [34]. Bogo et al. [5], when proposing SMPLify, applied a CNN-based method to predict 2D joint locations and then fitted a 3D body SMPL model to estimate 3D body shape

and pose. Other methods used regression of 3D human model parameters. They used deep neural networks as encoders to estimate the pose and shape parameters directly from images. For instance, Kolotouros et al. [6] proposed a deep network to infer SMPL parameters through iterating between learnable regression and the optimization-based approach SMPLify [5]. ExPose [35] is a deep neural network predicting the whole set of SMPL-X [36] parameters to overcome the problem of lacking training data for the human body model. Kanazawa et al. [7] employed adversarial learning by using a generator to predict parameters of SMPL, and a discriminator to distinguish the real mesh instances and the predicted ones. Other deep learning-based methods [37, 38, 39, 40] inferred the 3D body shape or mesh directly from color images using convolutional networks. Graph CNN method [37] first attached the extracted features from an input color image to the 3D vertex coordinates of a template mesh, and then predicted the vertex coordinates of the 3D body meshes using a convolutional mesh regression. Moon et al. [39] proposed a new heat-map representation, called "lixel", to recover 3D human meshes. [40] used image convolutional features and Transformers [41] to estimate a human mesh from a single RGB image. [42]

3D Human Shape and Pose using deep learning. Several methods enable to compute correspondences between human shapes in arbitrary poses [9, 10, 43]. Finding correspondences across images or point clouds is a fundamental building block for many 2D/3D computer vision tasks, such as reconstruction or tracking. Bogo et al. [14] proposed to optimize parameters of 3D human body model fitted through point cloud corresponded vertices. However, the correspondences were computed by a nearest-neighbor algorithm based on Euclidean distance, which demands a good initialization. Other works relied on an underlying parametric model of a human, such as SMPL, and directly performed a correspondence regression. A strong benefit of this was that the 3D model shares the same topology across different people. DensePose [8] showed that this can be learnt via gathering dense correspondences between the SMPL and body data using the COCO dataset, including simulated data [44]. Another popular type of method consisted in learning feature descriptors attached to RGB, depth, or points cloud. While early works used hand-crafted shape descriptors to identify geometric features [45], Huang et al. [43] used deep learning. They applied PointNet++ [46] to learn a representation of each point cloud, and further enforce local smoothness to compute dense correspondences across full or partial human shapes. They used a depth image as input and learned a descriptor for each pixel. Tan et al. [10] learned an embedding from RGB images that follows the geodesic properties of an underlying 3D surface, which enabled the inference of human correspondences.

Recent works proposed an alternative approach to dense correspondence, by using encoder to recover an implicit function of the human body surface based on sparse 3D points, and then fit a SMPL model [47, 48, 49]. These methods are able to jointly represent body pose, shape, and clothing geometry and obtained impressive results, even for fine details on the surface mesh.

In this paper, we explore the limits of coupling dense correspondence and model fitting [20] to handle background segmentation together with image-to-vertex correspondence. We also aim at demonstrating that combining this type of approach with SMPL model fitting should enhance the accuracy of the pose and shape reconstruction. However, this raises the question of finding a good manner to take into account this complex point-to-vertex correspondence in the model fitting method.

3. Our 3D human shape reconstruction from depth images

Given an input depth image containing a person with close-fitting clothes, 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. The formalization of the 3D model used in our approach is described in section 3.1. In the first step, a convolutional network (see section 3.2) predicts a segmentation of the input depth image into several human body parts, along with a correspondence map associating pixels to template mesh vertices. Then, pixel-to-vertex correspondences are established

using the segmentation and correspondence maps. In the second step, a parametric shape model is fitted to the depth image using the resulting correspondence maps (see section 3.3).

3.1. 3D Human Model

To model the 3D shape and pose of the character, we used the SMPL model. The shape of a human body is defined by a parametric deformable mesh $\mathcal{M}(\beta, \theta, \gamma)$. Shape parameters β are coefficients of low-dimensional shape space. γ is the global translation. The pose of the body is defined by a skeleton rig with 23 joints; pose parameters θ represent the relative rotation between parts. The model generates a triangle mesh $\mathcal{M}$ with 6890 vertices:

$$\begin{aligned} \mathcal{M}(\beta, \theta, \gamma) &= W(\mathcal{T}_p(\beta, \theta), J(\beta), \theta, \mathcal{W}) + \gamma, \\ \mathcal{T}_p(\beta, \theta) &= \mathcal{T} + B_S(\beta) + B_P(\theta), \end{aligned} \tag{1}$$

where W is a linear blend skinning function with vertex-joint assignment weights $\mathcal{W}$, and $J(\beta)$ is a joint location regressions function. The pose parameters θ encode the global rotation and the rotation angles of each skeleton joint, while the shape parameters β contain the coefficients of the ten most significant PCA components of the human shape learned from registered real human scans. $\mathcal{T}_p(\beta, \theta)$ is the deformed template mesh in a default body pose. It is expressed as the sum of a template mesh $\mathcal{T}$ with the shape and pose blendshape functions B_S and B_P, which add shape and pose-dependent vertex-wise corrective displacements to the skinning template in order to reduce the artifacts of linear blend skinning.

