Combining Human Body Shape And Pose Estimation For Robust Upper Body Tracking Using a Depth Sensor

Thomas Probst, Andrea Fossati and Luc Van Gool

Computer Vision Lab,
ETH Zurich
{probstt,fossati,vangool}@vision.ee.ethz.ch

Abstract. Rapid and accurate estimation of a person's upper body shape and real-time tracking of the pose in the presence of occlusions is crucial for many future assistive technologies, health care applications and telemedicine systems. We propose to tackle this challenging problem by combining data-driven and generative methods for both body shape and pose estimation. Our strategy comprises a subspace-based method to predict body shape directly from a single depth map input, and a random forest regression approach to obtain a sound initialization for pose estimation of the upper body. We propose a model-fitting strategy in order to refine the estimated body shape and to exploit body shape information for improving pose accuracy. During tracking, we feed refinement results back into the forest-based joint position regressor to stabilize and accelerate pose estimation over time. Our tracking framework is designed to cope with viewpoint limitations and occlusions due to dynamic objects.

Keywords: human pose estimation, human body shape, pose tracking, model fitting, real-time, occlusion handling, random forest, subspace

1 Introduction

Automatic perception of human subjects will play a key role in assistive technology and medical applications. For instance, systems for medical imaging, treatment planning, radiation therapy, interventional imaging and virtual reality benefit from a precise recognition of the patient in his/her distinct pose [1]. In general, potential advantages are more accurate interactions, compliant systems, alleviation for users and reduced costs. Future applications of computer-assisted surgery and telemedicine will provide even physical interaction with the patient by means of teleoperation techniques, and therefore rely crucially on a good localization and tracking of the subject.

Since the availability of low-cost depth sensors, real-time pose estimation has advanced fast and is present in many commercial videogame consoles [2]. However, the requirements regarding body poses, occlusions, field of view, body

shape fidelity and accuracy are quite different compared to professional and medical applications. In our work we investigate depth-based sensors in the scenario of tracking the torso of a person in lie-down poses. Our goal is to accurately estimate the upper body pose and surface of the subject in the presence of severe occlusions due to viewpoint limitations and potential occluding objects in front of the sensor.

In particular, we are involved in the *ReMeDi* research project that aims to develop a telediagnosis system: This will allow doctors to remotely perform physical and ultrasonography examinations by teleoperating a multifunctional robotic device at the patient side. Potential advantages are the provision of sparsely populated areas, enhanced availability of expert knowledge, more beneficial time schedules and reduced health care costs. In excess of teleconferencing, haptic interfaces, force-feedback and multisensory data representing the remote environment provide proactive support for the doctor. One goal of the project is to mimic the real examination process for the doctor as close as possible. In order to provide an intuitive and safe way of interaction with the patient, the robotic device has to perform certain tasks autonomously. Computer Vision methods serve the critical need of perceiving the patient in his/her distinct pose in order to estimate the position of the end effector with respect to the body. A Kinect sensor mounted on the robot's head is providing real-time depth data during the examination, while the patient is lying on a bed and the robot arm moving in front of the sensor. Our focus therefore is to accurately estimate the surface of the patient's upper body in order to determine the position of the examination probe relative to the torso. This knowledge can subsequently be used to map probe measurements to a human body model, providing an intuitive way of storing, visualizing and comparing examination results. The efficiency and usability of the teleoperation system directly depends on the speed and accuracy of this estimation process.

In general, there are discriminative (fast, lower accuracy) and generative (expensive, higher accuracy, but prone to local minima) approaches to both body shape and pose estimation problems. While we extend well known random forests for pose estimation [2–4], we propose to tackle prediction of body shape parameters from a single depth image by means of a linear subspace representation. To improve tracking accuracy and exploit shape information on top of our discriminative approach, we combine the two paradigms by subsequently performing additional model-fitting based refinement iterations.

The remainder of the paper is organized as follows: We provide an overview of the related work in Section 2. In Section 3, we present our framework for body shape estimation and pose tracking. Then we report quantitative and qualitative results of the proposed methods in Section 4, and conclude by discussing limitations and future work in Section 5.

2 Related Work

In is this section we relate our method to body shape and pose estimation approaches that have been proposed in the literature.

Body shape estimation from depth. Body shape estimation work is dominated by generative model-fitting approaches. Fitting a human body model to the observed depth data is very robust to noise. In this context, the SCAPE model [5] is the most popular one, but other models [6–9] have been investigated, or could potentially be used for this purpose. For instance, [10] and [11] follow a procedure based on the Iterative Closest Point (ICP) algorithm to fit a SCAPE model to observed point clouds. In [12] and [13], the SCAPE model is employed to obtain the human body shape under the garments worn by the subject. Weiss et al. [14] optimize the SCAPE parameters to jointly maximize the overlap of the projected model with the RGB-silhouette and minimize the distance between corresponding points on the model and on an input range image. Bogo et al. [15] propose a coarse-to-fine model fitting strategy.

Among the model-free approaches, methods similar to KinectFusion [16, 17] perform 3D body reconstructions from RGB-D sequences or multiple views [18, 19, 11, 20, 21].