3.2. Dense Correspondences

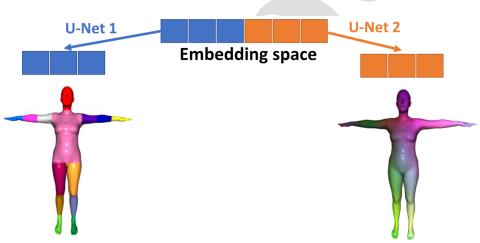


Fig. 2. Visualization of the six-dimensional template geometry embedding that we use to build pixel-to-vertex correspondences. It is a combination of a fifteen-part segmentation labeled with three-dimensional normalized color (left), and three normalized canonical vertex coordinates (right).

This section introduces the DL dense correspondence stage of our method: given a depth image and a template geometry mesh, a convolutional neural network predicts a dense mapping between the depth pixels and the SMPL template vertices. Mapping is obtained through the combination of a body part segmentation map and a pixel-to-vertex correspondence map.

Template Geometry Embedding. Our goal is to 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. As it is computationally expensive to learn a mapping from pixels to the entire span of the 6890 template vertices, we embedded the template geometry in a low dimensional space. We note this fixed embedding $E : \mathcal{T} \rightarrow [\![0, 1]\!]^6$. Building a reliable embedding E is crucial for our method, especially because neural networks can infer erroneous correspondences: switching body limbs due to the inherent symmetry of the human body, or confusing pixels belonging to adjacent body parts. E aims at mapping pixels to the template geometry, but images are 2D projections of the 3D world, this embedding must capture the underlying 3D geometry of the human shape under arbitrary poses and viewing angles. We started by defining the first 3 components of the embedding as the normalized 3 spatial coordinates of the template mesh in the canonical T-pose (see Fig. 2), by mapping the 3 normalized vertex coordinates to RGB values. As we found this representation insufficient to distinguish vertices in our experiments, we added 3 extra dimensions to our embedding to help us distinguish body parts more robustly. Hence, we divided the template geometry into 15 parts, including a background class, as shown in Fig. 2. We picked 15 distinctive RGB colors for each class, which represents the extra 3 coordinates of the embedding E. Next, we trained a deep neural network *slm* to map pixels to the template geometry embedding space: $slm : \Gamma \rightarrow E(\mathcal{T})$.

Neural Network. We stacked two U-Net [21] networks (image-to-image architecture) as illustrated in Fig. 1 to predict body part segmentation, and to regress normalized mesh colors.

These two outputs are concatenated to generate the pixel embedding values $slm(i) \in E(\mathcal{T})$. The first U-Net, called U-Net1 in Fig. 2, aims at segmenting the input depth image into 15 classes, one of which is the background. This segmentation label corresponds to the last three components of the embedding for each depth pixel. This embedding is then concatenated with the depth image and fed to the second U-Net, called U-Net2 in Fig. 2, to predict a 3-channel image corresponding to the 3 first components of the embedding. The network was trained using the combination of a cross-entropy loss on the output of the segmentation branch, and an L_2 loss on the output of the normalized color regression branch.

Pixel-to-Vertex Correspondence. To obtain correspondences v_c for a given depth image to the template geometry, we first map the image pixels to the low dimensional embedding using our convolutional neural network inference *slm*. The vertex j matching pixel i is then defined as the nearest template vertex in the embedding space, which writes:

$$v_c(i) = \arg\min_{j \in \mathcal{T}} ||slm(i) - E(j)||_2^2 \quad (2)$$

3.3. Model Fitting

In this section, we introduce the model-fitting stage of our method. Given an input depth image and pixel-to-vertex correspondences obtained from the previous stage, we fit the SMPL model to the depth image to recover the human shape and pose parameters of the adapted template mesh. To this end, we minimize the following objective function:

$$E(\theta, \beta, \gamma) = \lambda_D E_D(\theta, \beta, \gamma) + \lambda_\theta E_\theta(\theta) + \lambda_\beta E_\beta(\beta). \quad (3)$$

where E_D is the data term. The data term stands for minimizing a L_2 distance between pixel i's 3D point p_i (obtained using the intrinsic matrix and the pixel's depth value), and the corresponding vertex $v_c(i)$. This distance is summed over all pixels that belong to the body region $\Omega \subset \Gamma$ in the segmentation map:

$$E_D(\theta, \beta, \gamma) = \frac{1}{|\Omega|} \sum_{p_i \in \Omega} \rho(||p_i - v_c(i)||_2^2), \quad (4)$$

where $|\Omega|$ is the total number of pixels in Ω. Following previous work [5, 36], 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 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 $\lambda_D, \lambda_\theta, \lambda_\beta$ are trade-off weights of the objective function terms.

4. Comparison to previous works based on pre-segmented depth images

Prior to delving into the study of our approach in section 5, we would like to briefly recall the conclusions presented in [20], especially the comparison to previous works, and the ablation study. For a fair comparison with the state of the art, we carried out these evaluations on segmented simulated depth images without any background.

4.1. Benchmark and evaluation metrics

We conducted experiments on standard datasets of 3D human shapes with close-fitting clothes. Following protocols introduced in previous works [19, 18], we used the SURREAL [1], DFAUST [2] and also a subset of the AMASS [3] dataset entitled DanseDB[1]. To simulate depth images of the same resolutions, we rendered the single human 3D models contained in these datasets, placed at the scene's center and using the same resolutions with the same camera pose setting for all. Let us notice here that these 3D scenes did not contain any background. As a result, we have depth images for which the ground truth of the pose and shape of the person is known, since their models are the ones initially used in the rendering process. More information about the datasets is given in [20].

For fair comparison to [19, 18], the quality of our reconstruction method was assessed using the Mean Average Vertex Error in millimeters (*mm*), averaged subsequently overall testing frames:

$$\epsilon = \frac{1}{N} \sum_{i=1}^{N} \sqrt{||v_i - \hat{v}_i||_2^2}, \quad (5)$$

where vertices $\{v_i\}$ are the prediction, $\{\hat{v}_i\}$ are the ground truth, and N is the total number of vertices.

[1] http://dancedb.eu/

4.2. Comparison to the State-of-the-art

In Table 1, we compared our proposed approach to state-of-the-art methods that predict the human shape and pose from a single RGB or depth image. Using a similar evaluation protocol as in [18] (based on synthetic images), we replicated the reported performance of methods [4, 5, 9, 7, 50, 37]. The model-fitting method [4] deformed the SMPL model to the depth images using naive correspondences between the template and the input depth pixels. Kanazawa et al. [5] first detected 2D body joints from an RGB image and then fitted the SMPL model to the detected joints. Lassner et al. [38] and Kanazawa et al. [7] inferred SMPL parameters directly from RGB images. Wei et al. [9] built point correspondences by matching learned feature descriptors for pixels in the depth images. The 3D models were then generated by fitting the SMPL model to point correspondences. The rest of the methods in Table 1 directly inferred 3D meshes from RGB [37] or depth images [19, 18]. These two last methods based on depth images also used temporal information, which helps to reconstruct missing information and to obtain consistent shape along time, and continuity of the motion.

Methods	Input	SURREAL	DFAUST	DanseDB
Model fitting [4]	D	140.6	110.1	-
Lassner at al. [38]	RGB	155.5	-	-
Bogo et al. [5]	RGB	56.1	57.5	-
Wei et al. [9]	D	58.6	62.2	-
Kanaz. et al. [7]	RGB	54.3	58.1	-
Kanaz. et al. [50]	RGB	52.7	56.1	-
Kolot. et al. [37]	RGB	49.5	52.2	-
Wang et al. [18]	D + t	18.2	19.7	-
Jiang et al. [19]	D + t	15.5	8.1	-
Ours	D	**49.6**	**53.6**	**55.0**

Table 1. Comparison to methods predicting the mesh of a body with close-fitting clothes from a synthetic monocular RGB or depth (D) image, with time (+ t) or not, in terms of reconstruction errors (*mm*). For clarity, the error rates close, above and below 10% of our solution have been written respectively in blue, orange, and green.

Our method has similar results to most previous works, except for the two papers combining both depth data and temporal information [19, 18] noted (D+t). Model fitting approaches [4] based on depth images struggle to obtain good results due to the difficulty of getting a good initialization for the optimization using merely naive initialization heuristics.

Results reported in [20] showed that occlusions may lead to more important reconstruction errors, especially when the extremities of the body are hidden. Occlusions occurring to intermediate body parts can be more easily fixed by using the knowledge available for the previous and next body parts. For example, an occlusion of the hand leads to bigger errors compared to a forearm occlusion where the arm and the hand are visible. This reconstruction error may increase up to more than 20*cm* in some extreme cases.

Ablative analysis in [20] enabled us to evaluate the separated performance of each component of this hybrid approach. The results showed that the pose and shape reconstruction was enhanced compared to the reference model fitting method (even when using ground truth joints). Compared to using dense correspondence only, we decreased the reconstruction error from 59.7*mm* to 49.6*mm* thanks to adding human body segmentation. It avoids some mismatches between human parts to keep up the dense correspondence algorithm. These results demonstrated that coupling dense correspondence with model fitting (44.3*mm* error with GT dense correspondence) outperforms methods based on model fitting only using sparse joint position knowledge (80.1*mm* error with GT 3D joints positions).

5. Detailed evaluation of the method

Complementary to our previous paper [20], we introduce in this section new experiments to further evaluate the relevance of this hybrid approach. Firstly, we evaluate the impact of the method chosen for the SMPL fitting based on a preliminary dense correspondence. Secondly, we tested the ability of the hybrid approach to segment the background in non-segmented depth images.

5.1. Strategy used to infer correspondence

One key component of our method is the mapping between the dense point cloud of the depth image and the template vertex with the color embedding. However, the topology of the point cloud does not correspond to the template color embedding, based on the 6,098 points SMPL model. Indeed, several

points in the point cloud can correspond to the same SMPL vertex. Consequently, we need to design a method to associate each SMPL vertex to a unique 3D point coordinate in the optimization. To achieve this goal, we tested various strategies.