By contrast, we obtain an initial estimate of low dimensional body shape parameters (using a variant of the SCAPE body model) from a *single frame*. To this end, we propose to exploit depth image subspace features by training a regression forest to predict body shape parameters. Then we apply an ICP-based model fitting algorithm to refine the estimation and improve accuracy.

Human pose estimation from depth. In general, generative models attempt to fit an articulated body model to the observed data by finding 3D- contour- or silhouette-based correspondences with an ICP-like approach [22, 23, 14, 24]. Discriminative methods however try to directly infer pose information in a data-driven manner. Depth difference features as introduced by [2] enable training of random forest models that are capable of real-time inference [3, 4, 25]. Many hybrid approaches were proposed to combine the benefits of both worlds [26, 27, 10, 28] by using database look-up to obtain a good initialization. In the same sense, [29, 30] employ random forests to predict dense correspondences for a subsequent iterative model fitting procedure. Our work specializes the approach of [3] by improving the accuracy of the joint position estimation for upper body joints using a global forest refinement strategy. We further improve the upper body localization by fitting the estimated body shape to the observed data. We assume the torso shape to change rigidly with pose and therefore use an efficient rigid ICP-based alignment. To improve stability and robustness over time, the refined joint positions are fed back to the joint position estimation.

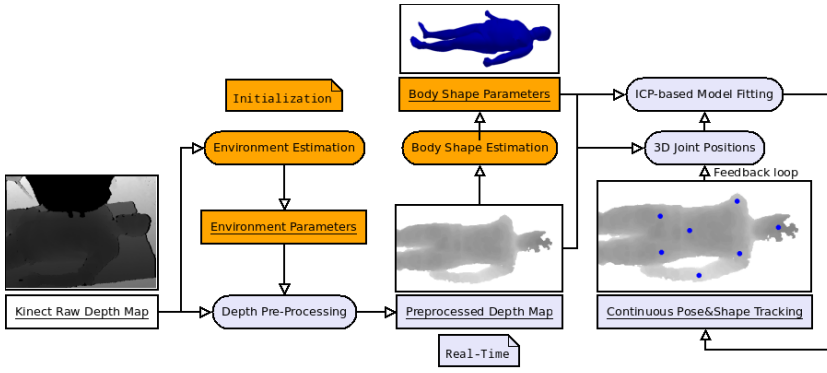


Fig. 1. Pipeline overview: At the initialization stage (orange) we estimate the environment and the patient's body shape. During the real-time phase (blue), we track the patient using the previously estimated parameters. To this end, we perform discriminative body pose estimation in combination with a model fitting procedure and introduce robustness and temporal consistency by a feedback loop.

3 Method

We now introduce our framework for human body shape estimation and upper body tracking. We first propose to make use of a subspace representation of canonical depth maps to predict a set of human body shape parameters. Second, we extend the random forest framework by Girshick et al. [3] to accurately estimate upper body joint positions. We then compute a coarse alignment using the predicted joint positions and initialize a model-fitting based refinement with our estimated upper body shape model. The whole pipeline is illustrated in Figure 1.

3.1 Preprocessing

As a first step, we estimate the 3D environment of the scene. In our medical scenario, we estimate the ground plane and the position and height of the patient's bed to define a volume of interest. This enables efficient pre-processing (planar clipping) of the depth images during the real-time phase. During the initialization phase, we require the patient to lie on the bed and the scene to be free of external objects.

3.2 Body shape estimation via subspace regression

The motivation for our approach is the observation that model-fitting procedures require a good initialization in order to converge to the correct minimum. As we show in our experiments, this is a serious disadvantage in terms of accuracy and run time. Therefore we propose a method to obtain a sound data-driven initialization to jump-start the model-fitting and to guide the algorithm towards the correct solution.

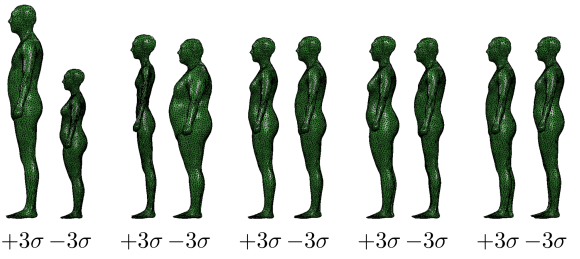


Fig. 2. Body shape variations captured by the first 5 PCA coefficients (from [7]). The renderings depict the deviation from the mean body shape resulting from setting each of the coefficients to $\pm 3\sigma$.

Our approach to rapidly obtaining an estimate of the body shape from a single frame is partly inspired by [27]. First, a normalized view of the subject is created by means of discriminative pose estimation. Then we compute a linear subspace of all canonical depth images to reduce complexity. The subspace coefficients ultimately serve as features for a random regression forest to predict a set of body shape parameters.

Body shape model In particular, we employ the MPII Human Shape model [7]. The shape parameters $\mathbf{s}$ of this statistical shape model represent a small number of directions with high variance from the mean body shape (principal components). They have been estimated using principal component analysis (PCA) on a varied dataset of aligned 3D scans of the human body. Figure 2 visualizes the variations captured by the first 5 shape parameters. For instance, the first component captures variations due to body height, while the second component $\mathbf{s}^1$ is dominantly influenced by the torso shape. This model allows to compactly represent body shapes and to generate corresponding mesh models for fitting purposes.