For this specific study, we changed the test conditions, with a smaller dataset compared to the previous section, which enabled us to carry out more numerous tests without huge training time. In the DFAUST dataset, we selected 104 motion clips for training and 25 for testing. The viewpoint was selected to ensure a front view of the character in all the images. Indeed, with such a smaller training dataset, it may be more difficult to deal with a wide range of viewpoints. In this section, we focus attention on the method used to associate a 3D point to a template vertex, and do not consider the ability to generalize to a wide set of possible viewpoints. The training dataset was then composed of the 100 first frames of each randomly selected clip, with 4Hz sampling, leading resulting in 10,000 depth images. 4Hz is used to avoid selecting poses that are too close together in the motion clip. We performed the same approach for the testing dataset, leading to 125 depth images, with similar training parameters than those described in [20]: learning rate 0.0016 and 10 epochs.

The first strategy tested in this work, called S_1 consists in selecting, for each 3D point of the point cloud, the closest vertex of the template SMPL model, as shown in equation 2. As a result, one point of the point cloud is associated with one template vertex, but one template vertex may be associated with several points (the number of points is generally greater than the number of template vertices). Hence, in the model fitting phase, the system tends to find a compromise between all the possible 3D candidate points. Let us notice that this strategy does not take the segmentation information into account, so that a point with a color embedding associated with a forearm could be linked to an arm template vertex, if this is the closest one.

The second strategy S_2 is similar to S_1, except that each point is associated to the closest template vertex with the same body part label. This way, a 3D point with a forearm color embedding could not be associated with an arm template vertex, even if this one is the closest one.

Unlike S_1 and S_2, S_3 is searching for each template vertex the closest 3D points among the candidates. This time, each template vertex is consequently associated with one unique 3D point. Some 3D points may not be associated with a template vertex. Hence, during the fitting phase the system has to deal with one unique pair of vertex and 3D point, for each vertex, unlike S_1 and S_2.

The last strategy S_4 is similar to S_3, except that we use the average position of all the 3D points candidates, instead of the one with the minimal distance to the template vertex.

Strategy of correspondence	reconstruct error (*mm*)
S_1: Closest vertex	47.79
S_2: S1 + body part seg.	47.00
S_3: Select Minimal distance	43.70
S_4: Select Average position	**43.27**

Table 2. Different strategies of inferring correspondences. The results were obtained with 10,000 training and 125 tests for each strategy.

Whatever the strategy, the remaining of the optimization process is unchanged, only the correspondences change. The results are presented in Table 2. In this table, for all the potential strategies, the final reconstruction error (in mm) is given using the same method than in the previous sections. These results show that using S_1 or S_2 does not significantly change the performance of the algorithm. Further experiments would be necessary to analyze these results, but we could imagine that the segmentation phase avoids a few big confusions between body parts, but does not improve significantly the overall accuracy of the reconstruction. However, when selecting only one point, with the minimal distance S_3 or the average point S_4, the reconstruction error decreased down to 43.7*mm* and 43.27*mm* respectively.

5.2. Background Segmentation

In the previous section 4, we tested our method on pre-segmented depth images. However, the dense correspondence method is also designed to segment the background, by introducing a special color encoding for the background. To evaluate the performance of the background segmentation, we carried-out complementary tests.

Fig. 3. Examples of our synthetic dataset. For each scene, the left image stands for the background image and the right one depicts an example of the final depth input after adding the character depth image.

To this end, we designed new depth images composed of a single character (DFAUST dataset, as in the previous sections) and a synthetic background. We simulated indoor scenes from the self-service synthetic dataset-generating platform [51], which provides photo-realism synthetic images based on 3D scenes, with different 3D objects, lighting, camera position, etc. We selected two main scenes: the bedroom and the living room. Setting a virtual camera position and orientation in each 3D scene, we can compute the corresponding RGB and depth image. By randomly changing these two scenes parameters (camera position, 3D assets placed in the scene at various positions), we obtained a total of 1480 depth images for training, and 298 depth images for testing.

We then added the depth image of the character in front of the 3D scene depth map, to ensure that no object in the scene could hide body parts. In other words, the character was placed in front of the scene, to avoid occlusions with the environment, although the depth of the character may not be fully consistent with the one of the background. We have chosen this solution to evaluate the theoretical impact of the system to segment the background in an ideal case where the characters have a different depth values compared to the environment. With real noisy images, confusions with points sharing the same depth value should occur, which would significantly decrease the accuracy of the segmentation and shape reconstruction. Fig.3 shows some examples of adding background to 125 test images.

In this test, we used the strategy S_4 presented above to deal with the dense correspondence, as it provided us with the best results. We applied the same test conditions than those used to test S_1 to S_4 strategies: 10,000 depth images selected with 4Hz

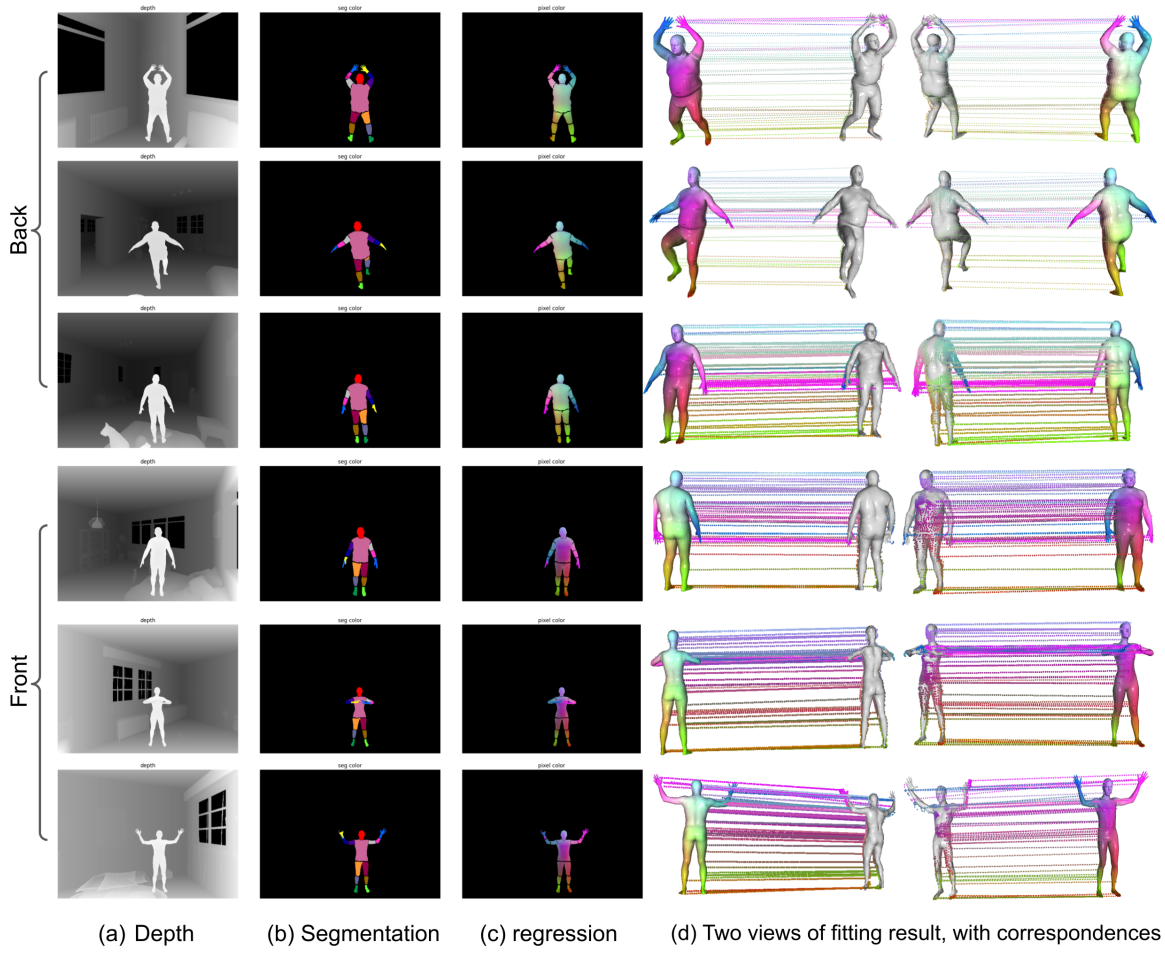


Fig. 4. Visualization of reconstruction results when adding a background. (a) Input depth image with a synthetic background; (b) Background (black) and body part segmentation; (c) Regressed template vertex color; (d) two views of the final reconstructed mesh with inferred dense correspondences. Each view contains the GT input mesh colored withe the template color embedding. The final reconstructed human body in lightgray is superimposed to this GT input mesh for comparison.

frequency among the clips, limited to front viewpoint of the character. As a result, the reconstruction error increased from 43.27*mm* to 53.24*mm* when adding a background.

Figure 4 shows some visual results of this test, at different stages of the method. In this figure, from left to right, one can see the depth input image including background (a), the human part segmentation with black color for the segmented background (b), the color embedding (c), and two different views of the final reconstructed pose (lightgray) and the GT human mesh (colored mesh using the template color embedding). In most of the cases, the background segmentation is fine. However, in some cases, points in body parts are badly labeled as background, such as the upper-limbs joints which are colored in black in the last bottom example. This type of segmentation error can explain the increase of reconstruction error after model fitting, when the input depth image includes a 3D background.

6. Conclusion

The main contribution of this paper is to explore the interest of combining model fitting and DL dense correspondence between depth pixels and human template vertices. The accuracy of the human pose and shape reconstruction from a single depth image, as shown in our results, demonstrated that thousands of correspondences used as inputs to the model fitting stage provide richer information than joint positions alone, confirming our hypothesis $H2$.

However, even a dense correspondence may contain errors due to the complexity of human pose, shape, symmetry, and camera viewpoint. In our approach, the body-part segmentation is assumed to also segment the background, but we only tested simple conditions where the depth of the feet is different from the depth of the ground. Our results partly support hypothesis $H1$, with a reconstruction error increasing by one centimeter approximately when adding such a background. In this test, we only considered background that cannot add partial external occlusions of the body. In more complex scenes, with objects hiding part of the body, and more challenging confusion in depth values between the ground and the feet, we could expect a significant decrease of performance. Further works are needed to evaluate this limitation more deeply.