Canonical view. The goal of this step is to normalize the depth camera's point of view w.r.t. the subject. To this end, we predict the joint positions of the subject using the method described in Section 3.3. We define a coordinate frame by a subset of the upper body joints, and define a canonical viewpoint as a rigid transformation originating from this frame. In particular, we choose to center the hip in the image and set the viewpoint 2 m in front of it. Choosing the distance is a trade-off between effective resolution, field-of-view size and robustness towards misalignment. Then we project the point cloud onto a virtual camera in this canonical pose. By contrast, the authors of [27] use the centroid and the principal directions of the point cloud to perform normalization, which is very problematic in the presence of (self-)occlusions.

We assume the point cloud data to be dense, and the range of rotation angles to be limited to roughly frontal/side views. The transformation to the canonical view can therefore be approximated with efficient point-wise rigid transformations and projections into the 2D virtual camera frame, without reconstructing the 3D surface. After applying the normalization, we obtain a training set of canonical depth images, showing different subjects in various poses from the same viewpoint and distance.

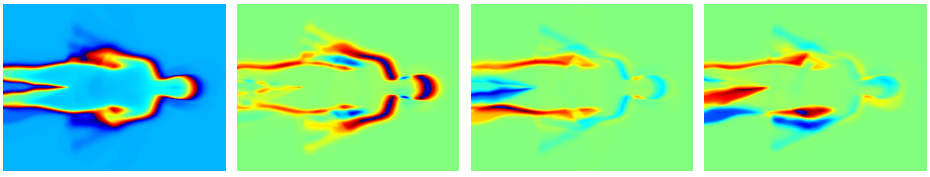


Fig. 3. Visualization of the mean canonical depth image and the first three eigen-images obtained by PCA. Note that the first eigen-image is sensitive to the global body size while the second dominantly captures variances in the lower body width. The third principal direction however seems to be mostly related to pose differences.

Learning. Intuitively, the appearance of the canonical depth images contains information about the body shape. To reduce complexity and exploit correlations between pixel locations, we compute a linear subspace using principal component analysis. To efficiently handle a big set of image data with high dimensionality, we use an approximation based on randomized projections.[1] We keep the N_d orthonormal directions (eigen-images) that contain the most variance in the data (highest eigenvalues). This yields a low-dimensional subspace representation of a depth image, denoted as a coefficient vector $\mathbf{c}$. Figure 3 visualizes the variances captured by the first three eigen-images. Note that variations are not only caused by shape, but also pose and misalignment noise are captured by the subspace coefficients. We therefore assume the relationship between the image subspace and the shape parameters to be nonlinear in general. Learning a dictionary like Ye et al. [27] resulted in bad generalization in our case. In order to learn which combination of coefficients is correlated with body shape, we employ a random forest model. We train an ensemble of regression forests to predict each of the body shape parameters in $\mathbf{s}$, using $\mathbf{c}$ as features. Note that we perform the PCA on rendered canonical images (without normalization artifacts), whereas normalized depth images are projected to the subspace (see next paragraph) and serve as input to the forest training.

Inference for a new image. At test time, we first estimate the set of 3D upper body joint positions (see Section 3.3) necessary to generate the canonical view. The task now is to find the most plausible subspace coefficients $\mathbf{c} \in \mathbb{R}^{N_d}$ explaining the observed canonical depth image. Due to the presence of occlusions and artifacts introduced by our normalization, considering all pixels would deteriorate the results. Instead of projecting the canonical depth map by performing the inner products with the eigen-images $P \in \mathbb{R}^{N_{\text{pixels}} \times N_d}$, we propose to minimize the squared reconstruction error only on visible points:

$$E(\mathbf{c}|I) = \frac{1}{2}(I_{\text{visible}} - \Pi(P\mathbf{c}))^2 \,. \tag{1}$$

$I_{\text{visible}} \in \mathbb{R}^{N_v}$ denotes the stacked intensity (depth) vector of all N_v visible pixels and $\mathbf{c}$ is the vector of subspace coefficients, while $\Pi(\cdot)$ selects and orders

[1] RedSVD implementation. https://code.google.com/archive/p/redsvd/

the reconstructed pixels according to the order in I_{visible}. We solve this convex minimization by performing gradient descent.[2] Using the optimal subspace coefficients $\mathbf{c}$ as features, the body shape parameters are finally predicted by the trained regression forest.

3.3 Joint position estimation

Our work on discriminative pose estimation is based on the efficient method by Girshick et al. [3]. The idea is to directly regress votes for predicting 3D joint positions from the depth map. First, a set of points is randomly sampled from the image. Then a random forest is trained using local depth features around each sample point. The regression target value is a 3D offset vector pointing from a sample point to the position of the body joint. The employed depth-difference features are invariant to translation and depth of the visible subject. We account for different viewpoint angles and body shapes by using high variety training data.