This hybrid approach aims at getting the benefits of DL-based dense correspondence, and model fitting by optimization. The communication between these two methods relies on the dense correspondence encoding, but the number of available 3D points is different from the number of template vertices. We have shown that the strategy used to associate a template vertex and 3D points has a significant influence on the final human pose and shape reconstruction, after optimization. The best accuracy was obtained when selecting the average of the available points associated with the same template vertex, before model fitting.

As in many previous works, we tested our approach with simulated depth images to accurately control the test conditions, and the corresponding ground truth. However, dealing with real depth images provided by depth sensors raises many difficult constraints, such as segmenting the character from the background, denoising the images, dealing with occlusions and clothes, etc. In this paper, we focused on the segmentation problem, but further evaluation is needed to see the behavior of the approach when dealing with real depth images. Preliminary results on real RGBD images (see Figure 5) tend to show that the segmentation step is very sensitive to noise, with several background pixels that were incorrectly labeled as body parts. The resulting SMPL model leads to incorrect surface shape reconstruction, with large errors. Further analysis is required to accurately quantify this error, in relation to the actual effect of noise and partial occlusions. From this quantization, one or several noise models could be defined to reflect different camera brands. Including noisy depth images in the training phase of the dual U-Net segmentation algorithm could improve these results. However, this would require either properly modeling the noise in the real depth images, or collecting a large amount of real depth images with ground truth denoised images.

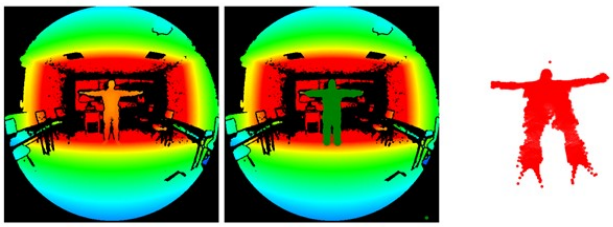


Fig. 5. Preliminary results obtained on real depth images provided by a Microsoft Kinect 4 Azure camera (left). The resulting reconstructed SMPL model is depicted in the middle image in green, and the segmented image used for the reconstruction is depicted in red on the right.

Although we have shown the interest of combining DL and model fitting, recent works using temporal information have shown very impressive results. This suggests that working on a single RGB or depth image is limited, and that the integration of temporal information needs to be explored in future work to determine whether the combination of DL and model fitting is still advantageous. Other recent works explored encoding an implicit function of the human body surface based on sparse 3D points [47, 48, 49], with impressive results to jointly represent body pose, shape and clothing geometry. More extensive evaluation is needed to evaluate the sensibility of these methods to noise, occlusion and complex background.

References

[1] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, C. Schmid, Learning from synthetic humans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 109–117.

[2] F. Bogo, J. Romero, G. Pons-Moll, M. J. Black, Dynamic faust: Registering human bodies in motion, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6233–6242.

[3] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, M. J. Black, AMASS: Archive of motion capture as surface shapes, in: International Conference on Computer Vision, 2019, pp. 5442–5451.

[4] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black, SMPL: A skinned multi-person linear model, ACM Trans. Graphics (Proc. SIGGRAPH Asia) 34 (6) (2015) 248:1–248:16.

[5] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M. J. Black, Keep it smpl: Automatic estimation of 3d human pose and shape from a single image, in: European conference on computer vision, Springer, 2016, pp. 561–578.

[6] N. Kolotouros, G. Pavlakos, M. J. Black, K. Daniilidis, Learning to reconstruct 3d human pose and shape via model-fitting in the loop, in: ICCV, 2019.

[7] A. Kanazawa, M. J. Black, D. W. Jacobs, J. Malik, End-to-end recovery of human shape and pose, in: Computer Vision and Pattern Recognition (CVPR), 2018.

[8] R. A. Güler, N. Neverova, I. Kokkinos, Densepose: Dense human pose estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7297–7306.

[9] L. Wei, Q. Huang, D. Ceylan, E. Vouga, H. Li, Dense human body correspondences using convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] F. Tan, D. Tang, M. Dou, K. Guo, R. Pandey, C. Keskin, R. Du, D. Sun, S. Bouaziz, S. Fanello, et al., Humangps: Geodesic preserving feature for dense human correspondences, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1820–1830.

[11] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, et al., Efficient human pose estimation from single depth images, IEEE transactions on pattern analysis and machine intelligence 35 (12) (2012) 2821–2840.

[12] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, CVPR 2011 (2011) 1297–1304.

[13] R. Bashirov, A. Ianina, K. Iskakov, Y. Kononenko, V. Strizhkova, V. Lempitsky, A. Vakhitov, Real-time rgbd-based extended body pose estimation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2807–2816.

[14] F. Bogo, M. J. Black, M. Loper, J. Romero, Detailed full-body reconstructions of moving people from monocular RGB-D sequences, in: International Conference on Computer Vision (ICCV), 2015, pp. 2300–2308.

[15] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, C. Theobalt, A data-driven approach for real-time full body pose reconstruction from a depth camera, in: 2011 International Conference on Computer Vision, 2011, pp. 1092–1099. doi:10.1109/ICCV.2011.6126356.

[16] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, Y. Liu, Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor, in: The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR), IEEE, 2018.

[17] K. Guo, F. Xu, Y. Wang, Y. Liu, Q. Dai, Robust non-rigid motion tracking and surface reconstruction using l0 regularization, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3083–3091. doi:10.1109/ICCV.2015.353.

[18] K. Wang, J. Xie, G. Zhang, L. Liu, J. Yang, 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.

[19] H. Jiang, J. Cai, J. Zheng, Skeleton-aware 3d human shape reconstruction from point clouds, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5431–5441.

[20] X. Wang, A. Boukhayma, S. Prevost, E. Desjardin, C. Loscos, F. Multon, Coupling dense point cloud correspondence and template model fitting for 3d human pose and shape reconstruction from a single depth image, in: 2022 International Conference on Interactive Media, Smart Systems and Emerging Technologies (IMET), 2022, pp. 01–08. doi:10.1109/IMET54801.2022.9929833.

[21] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[22] F. Mueller, M. Davis, F. Bernard, O. Sotnychenko, M. Verschoor, M. A. Otaduy, D. Casas, C. Theobalt, Real-time pose and shape reconstruction of two interacting hands with a single depth camera, ACM Trans. Graph. 38 (4). doi:10.1145/3306346.3322958.
URL https://doi.org/10.1145/3306346.3322958

[23] J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. A. Otaduy, D. Casas, C. Theobalt, Rgb2hands: Real-time tracking of 3d hand interactions from monocular rgb video, ACM Trans. Graph. 39 (6). doi:10.1145/3414685.3417852.
URL https://doi.org/10.1145/3414685.3417852

[24] T. B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Computer vision and image understanding 104 (2-3) (2006) 90–126.

[25] Y. Chen, Y. Tian, M. He, Monocular human pose estimation: A survey of deep learning-based methods, Computer Vision and Image Understanding 192 (2020) 102897.

[26] I. Jegham, A. B. Khalifa, I. Alouani, M. A. Mahjoub, Vision-based human action recognition: An overview and real world challenges, Forensic Science International: Digital Investigation 32 (2020) 200901.

[27] M. Ye, Y. Shen, C. Du, Z. Pan, R. Yang, Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2016) 1517–1532.

[28] G. Mishra, S. Saini, K. Varanasi, P. J. Narayanan, Human shape capture and tracking at home, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 390–399. doi:10.1109/WACV.2018.00049.

[29] Q. Zhang, B. Fu, M. Ye, R. Yang, Quality dynamic human body modeling using a single low-cost depth camera, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 676–683. doi:10.1109/CVPR.2014.92.

[30] A. Weiss, D. Hirshberg, M. J. Black, Home 3d body scans from noisy image and range data, in: 2011 International Conference on Computer Vision, 2011, pp. 1951–1958. doi:10.1109/ICCV.2011.6126465.

[31] R. A. Newcombe, D. Fox, S. M. Seitz, Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 343–352.

[32] Z. Zhang, L. Hu, X. Deng, S. Xia, Weakly supervised adversarial learning for 3d human pose estimation from point clouds, IEEE Transactions on Visualization and Computer Graphics 26 (5) (2020) 1851–1859. doi:10.1109/TVCG.2020.2973076.

[33] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, C. Theobalt, XNect: Real-time multi-person 3D motion capture with a single RGB camera, Vol. 39, 2020. doi:10.1145/3386569.3392410.
URL http://gvv.mpi-inf.mpg.de/projects/XNect/

[34] H. Choi, G. Moon, K. M. Lee, Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose, in: European Conference on Computer Vision (ECCV), 2020.

[35] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, M. J. Black, Monocular expressive body regression through body-driven attention, in: European Conference on Computer Vision, Springer, 2020, pp. 20–40.

[36] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, M. J. Black, Expressive body capture: 3d hands, face, and body from a single image, in: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.

[37] N. Kolotouros, G. Pavlakos, K. Daniilidis, Convolutional mesh regression for single-image human shape reconstruction, in: CVPR, 2019.

[38] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, P. V. Gehler, Unite the people: Closing the loop between 3d and 2d human representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
URL http://up.is.tuebingen.mpg.de

[39] G. Moon, K. M. Lee, I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image, in: European Conference on Computer Vision (ECCV), 2020.

[40] K. Lin, L. Wang, Z. Liu, End-to-end human pose and mesh reconstruction with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1954–1963.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[42] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, J. Davis, Scape: shape completion and animation of people, ACM Trans. Graph 24 (2005) 408–416.

[43] X. Huang, H. Yang, E. Vouga, Q. Huang, Dense correspondences between human bodies via learning transformation synchronization on

graphs, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 17489–17501.
URL https://proceedings.neurips.cc/paper/2020/file/ca7be8306ecc3f5fa30ff2c41e64fa7b-Paper.pdf

[44] T. L. Zhu, P. Karlsson, C. Bregler, Simpose: Effectively learning densepose and surface normals of people from simulated data, in: ECCV, 2020.

[45] R. Litman, A. Bronstein, Learning spectral descriptors for deformable shape correspondence, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (1) (2014) 171–180. doi:10.1109/TPAMI.2013.148.

[46] C. R. Qi, L. Yi, H. Su, L. J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 5105–5114.

[47] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, G. Pons-Moll, Combining implicit function learning and parametric models for 3d human reconstruction, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 311–329.