Girshick et al. [3] (re-)use the same forest structure (trained originally for body part classification) by dropping down samples and computing leaf statistics for the different joints using mean-shift. Similar to tree growing, this procedure follows a rather greedy strategy and does not obey a global cost function.

Multipurpose forest refinement. We therefore extend the forest refinement strategy by Ren et al. [31] to serve two goals: First, we aim to improve joint prediction accuracy for upper body joints by redistributing the prediction errors more favorably. Second, we exploit the refinement to make the forest structure reusable for multiple joints.

In the following formulation, every dimension of the 3D output vector is treated independently. Formally, let $\mathcal{P}_I = \{(\mathbf{x}_i^I, y_i^I)\}$ represent the set of samples belonging to image I, where y_i^I denotes one dimension of the target 3D offset. We model the refinement process as that of learning leaf weights $\mathbf{w}$ generating a per-sample offset prediction of the form

$$\hat{y}_i^I(\mathbf{w}|\mathbf{x}_i^I) = \mathbf{w}^T \phi(\mathbf{x}_i^I) \; . \tag{2}$$

By $\phi(\mathbf{x}_i)$ we denote the binary vector whose j^{th} position $\phi_j(\mathbf{x}_i)$ has value 1 if the local sample i has reached the corresponding leaf in the forest, and 0 otherwise[3]. Our goal of learning the optimal prediction value of each leaf can be thought of as a linear regression problem on the leaf weights $\mathbf{w}$, using the leaf indices corresponding to each sample as categorical features.

However, we found that directly applying the method of [31] worsens the overall results. This stems from the fact that the predictions are treated independently for each sample: the optimization overfits to sets of samples from

[2] Note that solving for $\mathbf{c}$ in equation $I(\mathbf{c}) = \Pi(P\mathbf{c})$ is overdetermined as long as there are more than N_d visible points/pixels. If all pixels are taken into account, the result of the optimization equals projecting the canonical depth map on the eigenimages.

[3] Note that this vector concatenates the leaves of all the trees in the forest.

images that are easy to predict, ignoring those samples which would improve prediction on difficult images.

Hence we extend the approach of [31], and account for combined predictions of the entire image, by defining an image-level prediction as the mean of all votes for the absolute joint position:

$$\bar{y}^I(\mathbf{w}) = \frac{1}{N_s} \sum_{\{\mathbf{x}_i^I\}} \left(p_i^I + \hat{y}_i^I(\mathbf{w}|\mathbf{x}_i^I) \right) \ , \tag{3}$$

where N_s is the number of sample points i in image I, and p_i^I the position of the sample point.

To account for the global image prediction in our refinement procedure, we therefore introduce an energy of the form

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{1}{2}\frac{\lambda_1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} (y^I - \bar{y}^I(\mathbf{w}))^2 + \frac{1}{2}\frac{\lambda_2}{|\{\mathbf{x}_i\}|} \sum_{\{\mathbf{x}_i\}} (y_i - \hat{y}_i(\mathbf{w}|\mathbf{x}_i))^2 \ , \tag{4}$$

where the first term regularizes the leaf weights, the second term optimizes for the combined votes on a *per image* base and the third one further regularizes the weights by minimizing the error of the offset predictions *per sample point*. We solve this sparse linear regression problem using stochastic gradient descent with momentum and a lazy update strategy for the L2 regularization. We report the hyper-parameters in Table 1.

Inference for a new image. Following [3], we evaluate our forest at randomly sampled center pixels according to Equation 2. Then we generate the votes for each of the 3 dimensions by

$$v_i = p_i^I + \hat{y}_i^I(\mathbf{w}|\mathbf{x}_i^I) \ , \tag{5}$$

and propose a vote weight of the form

$$u_i = \mathbf{v}^T \phi(\mathbf{x}_i^I) \text{ with } \mathbf{v}_l = e^{-E_l^2} \tag{6}$$

Each element $\mathbf{v}_l$ of the leaf confidence vector $\mathbf{v}$ depends on the average error E_l produced by the associated leaf l at training time.

Finally, we combine all 3D votes using a weighted mean-shift strategy. To exploit temporal consistency, we add a 3D Gaussian prior on the joint position of either the last time step or from the model fitting component feedback (see Section 3.4) to the mean shift weights. This smooths the tracking and introduces additional robustness towards occlusions.

3.4 Model fitting and tracking

For many professional applications, the coarse accuracy of discriminative methods is a significant drawback. We propose to combine our data-driven shape and pose estimation by means of a model-fitting strategy: by taking both the predicted body shape and the coarse alignment provided by the joint positions into account, we improve the accuracy of shape and pose estimation.

Body shape refinement Starting from our subspace-based estimates for the body shape coefficients, we now add a refinement step to further improve the torso shape and surface estimation accuracy. We propose a model fitting procedure with two alternating ICP-based methods. One iteration starts by correcting for the misalignment of the model with the observed data using standard rigid registration. Then the body shape is adapted by performing gradient descent on model parameters to fit corresponding points as closely as possible. This two-step process is repeated until convergence. Note that we already have coarse initializations for both the alignment and the shape parameters. We assume the upper body pose to change rigidly with pose and therefore use rigid (non-articulated) ICP on points associated with the upper body.

We then formulate the shape fitting as an energy minimization problem on the body shape coefficients $\mathbf{s}$:

$$E(\mathbf{s}|C, s_0) = \frac{\lambda}{2}\|\mathbf{s}\|_2^2 + \frac{1}{2}(\mathbf{s} - s_0)^T \Lambda(\mathbf{s} - s_0) + \frac{1}{2}\frac{1}{N_p}\|(C - (M + V\mathbf{s}))\|_2 \,. \quad (7)$$

Given the observed points $C \in \mathbb{R}^{3N_p}$ corresponding to the N_p model points, we minimize the squared error to the model, which is computed as the mean shape $M \in \mathbb{R}^{3N_p}$ plus the shape variations $V \in \mathbb{R}^{3N_p \times 20}$ according to the current shape coefficients $\mathbf{s}$. The first term penalizes the distance from the mean shape to avoid unlikely shapes. If prior information about the shape s_0 is available, we incorporate it using a diagonal regularization matrix Λ. In doing so, we are able to control which parameters of $\mathbf{s}$ should be primarily refined by the process, given that some parameters of the prior s_0 are already estimated with high confidence and should not be affected during optimization. Both regularizations effectively serve as (anisotropic) Gaussian priors. This convex optimization problem can be solved efficiently via gradient descent techniques.

Model-based surface tracking and feedback Fitting the refined body model to the observed data is a straight-forward and effective way to exploit body shape information for tracking. In every new frame, we make use of our discriminative 3D joint position estimates and initialize an ICP-based rigid alignment to improve accuracy.

The refined joint positions are then fed back to serve as a prior for the mean-shift procedure in Section 3.3. Only a few ICP iterations are enough to improve accuracy, since the initialization can be assumed to already be reasonably good, especially once the feedback loop is closed. This approach is very robust towards dynamic occlusions and enables efficient model-based tracking of the upper body surface.

4 Experiments

4.1 Dataset

To evaluate our methods on a large set of different body shapes and poses, we use the *HumanVP* dataset created from synthetically generated mesh data. The

Tree training	
Sample points	512 per image
Trees	3
Tree depth	24
Decision Functions	
Probe offset range	+-50px
Threshold range	+-5cm
Candidates	128 per split
min. #Samples	50 to split

Refinement	
Method	SGD+Momentum
Initial LR	1
λ_1	10^{-3}
λ_2	10^{-4}
Momentum	0.8
Epochs	48
Batch size	1
LR decay	0.9

Table 1. Hyperparameters for our 3D body joint estimation algorithm.

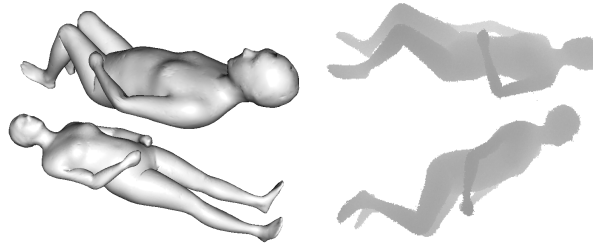


Fig. 4. Example posed body meshes from the MPII Human Shape Model [7] and rendered synthetic depth images from our *HumanVP* dataset.

motivation behind this is that it allows us to easily annotate the ground-truth shape parameters and joint positions, and carefully evaluate the behavior of our methods under different conditions, such as different body shapes, different poses, and different levels of occlusions.

In particular, we employed the MPII Human Shape Model [7]. This rigged statistical shape model was created from the CAESAR dataset [32], which contains a wide variety of body shapes represented as 3D meshes. These meshes are in vertex correspondence, and annotations for body parts and joint positions are provided. The depth data was obtained by sampling from the 4000 CAESAR-fitted body meshes of [7]. We randomly assigned a pose combination chosen from a set of 750 sub-poses of upper body (bending, torsion), arm (straight, angled, supporting head, ...) and leg (straight, angled) to each one of these meshes. Two example meshes are shown in Fig. 4a. We rendered depth images from 12 different viewpoints (random rotations around 2 rotation axes) using OpenGL. To this end, we employed a virtual camera that mimics the projection properties (FoV angles, resolution, aspect ratio) of the Kinect sensor. We further used the noise model of [33], which has been shown to yield synthetic depth maps that are very similar to real ones. This resulted in a total of about 50 000 images. Figure 4b shows some example images. In our experiments, we partitioned these images into training and test sets based on the 4000 mesh models, and thus divided the images created from the models accordingly. We used 70% of the models for training and 30% for testing.

4.2 Body shape estimation from a single frame

We compare our subspace-based shape estimation method on the *HumanVP* dataset with plain model-fitting and their combination. To this end, we report estimation errors on the first 5 shape coefficients. Since these coefficients represent a body shape in a linear subspace of a high-dimensional statistical shape model, the prediction errors can be seen as a proxy for the average per-vertex error between the reconstructed body model and the ground truth model. The second shape parameter $\mathbf{s}^1$ mainly captures the belly/upper body shape (see Figure 2) and is therefore the most interesting in this context. All methods are provided with the same initial 3D joint positions. Our reported values are normalized by the respective PCA standard deviation (see Section 3.2).

We learn $N_d = 20$ eigen-images for our subspace representation of the canonical images (640×480 px) and train a forest of 16 trees to predict the body shape coefficients using the MATLAB TreeBagger implementation. For the model fitting (see Equation 7), we chose $\lambda = 0.1$ and the maximum correspondence distance for ICP to 0.1 m.

Subspace regression vs. model fitting. To enable a fair comparison and to avoid distortion of the results due to local minima during model fitting, we take the following measures: We start from a set of initial shapes and select the best fitting solution. Also, we reject unlikely shape results, if at least one predicted shape parameter exceeds the $\pm 3\sigma$ threshold. We therefore discarded about 17 % of test images for the model fitting results.

Figure 5 shows the resulting RSME (b) and error standard deviation (c) for the first 5 shape coefficients. Our subspace-based regression method consistently provides lowest errors on all shape parameters and poses a very reasonable initialization. We can see that on the most important coefficient $\mathbf{s}^1$, the methods perform on a similar level, model fitting however induces slightly less RSME and less error deviation. Due to the fact that we only fit to the upper body, model fitting alone is not able to recover shape parameters that are less related to the torso shape. Also, the higher order coefficients capture smaller shape details which tend to result in more noisy estimates. While the baseline needs about 12 s per frame, our subspace-based method runs significantly faster, demanding less than 2 s on average.[4]

Combining subspace regression and model fitting. Following our paradigm of combining discriminative and generative methods, we investigated the benefits of using our subspace-based method to initialize the model fitting procedure. We set $\Lambda_{i \neq 1} = \mathrm{diag}(1)$, $\Lambda_1 = 10^{-3}$ to focus on refining the torso shape while keeping the other coefficients close to the initial estimate (see Equation 7). Our results show that our refinement strategy significantly improves the accuracy of the torso shape coefficient $\mathbf{s}^1$ compared to the initial shape provided by our subspace regression method. As desired, prediction errors of other parameters

[4] Evaluation hardware: Intel Core i7-4790K CPU 4.00GHz, 16Gb RAM.

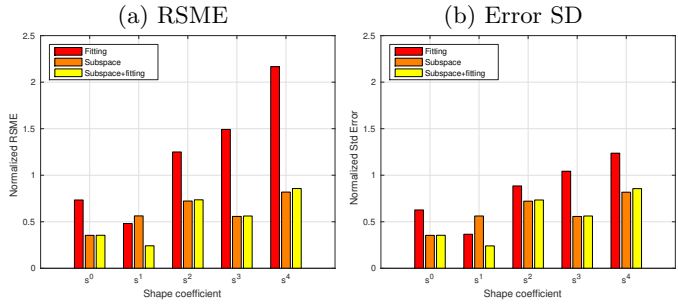


Fig. 5. Body shape estimation: (a) normalized RSME and (b) standard deviation (SD) of our subspace method, model-fitting, and their combination on *HumanVP*.

change only marginally. Due to the effective regularization, the refinement converges after about 6 s on average.

Shape estimation efficiency To gauge the dense surface accuracy of the estimated upper body shape, we evaluated the 3D displacement error per model vertex and the error standard deviation. Note that we consider the complete set of vertices of the full upper body model. Since a significant portion of points is not visible in a single image, this measure is a very interesting accuracy benchmark and shows the strength of model-based approaches. Our combined approach achieves an average error of approximately 9 mm ($\sigma = 6$ mm) per model vertex, compared to 12 mm ($\sigma = 9$ mm) without model-fitting. Since the data-driven subspace method does not consider the domain shift from synthetic data, it is however safe to assume that this difference will increase when dealing with real data. We can therefore conclude that our approach yields upper body shape estimates of reasonable accuracy.

4.3 Joint position estimation from a single frame

To evaluate the accuracy of the predicted 3D salient point positions, we report both the joint detection rate (true positive, if the prediction is within 3 cm from the GT) and the RSME across 6 upper body joints/salient points. As our method extends the one proposed by Girshick et al. [3], it is a natural choice to compare the two methods on our *HumanVP* dataset. We use the same hyper-parameters for both methods and optimize the mean-shift procedure to the upper body joints. In order to better handle occlusions, we reduced the range of probing offsets to 50 px (see Table 1), without noticeable loss in prediction accuracy for upper body joints. The results in Figure 6 show that our forest refinement strategy consistently improves prediction accuracy on all upper body joints but the belly reference point. This could indicate an upper bound on the attainable accuracy with this random forest model, since the belly point predictions are the most accurate among all predicted points. Figure 7 depicts some qualitative results on real Kinect data. Combining discriminative pose estimation with ICP-based refinement using our estimated body model (Figure 6b) significantly boosts localization accuracy.

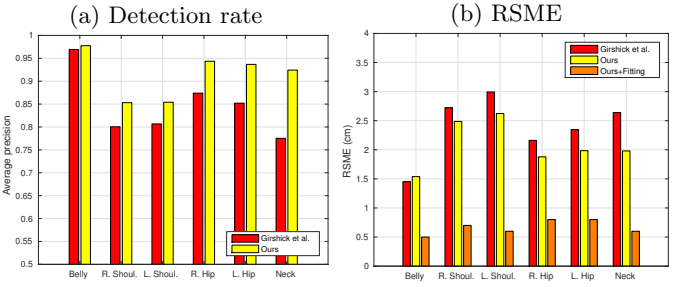


Fig. 6. 3D Joint position estimation: Comparison of our methods with [3] on our *HumanVP* dataset.
(a) Detection rate (3 cm threshold)
(b) Localization error

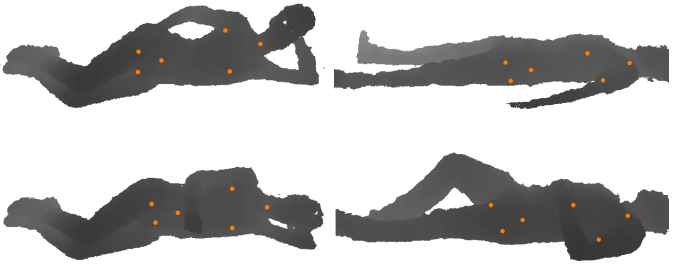


Fig. 7. Visualization of discriminative joint position estimates on real data, showing different exemplary subjects and poses. They provide a sound initialization for subsequent model fitting.

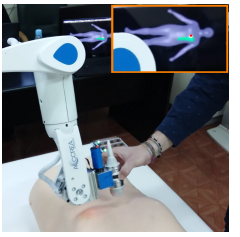
4.4 Robust upper body tracking framework

We now give a preliminary qualitative impression of our complete tracking framework, combining shape and pose information with temporal feedback. We visualize results on real data, since we do not have annotated depth sequences for evaluations yet. Our data originates from a mobile robot platform that features a 7 DoF robotic arm used for performing remote-controlled ultrasonography examinations within the *ReMeDi* project. Our task is to localize the end-effector w.r.t. the patient under severe occlusions due to viewpoint limitations and the robot arm moving in front of the sensor. We implemented our method as three processing nodes (shape estimation, pose estimation and model-fitting) within the ROS[5] framework. Without further optimization, the tracking update frequency is about 10Hz on a conventional laptop [6]. The Kinect coordinate frame has been calibrated to the robot base frame to enable transformations between the 3D data and the robot arm. Looking at the random forest estimates for the joint position in Figure 7 reveals that they pose a sound initialization for the subsequent model fitting. Figure 8a visualizes the localization of the patient w.r.t. to the robotic platform while the robotic arm is partly occluding the subject.

Hands-on experiments showed that for typically slow upper body motions of the patient, our approach yields practical results for online mapping of probe measurements to a position on the 3D body model during examination (see Figure 8b). In contrast, using only discriminative pose estimation produces qualita-

[5] Robot Operating System (ROS). http://www.ros.org/
[6] Hardware: Intel Core i7-4510U 2.0Ghz, 8Gb RAM.

(a) Overlaying the estimated body model with the synchronous Kinect RGB-D stream.



(b) Online probe mapping (red dot) and tracking history (green)

Fig. 8. Visualization of the estimated localized patient w.r.t. the robot platform (a). During examination, the robotic arm is partly occluding the subject. Note that we only fit to the torso of the patient. (b) Estimated probe position w.r.t to mannequin upper body. Our approach remains robust towards occluded/missing body parts.

tively more unstable and inaccurate results on the body surface. This confirms our results on the synthetic data (Figure 6b). ICP-based refinement with an adequate body model is able to correct for surface misalignment and introduces significant robustness towards occlusions and missing body parts.

5 Conclusion and Future Work

We have proposed a hybrid approach towards rapid shape estimation and real-time pose tracking of the human upper body. We employ fast data-driven methods in combination with a model fitting-based refinement strategy to exploit body shape for accurate torso tracking in real-time.

We introduced a subspace-based algorithm to estimate body shape parameters directly from a single depth image and showed that it provides a sound initialization for model fitting methods. Our second contribution is the development of a suitable random forest refinement strategy for the well known body joint position estimation framework by Girshick et al. [3]. Our experiments show that the prediction error is distributed advantageously across training images and the method therefore generalizes better on upper body joints.

Moreover, we provided our tracking framework with an ICP-based refinement for both upper body shape and pose, and presented qualitative results on real data. To encourage temporal consistency and induce robustness towards occlusions due to dynamic objects in the scene, we proposed a feedback mechanism that improves the interplay between data-driven and model-based torso tracking.

For future work, we plan to evaluate our tracking method on depth image sequences in a more quantitative manner. We also intend to tackle the problem of non-rigid shape changes induced by bending and torsion poses and aim to investigate novel methods for predicting body shape directly from depth data.

Acknowledgment

This work is funded by the EU Framework Seven project *ReMeDi* (grant 610902).

References

1. Bauer, S., Seitel, A., Hofmann, H.G., Blum, T., Wasza, J., Balda, M., Meinzer, H.P., Navab, N., Hornegger, J., Maier-Hein, L.: Real-Time Range Imaging in Health Care: A Survey. In: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications. (2013) 228–254
2. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. Studies in Computational Intelligence **411** (2013) 119–135
3. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient regression of general-activity human poses from depth images. Proceedings of the IEEE International Conference on Computer Vision (2011) 415–422
4. Jung, H.Y., Lee, S., Comp, D., Eng, E.S.: Random Tree Walk toward Instantaneous 3D Human Pose Estimation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2015)
5. Anguelov, D., Srinivasan, P., Thrun, S., Daphne, K., Davis, J., Rodgers, J.: SCAPE: Shape Completion and Animation of People. Lecture Notes in Computer Science **7729 LNCS**(PART 2) (2013) 133–147
6. Hasler, N., Stoll, C.: A statistical model of human pose and body shape. Eurographics **28**(2) (2009) 1–10
7. Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., Schiele, B.: Building Statistical Shape Spaces for 3D Human Modeling. In: arXiv. (2015)
8. Zuffi, S., Black, M.J.: The Stitched Puppet : A Graphical Model of 3D Human Shape and Pose. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
9. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A Skinned Multi-Person Linear Model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6) (oct 2015) 248:1—-248:16
10. Helten, T., Baak, A., Bharaj, G., Muller, M., Seidel, H.P., Theobalt, C.: Personalization and evaluation of a real-time depth-based full body tracker. In: International Conference on 3D Vision (3DV). (2013)
11. Zhang, Q., Fu, B., Ye, M.: Quality Dynamic Human Body Modeling Using a Single Low-cost Depth Camera. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
12. Xu, H., Yu, Y., Zhou, Y., Li, Y., Du, S.: Measuring accurate body parameters of dressed humans with large-scale motion using a Kinect sensor. Sensors **13**(9) (2013) 11362–11384
13. Perbet, F., Johnson, S., Pham, M.T., Stenger, B.: Human Body Shape Estimation Using a Multi-resolution Manifold Forest. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
14. Weiss, A., Hirshberg, D., Black, M.J.: Home 3D Body Scans from Noisy Image and Range Data. In: International Conference on Computer Vision (ICCV). (2011)
15. Bogo, F., Black, M.J., Loper, M., Romero, J.: Detailed Full-Body Reconstructions of Moving People from Monocular RGB-D Sequences. ICCV (2015)
16. Newcombe, R.a., Fox, D., Seitz, S.M.: DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
17. Newcombe, R.a., Molyneaux, D., Kim, D., Davison, A.J., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: Real-Time Dense Surface Mapping and Tracking. Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (2011)

18. Cui, Y., Chang, W., Tobias, N.: KinectAvatar:Fully Automatic Body Capture Using a Single Kinect. In: ACCV Workshop on Color Depth Fusion in Computer Vision. (2012)
19. Zeng, M., Zheng, J., Cheng, X., Liu, X.: Templateless Quasi-rigid Shape Modeling with Implicit Loop-Closure. 2013 IEEE Conference on Computer Vision and Pattern Recognition (2013) 145–152
20. Tong, J., Zhou, J., Liu, L., Pan, Z., Yan, H.: Scanning 3D Full Human Bodies using Kinects. IEEE Transactions on Visualization & Computer Graphics (2012)
21. Li, H., Vouga, E., Gudym, A., Luo, L., Barron, J.T., Gusev, G.: 3D Self-Portraits
22. Ganapathi, V., Plagemann, C.: Real-time human pose tracking from range data. In: ECCV. (2012) 1–14
23. Gall, J., Stoll, C., De Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P.: Motion capture using joint skeleton tracking and surface estimation. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009 (2009) 1746–1753
24. Grest, D., Krüger, V., Koch, R.: Single view motion tracking by depth and silhouette information. Image Analysis (2007) 719–729
25. Sun, M., Kohli, P., Shotton, J.: Conditional regression forests for human pose estimation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2012) 3394–3401
26. Baak, A., Muller, M., Bharaj, G., Seidel, H.P., Theobalt, C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. Proceedings of the IEEE International Conference on Computer Vision (2011) 1092–1099
27. Ye, M., Yang, R., Pollefeys, M.: Accurate 3D pose estimation from a single depth image. 2011 International Conference on Computer Vision (2011) 731–738
28. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real time motion capture using a single time-of-flight camera. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2010) (2010) 755–762
29. Pons-moll, G., Javier, R., Mahmood, N., Black, M.J.: Dyna: A Model of Dynamic Human Shape in Motion. ACM Transactions on Graphics (2015) 1–14
30. Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.: The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 103–110
31. Ren, S., Cao, X., Wei, Y., Sun, J.: Global Refinement of Random Forest. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
32. Robinette, K.M., Daanen, H., Paquet, E.: The CAESAR project: a 3-D surface anthropometry survey. In: International Conference on 3-D Digital Imaging and Modeling. (1999)
33. Nguyen, C.V., Izadi, S., Lovell, D.: Modeling kinect sensor noise for improved 3D reconstruction and tracking. In: International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT). (2012)