[48] E. Corona, A. Pumarola, G. Alenya, G. Pons-Moll, F. Moreno-Noguer, Smplicit: Topology-aware generative model for clothed people, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11875–11885.

[49] Z. Dong, C. Guo, J. Song, X. Chen, A. Geiger, O. Hilliges, Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20470–20480.

[50] A. Kanazawa, J. Y. Zhang, P. Felsen, J. Malik, Learning 3d human dynamics from video, in: Computer Vision and Pattern Recognition (CVPR), 2019.

[51] Ai verse, https://www.ai-verse.com/.

Computers & Graphics (2023)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Evaluation of hybrid deep learning and optimization method for 3D human pose and shape reconstruction in simulated depth images

Xiaofang Wang[a,b], Stéphanie Prévost[b], Adnane Boukhayma[c], Eric Desjardin[b], Céline Loscos[b], Benoit Morisset[a], Franck Multon[c,d]

[a]*AI Verse*
[b]*University of Reims Champagne Ardenne, Reims, France*
[c]*Inria, Univ. Rennes, CNRS, IRISA, Rennes, France*
[d]*Univ. Rennes, M2S, Rennes, France*

ARTICLE INFO

Article history:
Received February 10, 2023

Keywords: Human motion capture, shape reconstruction, deep learning, computer vision, depth sensor

ABSTRACT

In this paper, we address the problem of capturing both the shape and the pose of a human character using a single depth sensor. Some previous works proposed to fit a parametric generic human template into the depth image, while others developed deep learning (DL) approaches to find the correspondence between depth pixels and vertices of the template. We designed a hybrid approach, combining the advantages of both methods, and conducted extensive experiments on the SURREAL [1], DFAUST datasets [2] and a subset of AMASS [3]. Results show that this hybrid approach enables us to enhance pose and shape estimation compared to using DL or model fitting separately. We also evaluated the ability of the DL-based dense correspondence method to segment also the background - not only the body parts. We also evaluated 4 different methods to perform the model fitting based on a dense correspondence, where the number of available 3D points differs from the number of corresponding template vertices. These two results enabled us to better understand how to combine DL and model fitting, and the potential limits of this approach to deal with real depth images. Future works could explore the potential of taking temporal information into account, which has proven to increase the accuracy of pose and shape reconstruction based on a unique depth or RGB image.

Acknowledgements

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

1. Introduction

Reconstructing the pose and shape of a human character using a single camera is of great interest in applications where the user is represented by a realtime avatar, such as in immersive social media or mixed reality videogames.

Several approaches proposed to reconstruct the 3D shape and pose of a human character using a single RGB image, by fitting parametric models (such as SMPL [4], SMPLify [5]) to the RGB information, or by directly learning parameters of parametric models [6, 7]. Previous works demonstrated that Deep Learning (DL) was promising to learn the correspondence

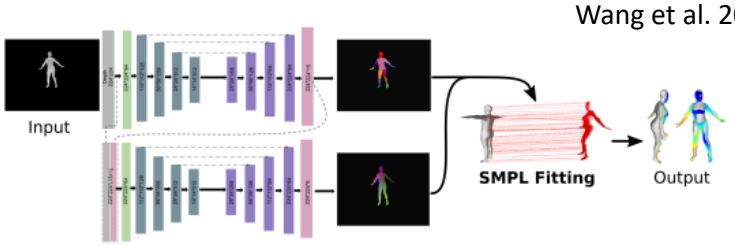
Highlights

- Dense correspondence and model fitting enhance 3D shape and pose reconstruction
- Bodypart and background color embedding enables character segmentation in depth map
- Averaging 3D candidate points before model fitting offers the best reconstruction

Evaluation of hybrid deep learning and optimization method for 3D human pose and shape reconstruction in simulated depth images

Wang et al. 2021

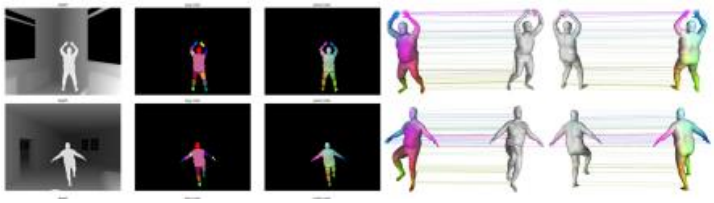
Hybrid method tested



Double U-Net: pixel-to-vertex correspondence

Optimization

Tests on simulated images with synthetic background



Conclusion:

1) Depth image background segmentation is performed jointly with the dense correspondence
2) Using dense correspondence as an input of the model fitting algorithm improves the performance of pose and shape reconstruction
3) Future direction using temporal information and robustness to noisy real images

CRediT author statement for the paper “Evaluation of hybrid deep learning and optimization method for 3D human pose and shape reconstruction in simulated depth images”

Xiaofang Wang: Conceptualization, Methodology, Software

Stéphanie Prévost: Conceptualization, Methodology, Writing

Adnane Boukhayma: Conceptualization, Methodology, Writing

Eric Desjardin: Conceptualization, Methodology, Writing

Céline Loscos: Conceptualization, Methodology, Writing, Supervision

Benoit Morisset: Data curation

Franck Multon: Conceptualization, Methodology, Writing, Supervision

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: