HAL open science

Building Statistical Shape Spaces for 3D Human Modeling

Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, Bernt Schiele

To cite this version:

Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, Bernt Schiele. Building Statistical Shape Spaces for 3D Human Modeling. Pattern Recognition, 2017, 67, pp.276-286. hal-01136221v1

HAL Id: hal-01136221

https://inria.hal.science/hal-01136221v1

Submitted on 27 Mar 2015 (v1), last revised 3 Mar 2017 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Building Statistical Shape Spaces for 3D Human Modeling

Leonid Pishchulin, *Student Member, IEEE,* Stefanie Wuhrer, Thomas Helten, Christian Theobalt, *Member, IEEE,* Bernt Schiele, *Member, IEEE*

Abstract—Statistical models of 3D human shape and pose learned from scan databases have developed into valuable tools to solve a variety of vision and graphics problems. Unfortunately, most publicly available models are of limited expressiveness as they were learned on very small databases that hardly reflect the true variety in human body shapes. In this paper, we contribute by rebuilding a widely used statistical body representation from the largest commercially available scan database, and making the resulting model available to the community (visit *http:// humanshape.mpi-inf.mpg.de*). As preprocessing several thousand scans for learning the model is a challenge in itself, we contribute by developing robust best practice solutions for scan alignment that quantitatively lead to the best learned models. We make implementations of these preprocessing steps also publicly available. We extensively evaluate the improved accuracy and generality of our new model, and show its improved performance for human body reconstruction from sparse input data.

Index Terms—statistical human body model, non-rigid template fitting

1 Introduction

Statistical human shape models represent variations in human physique and pose using low-dimensional parameter spaces, and are valuable tools to solve difficult vision and graphics problems, e.g. in pose tracking or animation. Despite significant progress in modeling the statistics of the complete 3D human shape and pose [1], [3], [12], [9], [20], [16], [18] only few publicly available statistical 3D body shape spaces exist [16], [18]. Further on, the public models are often learned on only small datasets with limited shape variations [16]. The reason is a lack of large representative *public* datasets and the significant effort required to process and align raw data scans prior to learning a statistical shape space.

This paper contributes by systematically constructing a model of 3D human shape and pose from the largest *commercially* available dataset of 3D laser scans [24] and making it publicly available to the research community (Section 2). Our model is based on a simplified and efficient variant of the SCAPE model [3] (henceforth termed *S-SCAPE space*) that was described by Jain et al. [18] and used for different applications in computer vision and graphics [18], [23], [22], [17], [19], but was never learned from such a complete dataset. This compact shape space learns a probability distribution from a dataset of 3D human laser scans. It models variations due to changes in identity using a principal component analysis (PCA) space, and variations due to pose using a skeleton-based surface skinning approach. This representation makes the model versatile and computationally efficient.

Prior to statistical analysis, the human scans have to be processed and aligned to establish correspondence. We contribute by evaluating different variants of state-of-the-art techniques for non-rigid template fitting and posture normalization to process the raw data [1], [16], [35], [20]. Our findings are not entirely new methods, but best practices and specific solutions for automatic preprocessing of large scan databases for learning the *S-SCAPE* model in the best way (Section 3). First, shape and posture fitting of an initial shape model to a raw scan prior to non-rigid deformation considerably improves the results. Second, multiple passes over the dataset improve initialization and thus increase the overall fitting accuracy and statistical model qualities. Third, posture normalization prior to shape space learning leads to much better generalization and specificity.

The main contribution of our work is a set of *S-SCAPE* spaces learned from the largest database that is currently commercially available [24]. The differences in our *S-SCAPE* spaces stem from differences in the registration and pre-alignment of the human body scans. We evaluate different data processing techniques in Section 4 and the resulting shape spaces in Section 5. Finally, we compare our *S-SCAPE* spaces to the state-of-the-art, which is a *S-SCAPE* space learned from a publicly available database [16] for the application of reconstructing full 3D body models from monocular depth images in Section 6. Our experimental evaluation clearly demonstrates the advantages of our more expressive shape models in terms of shape space quality and for the task of reconstructing 3D human body shapes from monocular depth images (Section 6).

We release the new shape spaces with code to (1) preprocess raw scans and (2) fit a shape space to a raw scan

- *L. Pishchulin, C. Theobald and B. Schiele are with the Max Planck Institute for Informatics, Germany*
- *S. Wuhrer is with INRIA Grenoble Rhône-Alpes, France*
- *T. Helten is with GoalControl GmbH, Germany*

for public usage. We believe this contribution is required for future development in human body modeling. Visit *http://humanshape.mpi-inf.mpg.de* to get code and models.

Related work

Datasets

Several datasets have been collected to analyze populations of 3D human bodies. Many publicly available research data sets allow for the analysis of shape and posture variations jointly; unfortunately they feature data of at most 100 individuals [3], [16], [7], which limits the range of shape variations. We therefore use the CAESAR database [24], the largest commercially available dataset to date that contains 3D scans of over 4500 American and European subjects in a standard pose, which is a much richer sample of the human physique.

Statistical shape spaces of 3D human bodies

Building statistical shape spaces of human bodies is challenging, as there is strong and intertwined 3D shape and posture variability, yielding a complex function of multiple correlated shape and posture parameters. Methods to learn this shape space usually follow one of two routes. The first group of methods learn shape- and posture-related deformations separately and combine them afterwards [3], [12], [9], [20], [18]. These methods are inspired by the SCAPE model [3], which couples a shape space capturing variation in body shape with a posture space learned from deformations of a single subject. All SCAPE-like models use a set of transformations per triangle to encode shape variations in a shape space. Hence, to convert between the vertex coordinates of a processed scan and its representation in shape space, a computationally demanding optimization problem needs to be solved. To overcome this difficulty, a simplified version of the SCAPE model (*S-SCAPE*) was proposed that models pose variation with an efficient skeleton-based surface skinning approach [18], [17], [19].

Another group of methods intends to perform simultaneous analysis of shape and posture variations [2], [16]. These methods learn skinning weights for corrective enveloping of posture-related shape variations, which allows to explore both shape and posture variations using a single shape space. Furthermore, it allows for realistic muscle bulging as shape and posture are correlated [21]. It has been shown, however, that for many applications in computer vision and graphics this level of detail is not required and simpler and computationally more efficient shape spaces can be used [18], [17], [23], [22].

Mesh registration

Mesh registration is performed on the scans to bring them in correspondence for statistical analysis. Two surveys [31], [30] review such techniques, and a full review is beyond the scope of this paper. Allen et al. [1] use non-rigid template fitting to compute correspondences between human body shapes in similar posture. This technique has been extended to work for varying postures [3], [2], [16], and in scenarios where no landmarks are available [34]. In this work, we evaluate a non-rigid template fitting approach inspired by Allen et al. [1].

Applications

Statistical spaces of human body shape and posture are applicable in many areas including computer vision, computer graphics, and ergonomic design; our new model that was learned on a large commercially available dataset is beneficial in each of these applications. Statistical shap spaces have been used to predict body shapes from partial data, such as image sequences and depth images [27], [6], [28], [36], [13], [14], [32], [8], [17] and semantic parameters [26], [1], [10], [4], [33], [25]. Furthermore, they have been used to estimate body shapes from images [5] and 3D scans [15], [**?**] of dressed subjects. Given a 3D body shape, statistical shape spaces can be used to modify input images [37] or videos [18], to automatically generate training sets for people detection [23], [22], or to simulate clothing on people [12].

2 Statistical modeling with SCAPE

We briefly recap the efficient version of the SCAPE model [18] which we use and discuss its differences to the original SCAPE model [3] in more detail. For learning the model, both methods assume that a template mesh $\mathbf{T}$ has been deformed to each raw scan in a database. All scans of the database are assumed to be rigidly aligned, e.g. by Procrustes Analysis [11].

2.1 Original SCAPE model

In the original SCAPE model, the transformation of each triangle of $\mathbf{T}$ is modeled as combination of two linear transformations $\mathbf{R}_{m,i} \in \mathrm{SO}(3)$ and $\mathbf{Q}_{m,i} \in \mathbb{R}^{3\times 3}$. Index i indicates one particular scan $\mathbf{T}$ is fitted to and we refer to the fitting result after rigid alignment with $\mathbf{T}$ as instance mesh $\mathbf{M}_i$. While $\mathbf{R}_{m,i}$ represents the posture of the person as global rotation induced by the deformation of an underlying rigid skeleton, $\mathbf{Q}_{m,i}$ encodes the individual deformations of each triangle that originates from varying body shape or non-rigid posture dependent surface deformations such as muscle bulging. Computing $\mathbf{Q}_{m,i}$ for each vertex separately is an underconstrained problem. Therefore, smoothing is applied so that $\mathbf{Q}_{m,i}$ of neighboring vertices are dependent. Finally, by applying dimensionality reduction techniques to the transformations $\mathbf{R}_{m,i}$ and $\mathbf{Q}_{m,i}$, one obtains a flexible model that covers a wide range of possible surface deformations. However, as the model does not explicitly encode vertex position, one needs to solve a complex least squares problem to reconstruct the mesh surface.

2.2 Simplified SCAPE (*S-SCAPE*) space

The aforementioned computational overhead is often prohibitive in applications where speed is more important than the overall reconstruction quality, or when

many samples need to be drawn from the shape space. *S-SCAPE* space [18] reconstructs vertex positions in a given posture and shape without needing to solve a Poisson system. To learn the model, only laser scans in a standard posture χ_0 are used. A PCA model of the meshes $\mathbf{M}_i$ is learned, which represents each shape using a parameter vector φ. This shape space only covers variations in overall body shape and not posture. An articulated skeleton is fitted to the average human shape, and linear blend skinning weights to attach the surface to the bones are computed. The skeleton scales in accordance to the body shape by expressing joint locations relative to nearby surface vertex locations.

For reconstructing a model of shape φ in skeleton pose χ (joint angle parameters), the method first calculates a personalized mesh $\mathbf{M}_{\varphi,\chi_0}$ using φ. Then a linear blend skinning is applied to the personalized mesh to obtain the final mesh $\mathbf{M}_{\varphi,\chi}$ in pose χ. While *S-SCAPE* approach shows much faster reconstruction speed, especially when the personalized mesh and skeleton can be precomputed, its reconstruction quality is inferior to the original SCAPE approach. In the rest of this paper, we use this simple and efficient shape space.

3 Data processing

This section describes how to preprocess a set of raw body scans to establish correspondence and pre-align the models. We show best-practice ways how non-rigid template fitting can be used to register raw scans, how to initialize the template fitting, how bootstrapping can help to improve the correspondence, and how the postures of registered scans can be normalized. Tools to reproduce these steps will be made publicly available.

3.1 Non-rigid template fitting

Our method to fit a human shape template $\mathbf{T}$ to a human scan $\mathbf{S}$ is inspired by Allen et al. [1]. In non-rigid template fitting (henceforth abbreviated *NRD*), each vertex $\mathbf{p}_i$ of $\mathbf{T}$ is transformed by a 4×4 affine matrix $\mathbf{A}_i$, which allows for twelve degrees of freedom during the transformation. The aim is to find a set of matrices $\mathbf{A}_i$ that define vertex positions in a deformed template matching well with $\mathbf{S}$. The fitting is done by minimizing a combination of data, smoothness and landmark errors.

Data term

The data term requires each vertex of the transformed template to be as close as possible to its corresponding vertex of $\mathbf{S}$, and takes the form

$$E_d = \sum_{i=1}^{N} w_i ||\mathbf{A}_i \mathbf{p}_i - NN_i(\mathbf{S})||_F^2, \tag{1}$$

where N is the number of vertices in $\mathbf{T}$, w_i weights the error contribution of each vertex, $||.||_F$ denotes the Frobenius norm, and NN_i is a closest compatible point in $\mathbf{S}$. If surface normals of closest points are less than $60°$ apart and the distance between the points is less than 20 mm, we set w_i to 1, otherwise to 0.

Smoothness term

Fitting using E_d only may lead to situations where neighboring vertices of $\mathbf{T}$ match to disparate vertices in $\mathbf{S}$. To enforce smooth surface deformations we use a smoothness term E_s that requires affine transformations applied to connected vertices to be similar, i.e.

$$E_s = \sum_{\{i,j|(\mathbf{p}_i,\mathbf{p}_j)\in edges(\mathbf{T})\}} ||\mathbf{A}_i - \mathbf{A}_j||_F^2. \tag{2}$$

Landmark term

Although using E_d and E_s would suffice to fit two surfaces that are close to each other, the optimization gets stuck in a local minimum when $\mathbf{T}$ and $\mathbf{S}$ are far apart. A remedy is to identify a set of points on $\mathbf{T}$ corresponding to known anthropometric landmarks on $\mathbf{S}$. In each CAESAR scan these are obtained by placing markers on each subject prior to scanning. Our landmark term penalizes misalignments between landmark locations

$$E_l = \sum_{i=1}^{M} ||\mathbf{A}_{k_i} \mathbf{p}_{k_i} - \mathbf{l}_i||_F^2, \tag{3}$$

where k_i is the landmark index on $\mathbf{T}$, and $\mathbf{l}_i$ is the landmark point on $\mathbf{S}$. Although there are only 64 landmarks compared to the total number of 6449 vertices, good landmark fitting is enough to get the deformed surface of $\mathbf{T}$ close to $\mathbf{S}$, and avoid local convergence.

Combined energy

The three terms are combined into a single objective

$$E = \alpha E_d + \beta E_s + \gamma E_l. \tag{4}$$

For optimization we use L-BFGS-B [38]. We vary the weights α, β and γ according to the following empirically found schedule. We first perform a single iteration of optimization without data term by setting $\alpha = 0$, $\beta = 10^6$, $\gamma = 10^{-3}$, which allows to bring the surfaces into a rough correspondence. We then allow the data term to contribute by setting $\alpha = 1$, $\beta = 10^6$, $\gamma = 10^{-3}$. In addition, we relax smoothness and landmark weights after each iteration of fitting to $\beta := 0.25\beta$ and $\gamma := 0.25\gamma$, thus allowing the data term to dominate. This is repeated until $\beta \leq 10^3$. Reducing β increases the flexibility of deformation and allows $\mathbf{T}$ to better reproduce fine details, while reducing γ is necessary due to unreliable placement of landmarks in some scans.

3.2 Initialization

For non-rigid template fitting to succeed, $\mathbf{T}$ should be pre-aligned to $\mathbf{S}$. We explore two initialization strategies.

A first standard way to initialize *NRD* is to use a static template with annotated landmarks. Corresponding landmarks are then used to rigidly align $\mathbf{S}$ to $\mathbf{T}$.

A second way to initialize the fitting is to start with a *S-SCAPE* space that was learned from a small registered dataset. Fitting the shape space to a scan is achieved

in three steps. First, $\mathbf{S}$ is rigidly aligned to $\mathbf{T}$. Second, the shape parameters of the *S-SCAPE* space are fixed while the posture parameters are adjusted to minimize the landmark term given in Eq. 3. Third, the shape and posture parameters of the *S-SCAPE* space are optimized iteratively to minimize the data term in Eq. 1. In each iteration, the posture is fitted in a first step and the shape is fitted in a second step. After each fitting step, the set $NN_i(\mathbf{S})$ is recomputed. This iterative procedure is repeated until E_d does not change significantly. For optimization, we use the iterative interior point method.

3.3 Bootstrapping

In many cases, even after *NRD*, $\mathbf{T}$ is far from $\mathbf{S}$. Using registered scans with a high fitting error for shape space learning may lead to unrealistic shape deformations in the learned space. A remedy is to visually examine each fitting, discard fittings of low quality, and learn a *S-SCAPE* space using the samples that passed the visual inspection. This *S-SCAPE* space is then used as initialization to perform a fitting during the next pass. This bootstrapping process is performed until nearly all registered scans pass the visual inspection. Note that visual inspection is required, as low average fitting errors do not always correspond to good results, since the fitting of localized areas may be inaccurate.

3.4 Posture normalization

The *S-SCAPE* space used in this study aims to represent shape and posture variations independently. However, by performing PCA over the vertex coordinates of processed scans captured in a standard posture, the shape space capturing variations caused by different identities is not normalized for posture. This may cause problems because the scans in standing posture present in the CAESAR and MPI Human Shape databases inevitably contain slight posture variations, mostly in the areas of the arms. To account for these variations, we compare the statistical shape space learned on the registered data directly to the one learned on data that was modified to remove posture variations. We consider two recent posture normalization approaches.

Wuhrer et al. [35] factor out variations caused by posture changes by performing PCA on localized Laplacian coordinates. While this approach leads to better shape spaces than performing PCA on the vertex coordinates, it is difficult to compare this shape space to the *S-SCAPE* space. We therefore modify each model $\mathbf{M}_i$ obtained by fitting $\mathbf{T}$ to $\mathbf{S}_i$ by initializing each shape to the mean shape $\overline{\mathbf{M}}$ and by optimizing the localized Laplacian coordinates to be as close as possible to the ones computed on $\mathbf{M}_i$. This leads to models that have the body shape of $\mathbf{M}_i$ in the posture of $\overline{\mathbf{M}}$.

Neophytou and Hilton [20] normalize the posture of each processed scan using a skeleton model and Laplacian surface deformation. While this type of normalization may introduce artefacts around joints when the posture is changed significantly, this approach is suitable to normalize the posture of models of the CAESAR database as the posture variations are minor. We use this method to modify the posture of each $\mathbf{M}_i$.

4 Evaluation of Template Fitting

We now evaluate the different components of our registration procedure on the CAESAR dataset [24]. Each CAESAR scan contains 73 manually placed landmarks. We exclude several landmarks located on open hands, as those are missing for our template, resulting in 64 landmarks used for registration. Furthermore, we remove all laser scans without landmarks and corrupted scans, resulting in 4308 scans.

4.1 Implementation details

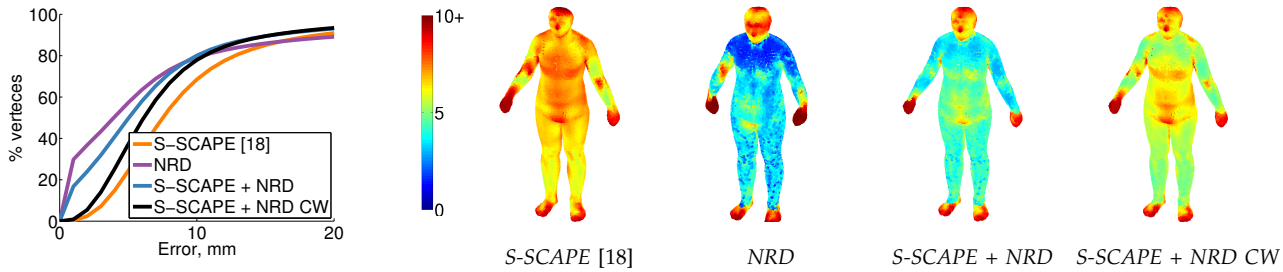
Non-rigid template fitting requires a human shape template as input, and the initialization procedure requires an initial shape space. We use registered scans of 111 individuals in neutral posture of the MPI Human Shape dataset to compute these initializations.

However, these data have artifacts in non-smooth areas at the head and neck. We smooth these areas by identifying problematic vertices and by iteratively recomputing their positions as an average position of direct neighbors. Furthermore, due to privacy reasons, head vertices of each human scan were replaced by the same dummy head, which is not representative and of low quality at the backside. We adjust the vertex compatibility criteria to compute nearest neighbors during *NRD* by allowing $30°$ deviation of the head face normals while increasing the distance threshold to 50 mm.

We employ the algorithm from Section 3.1 to compute correspondences for the CAESAR dataset. One inconsistency between the datasets is that the hands in the MPI Human Shape dataset are closed, while they are open in the CAESAR dataset. As remedy, we set α and γ to zero for hand vertices in Eq. 4, thus only allowing E_s to contribute. Prior to fitting, we sub-sample each CAESAR scan to have a total number of vertices that exceeds the number of vertices of $\mathbf{T}$ by a factor of three (6449 vertices in $\mathbf{T}$ vs. 19347 vertices in $\mathbf{S}$). This gives a good trade-off between fitting quality and computational efficiency.

4.2 Quality measure

Measuring the accuracy of surface fitting is not straightforward, as no ground truth correspondence between $\mathbf{S}$ and $\mathbf{T}$ is available. We evaluate the fitting accuracy by finding the nearest neighbor in $\mathbf{S}$ for each fitted template vertex. If this neighbor not further from its correspondence in $\mathbf{T}$ than 50 mm and its face normals do not deviate more than $60°$, the Euclidean vertex-to-vertex distance is computed. In our experiments we report both the proportion of vertices falling below a certain threshold and the distance per vertex averaged over all fitted templates. In the following, we first show the effects of



(a) Total fitting error (b) Mean fitting error

Fig. 1. Fitting error on the CAESAR dataset when using the *S-SCAPE* space [18] alone, *NRD* alone, and initializing using *S-SCAPE* prior to *NRD* with different weighting schedules (*S-SCAPE + NRD*, *S-SCAPE + NRD CW*). Shown are (a) the proportion of vertices [%] with fitting error below a threshold and (b) the average fitting error per vertex.

various types of initialization and weighting schemes in the *NRD* procedure on the fitting error. Second, we show the effect of performing multiple bootstrapping rounds.

4.3 Initialization

First, we evaluate two different initialization strategies used in our fitting procedure. We compare the results when using an average human template (*NRD*) to the result when using the *S-SCAPE* space learned on the MPI Human Shape dataset (*S-SCAPE + NRD*) for initialization. We compare the results by both non-rigid deformation schemes to the fitting accuracy when fitting the publicly available *S-SCAPE* space by Jain et al. [18] without any non-rigid deformation (*S-SCAPE*).

The results are shown in Fig. 1. The total fitting error in Fig. 1(a) shows that *NRD* achieves good fitting results in the low error range of 0 – 10 mm, as it can produce good template fits for the areas where $\mathbf{T}$ is close to $\mathbf{S}$. However, as *NRD* is a model-free method, the smooth topology of $\mathbf{T}$ may not be preserved during the deformation, e.g. convex surfaces of $\mathbf{T}$ may be deformed into non-convex surfaces after *NRD*. This leads to large fitting errors for areas of $\mathbf{T}$ that are far from $\mathbf{S}$. *S-SCAPE + NRD* uses a shape space fitting prior to NRD, which allows for a better initial alignment of $\mathbf{T}$ to $\mathbf{S}$. Note that *S-SCAPE + NRD* results in a better fitting accuracy in the high error range of $10 - 20\ mm$. The fitting result by *S-SCAPE + NRD* favorably compares against using *S-SCAPE* alone. Although *S-SCAPE* results in deformations preserving the human body shape topology, the shape space is learned from the relatively specialized MPI Human Shape dataset containing mostly young adults and thus cannot represent all shape variations.

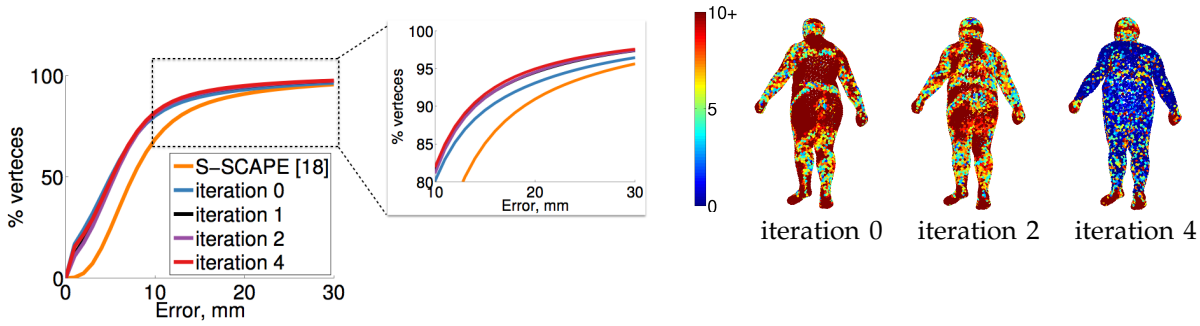
We also analyze the differences in the mean fitting errors per vertex in Fig. 1(b). *NRD* achieves good fitting results for most of the vertices. However, the arms are not fitted well due to differences in body posture of $\mathbf{T}$ and $\mathbf{S}$. Furthermore, the average fitting error is not smooth, which shows that despite using E_s, *NRD* may produce non-smooth deformations. In contrast, the result of *S-SCAPE + NRD* is smoother and has a lower fitting error for the arms. Clearly, the average fitting error of *S-SCAPE* is much higher, with notably worse fitting results for arms, belly and chest.

4.4 *NRD* parameters

Second, we evaluate the influence of the weight relaxation during *NRD* on the fitting accuracy. Specifically, we compare the standard weighting scheme where weights are relaxed in each iteration (*S-SCAPE + NRD*) to the case where the weights stay constant (*S-SCAPE + NRD CW*). Fig. 1(a) shows that the total fitting error of *S-SCAPE + NRD* is lower than *S-SCAPE + NRD CW*. This is because *S-SCAPE + NRD CW* enforces higher localized rigidity by keeping weights constantly high, while *S-SCAPE + NRD* relaxes the weights so that $\mathbf{T}$ can fit more accurately to $\mathbf{S}$. This explanation is supported by consistently higher per-vertex mean fitting errors in case of *S-SCAPE + NRD CW* compared to *S-SCAPE + NRD*, as shown in Fig. 1(b). The highest differences are in the areas of high body shape variability, such as belly and chest. Different weight reduction schemes such as $\beta := 0.5\beta$, $\gamma := 0.5\gamma$ and $\beta := 0.25\beta$, $\gamma := 0.25\gamma$ lead to better fitting accuracy compared to constant weights, with the latter scheme achieving slightly better results and faster convergence rates. We thus use the proposed weight reduction scheme in the following.

4.5 Bootstrapping

Third, we evaluate the fitting accuracy before and after performing multiple rounds of bootstrapping. To that end, we use the output of *S-SCAPE + NRD* (iteration 0) to learn a new statistical shape space, which is in turn used to initialize *NRD* during the second pass over the data (iteration 1). This process is repeated for five passes. The number of registered scans that survived the visual inspection after each round is 1771, 3253, 3641, 4237 and 4301, respectively. This results show that bootstrapping allows to register and thus to learn from an increasing number of scans. Fitting results are shown in Fig. 2. The close-up shows that although the overall fitting accuracy before and after bootstrapping is similar, bootstrapping allows to slightly improve the fitting accuracy in the range of $10 - 30\ mm$. Fitting results after three passes over the dataset (iteration 2) are



(a) Total fitting error

(b) Mean fitting error on a sample scan

Fig. 2. Fitting error after up to four bootstrapping rounds over the CEASAR database when using the *S-SCAPE* [18] as initialization for iteration 0. Shown is the proportion of vertices [%] with fitting error below a threshold.

slightly better compared to the initial fitting (iteration 0), and the accuracy is further increased after five passes (iteration 4). Fig. 2 (b) shows sample fitting results before and after several bootstrapping rounds. Largest improvements are achieved for the belly and chest; these are areas with large variability. The fitting improves with an increasing number of bootstrapping rounds. We use the fitting results after five passes (iteration 4) to learn the *S-SCAPE* space used in the following.

5 Evaluation of Statistical Shape Space

In this section, we evaluate the *S-SCAPE* space using the statistical quality measures generalization and specificity [29].

5.1 Quality measure

We use two complementary measures of shape statistics. Generalization evaluates the ability of a shape space to represent unseen instances of the object class. Good generalization means the shape space is capable of learning the characteristics of an object class from a limited number of training samples, poor generalization indicates overfitting of the training set. Generalization is measured using leave-one-out cross reconstruction of training samples, i.e. the shape space is learned using all but one training sample and the resulting shape space is fitted to the excluded sample. The fitting error is measured using the mean vertex-to-vertex Euclidean distance. Generalization is reported as mean fitting error averaged over the complete set of trials, and plotted as a function of the number of shape space parameters. It is expected that the mean error decreases until convergence as the number of shape space parameters increases.

Specificity measures the ability of a shape space to generate instances of the object class that are similar to the training samples. The specificity test is performed by generating a set of instances randomly drawn from the learned shape space and by comparing them to the training samples. The error is measured as average distance of the generated instances to their nearest neighbors in the training set. It is expected that the mean distance increases until convergence with increasing number of shape space parameters. We follow Styner et al. [29] and generate 10,000 random samples.

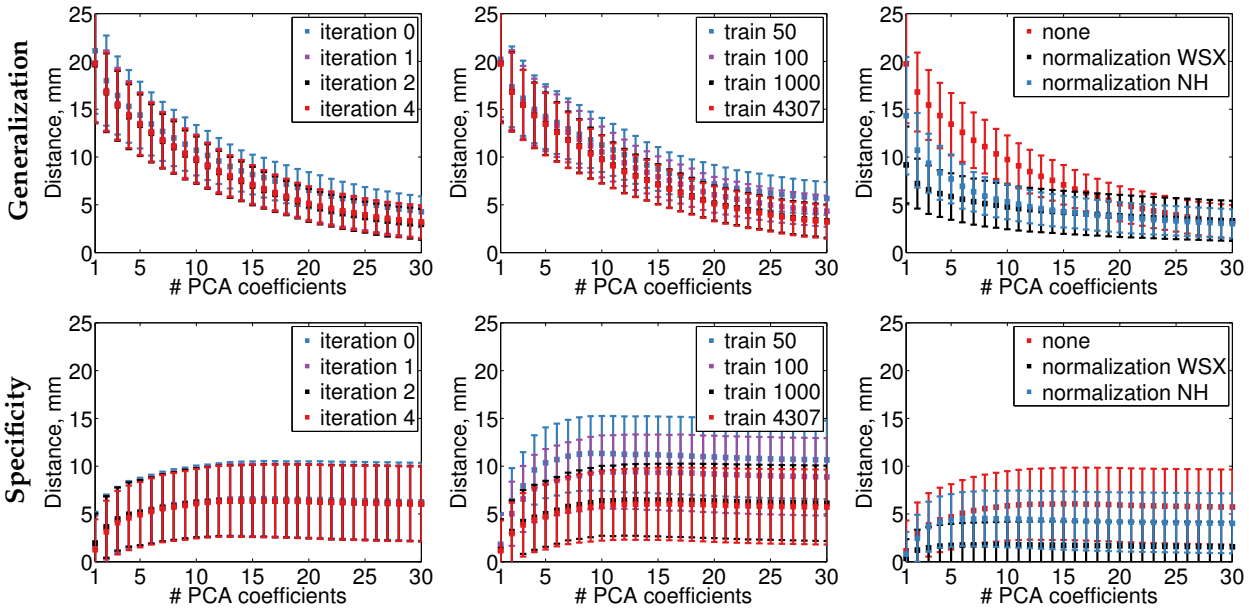
5.2 Bootstrapping

We evaluate the influence of bootstrapping on the quality of the statistical shape space by comparing models obtained after zero, one, two and four iterations of bootstrapping. The geometry of the training samples changes in each bootstrapping round, which makes the generalization and specificity results incomparable across different shape spaces. We thus use the training samples obtained after four iterations of bootstrapping as "ground truth", i.e., the reconstruction error of generalization and the nearest neighbor distance of specificity for each shape space is computed w.r.t. fitting results after four bootstrapping rounds. This allows for a fair comparison across different statistical shape spaces.

The results are shown in Fig. 3(a). Generalization error is already low after a single iteration of bootstrapping because after one iteration, the shape space is learned from a significantly larger number of training samples, thereby using samples with higher shape variation that were discarded in the 0th iteration. The following rounds of bootstrapping have little influence on generalization and specificity, with the shape space after four iterations resulting in a slightly lower specificity error than for previous iterations for a small number of shape parameters.

5.3 Number of training samples

To evaluate the influence of the number of training samples, we vary the number of samples obtained after four bootstrapping iterations. Specifically, we consider subsets of 50, 100, $1,000$ and $4,307$ $(all - 1)$ training samples. To compute a shape space, the desired number of training shapes are sampled from all training samples according to a learned PCA space. For generalization, we cross-evaluate on all $4,308$ training samples by leaving one sample out and by sampling the desired number of training shapes from the remaining samples. For specificity, we compute the nearest-neighbor distances to all $4,308$ training samples to find the closest sample.



(a) bootstrapping (b) # training samples (c) posture normalization

Fig. 3. Influence of different design choices on statistical quality measures. Shown are influence of (a) bootstrapping, (b) number of training samples and (c) posture normalization on generalization (top row) and specificity (bottom row). Best viewed with zoom on the screen.

The results are shown in Fig. 3(b). The shape space learned from the smallest number of samples performs worst. Increasing the number of samples consistently improves the performance with the best results achieved when using the maximum number. Both generalization and specificity error reduction is most pronounced when increasing the number of samples from 50 to 100. Further increasing the number of samples to $1,000$ affects specificity much stronger than generalization. This shows that the shape space learned from only 100 samples generalizes well, while its generative qualities are poor. Increasing the number of samples from $1,000$ to $4,307$ only slightly reduces both generalization and specificity errors, which shows that a high-quality statistical shape space can be learned from $1,000$ samples.

5.4 Posture normalization

Finally, we evaluate the generalization and specificity of the shape space obtained when performing posture normalization using the methods of Wuhrer et al. [35] (*WSX*) and Neophytou and Hilton [20] (*NH*). The results are shown in Fig. 3 (c). Posture normalization significantly improves generalization and specificity, with *WSX* achieving the best result. The reduction of the average fitting error in case of generalization is highest for a low number of shape parameters. This is because both *WSX* and *NH* lead to shape spaces that are more compact compared to the shape space obtained with unnormalized training shapes. Additionally, both posture-normalized shape spaces exhibit much better specificity. Compared to the shape space trained before posture normalization, randomly generated samples from both shape spaces trained after *WSX* and *NH* exhibit less variation in posture and are thus more similar to their corresponding posture-normalized training samples.

Finally, we qualitatively examine the first five PCA components learned by the following *S-SCAPE* spaces: the current state-of-the-art shape space *S-SCAPE* [18], our shape space without posture normalization and with posture normalization using *WSX* and *NH*. The results are shown in Fig. 4. Many of the major modes of shape variation by *S-SCAPE* (row 1) are affected by global and local posture-related deformations, such as moving of arms or tilting the body. In contrast, the principal components of variation by our shape space (row 2) are mostly due to shape changes thanks to a better template fitting procedure and a more representative training set. However, small posture variations are still part of the learned shape space. Performing posture normalization of the training samples prior to learning the shape space completely factors out changes due to posture, as can be seen in the shape spaces learned using *WSX* (row 3) and *NH* (row 4).

6 Human Body Reconstruction

Finally, we evaluate our improved *S-SCAPE* spaces in the specific application of estimating human body shape from sparse visual input. We follow the approach by Helten et al. [17] to estimate the body shape of a person from two sequentially taken front and back depth images. First, body shape and posture are fitted independently to each depth image. Second, the obtained results are used as initialization of a method that jointly optimizes over shape and independently optimizes over posture parameters. This optimization strategy is used because

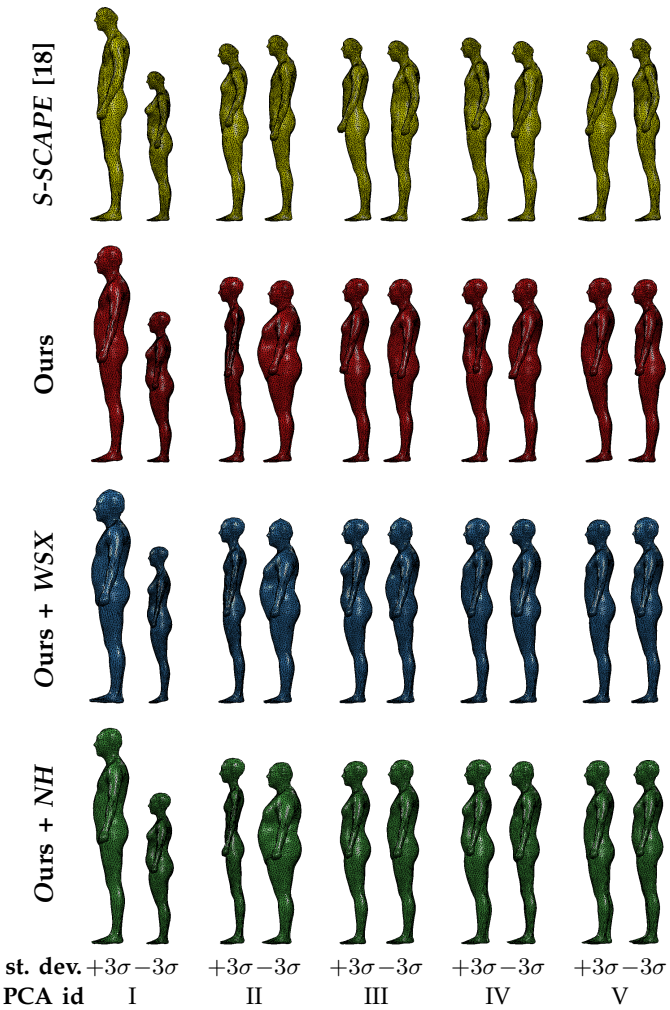


Fig. 4. Visualization of the first five PCA eigenvectors scaled by $\pm 3\sigma$ (standard deviation). Shown are eigenvectors of the *S-SCAPE* space [18] (row 1) and the *S-SCAPE* spaces trained using our processed data without (row 2) and with posture normalization using *WSX* [35] (row 3) and *NH* [20] (row 4).

the shape in both depth scans is of the same person, but the pose may differ.

6.1 Dataset and experimental setup

We use a publicly available dataset [17] containing Kinect body scans of three males and three females. Examples of the Kinect scans are shown in Fig. 6(a). For each subject, a high-resolution laser scan was captured, which is used to determine "ground truth" body shape by fitting a *S-SCAPE* space to the data. We follow the evaluation protocol of Helten et al. [17], which computes the fitting error as a difference between the results of fitting a *S-SCAPE* space to the depth data and the "ground truth" computed as a vertex-to-vertex Euclidean distance. As the required landmarks are not available for this dataset, we manually placed 14 landmarks on each depth and laser scan.

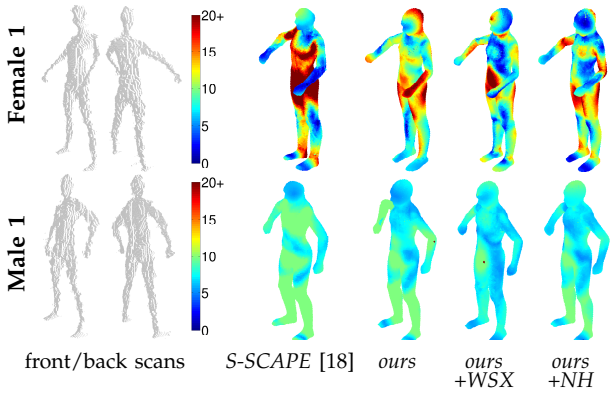


Fig. 6. Per-vertex shape fitting error (mm) of multiple methods on sample individuals of Helten et al. [17].

6.2 Quantitative evaluation

For quantitative evaluation, we compare the following four shape models presented above: the current state-of-the-art shape space [18], our shape space without posture normalization and with posture normalization using *WSX* and *NH*. In our experiments, we vary the number of shape space parameters and the number of training samples the *S-SCAPE* models are learned from. To evaluate the fitting accuracy, we report the proportion of vertices below a certain threshold.

The results are shown in Fig. 5, where the number of shape space parameters varies in the columns and the number of training samples varies in the rows. In all cases our *S-SCAPE* spaces learned from the CAESAR dataset significantly outperform the shape space by Jain et al., which is learned from the far less representative MPI Human Shape dataset. Our models achieve good fitting accuracy when using as few as 20 shape parameters, and the performance stays stable when increasing the number of shape parameters up to 50 (first row). In contrast, the performance of the shape space by Jain et al. drops, possibly due to overfitting to unrealistic shape deformations in noisy depth data. Interestingly, better performance by our models is evident even in the case when all models are learned from the same number of training samples (third and fourth rows). This shows that the CAESAR data has higher shape variability than the MPI Human Shape data. In the majority of cases, the shape space learned from the posture-normalized samples with *NH* outperforms the shape space learned from samples without posture normalization. This shows that the posture normalization method of Neophytou and Hilton [20] helps to improve the accuracy of fitting to noisy depth data. Surprisingly, the shape space learned from samples without posture normalization outperforms the shape space learned from the posture-normalized samples with *WSX* in most cases. Overall, the quantitative results show the advantages of our approach of building *S-SCAPE* spaces learned from a large representative set of training samples with additional posture normalization.

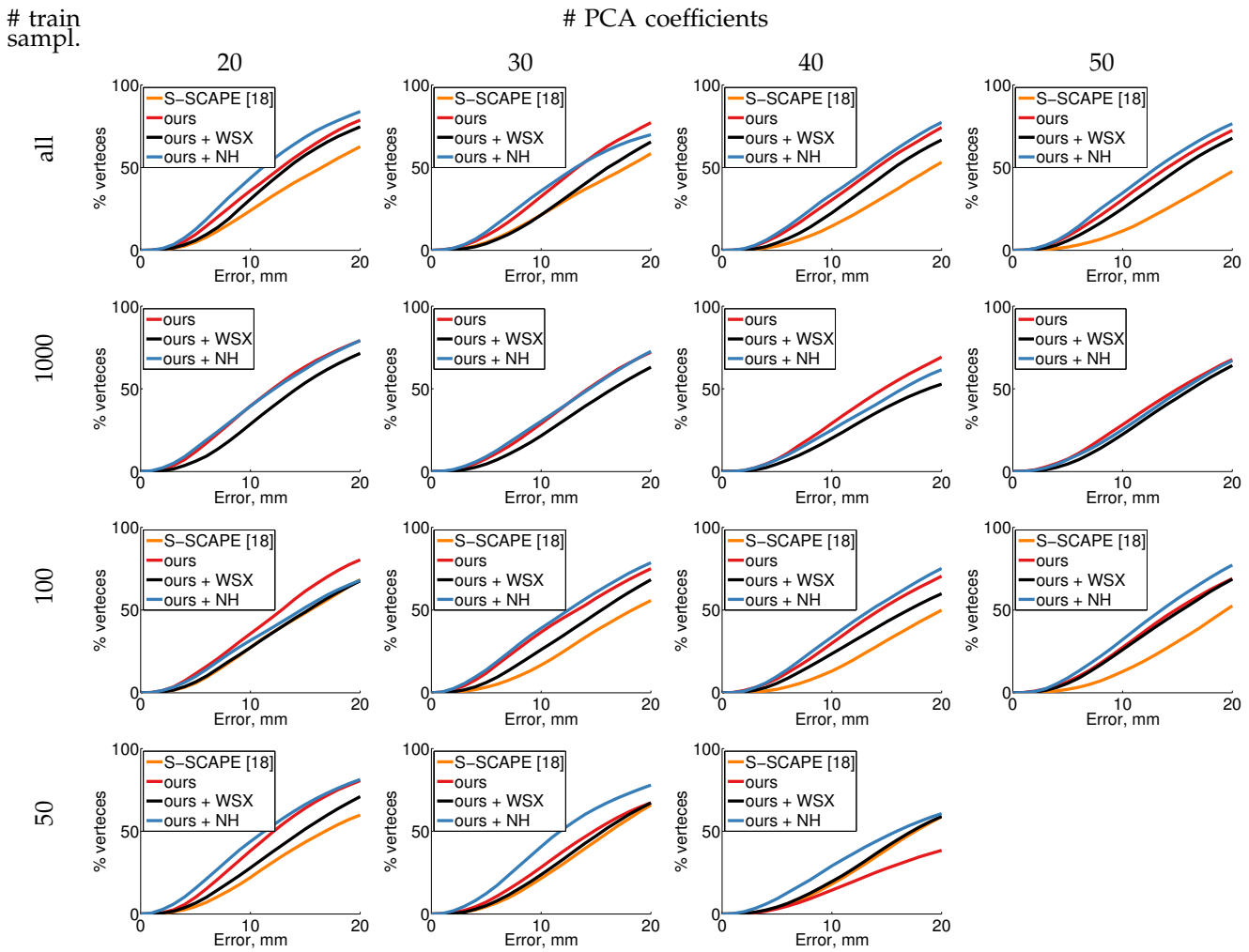


Fig. 5. Fitting error on dataset of depth scans [17] of *S-SCAPE* spaces by Jain et al. [18] and *S-SCAPE* spaces trained using our processed data without posture normalization and with posture normalization using WSX and NH. Shown is the proportion of vertices [%], for which the fitting error falls below a threshold.

6.3 Qualitative evaluation

To qualitatively evaluate the fitting, we visualize the per-vertex fitting errors. We consider the *S-SCAPE* spaces learned from all available training samples and use 20 shape space parameters. For visualization we choose two subjects, male and female, where the differences among the shape spaces are most pronounced.

Results are shown in Fig. 6. Our shape spaces better fit the data, in particular in the areas of belly and chest. This is to be expected, as we learn from the larger and more representative CAESAR dataset. Both shape spaces trained from posture normalized models can better fit the arms compared to non-normalized models.

7 Conclusion

In this work we address the challenging problem of building an efficient and expressive 3D body shape space from the largest commercially available 3D body scan dataset [24]. We carefully design and evaluate different data preprocessing steps required to obtain high-quality body shape models. To that end we evaluate different template fitting procedures. We observe that shape and posture fitting of an initial shape space to a scan prior to non-rigid deformation considerably improves the fitting results. Our findings indicate that multiple passes over the dataset improve initialization and thus increase the overall fitting accuracy and statistical shape space qualities. Furthermore, we show that posture normalization prior to learning a shape space leads to significantly better generalization and specificity of the *S-SCAPE* spaces. Finally, we demonstrate the advantages of our learned shape spaces over the state-of-the-art shape space of Jain et al. [18] learned on largest publicly available dataset [16] for the task of human body tracking from monocular depth images.

We release our *S-SCAPE* spaces, raw scan preprocessing code, code to fit a *S-SCAPE* space to a raw scan and evaluation code for public usage[1]. We believe this contribution is required for future development in human body modeling.

1. Available at *http://humanshape.mpi-inf.mpg.de*

ACKNOWLEDGEMENTS

We thank Alexandros Neophytou and Adrian Hilton for sharing their posture normalization code. This work was partially funded by the Cluster of Excellence *MMCI*.

REFERENCES

[1] B. Allen, B. Curless, and Z. Popović. The space of human body shapes: reconstruction and parameterization from range scans. *TG'03*.
[2] B. Allen, B. Curless, Z. Popović, and A. Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *SCA'06*.
[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. *TG'05*.
[4] S.-Y. Baek and K. Lee. Parametric human body shape modeling framework for human-centered product design. *CAD'12*.
[5] A. Balan and M. Black. The naked truth: Estimating body shape under clothing. In *ECCV'08*.
[6] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *CVPR'07*.
[7] F. Bogo, J. Romero, M. Loper, and M. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *CVPR'14*.
[8] J. Boisvert, C. Shu, S. Wuhrer, and P. Xi. Three-dimensional human shape inference from silhouettes: Reconstruction and validation. *MVAP'13*.
[9] Y. Chen, Z. Liu, and Z. Zhang. Tensor-based human body modeling. In *CVPR'13*.
[10] C.-H. Chu, Y.-T. Tsai, C. Wang, and T.-H. Kwok. Exemplar-based statistical model for semantic parametric design of human body. *Comp. in Ind.'10*.
[11] C. Goodall. Procrustes Methods in the Statistical Analysis of Shape. *J. R. Stat. Soc. Ser. B Stat. Methodol.'91*.
[12] P. Guan, L. Reiss, D. Hirshberg, A. Weiss, and M. Black. DRAPE: DRessing Any PErson. *TG'12*.
[13] P. Guan, A. Weiss, A. Bălan, and M. Black. Estimating human shape and pose from a single image. In *ICCV'09*.
[14] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H.-P. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *CVPR'10*.
[15] N. Hasler, C. Stoll, B. Rosenhahn, T. Thormählen, and H.-P. Seidel. Estimating body shape of dressed humans. *Comput. & Graph.'09*.
[16] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. *CGF'09*.
[17] T. Helten, A. Baak, G. Bharai, M. Müller, H.-P. Seidel, and C. Theobalt. Personalization and evaluation of a real-time depth-based full body scanner. In *3DV'13*.
[18] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. MovieReshape: tracking and reshaping of humans in videos. *TG'10*.
[19] L. Mündermann, S. Corazza, and T. Andriacchi. Accurately measuring human movement using articulated icp with soft-joint constraints and a repository of articulated models. In *CVPR'07*.
[20] A. Neophytou and A. Hilton. Shape and pose space deformation for subject specific animation. In *3DV'13*.
[21] T. Neumann, K. Varanasi, N. Hasler, M. Wacker, M. Magnor, and C. Theobalt. Capture and statistical modeling of arm-muscle deformations. *CGF'13*.
[22] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR'12*.
[23] L. Pishchulin, A. Jain, C. Wojek, T. Thormaehlen, and B. Schiele. In good shape: Robust people detection based on appearance and shape. In *BMVC'11*.
[24] K. Robinette, H. Daanen, and E. Paquet. The CAESAR project: A 3-D surface anthropometry survey. In *3DIM'99*.
[25] C. Rupprecht, O. Pauly, C. Theobalt, and S. Ilic. 3d semantic parameterization for human shape modeling: Application to 3d animation. In *3DV'13*.
[26] H. Seo and N. Magnenat-Thalmann. An automatic modeling of human bodies from sizing parameters. In *I3D '03*.
[27] H. Seo, Y. In Yeo, and K. Wohn. 3D body reconstruction from photos based on range scan. In *Edutainment'06*.
[28] L. Sigal, A. Balan, and M. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS'07*.
[29] M. Styner, K. Rajamani, L.-P. Nolte, G. Zsemlye, G. Székely, C. Taylor, and R. Davies. Evaluation of 3D correspondence methods for model building. In *IPMI'03*.
[30] G. Tam, Z.-Q. Cheng, Y.-K. Lai, F. Langbein, Y. Liu, D. Marshall, R. Martin, X.-F. Sun, and P. Rosin. Registration of 3D point clouds and meshes: A survey from rigid to non-rigid. *TVCG'13*.
[31] O. van Kaick, H. Zhang, G. Hamarneh, and D. Cohen-Or. A survey on shape correspondence. *CGF'11*.
[32] A. Weiss, D. Hirshberg, and M. Black. Home 3D body scans from noisy image and range data. In *ICCV'11*.
[33] S. Wuhrer and C. Shu. Estimating 3d human shapes from measurements. *MVAP'13*.
[34] S. Wuhrer, C. Shu, and P. Xi. Landmark-free posture invariant human shape correspondence. *The Vis. Comput.'11*.
[35] S. Wuhrer, C. Shu, and P. Xi. Posture-invariant statistical shape analysis using Laplace operator. *Comput. & Graph.'12*.
[36] P. Xi, W.-S. Lee, and C. Shu. A data-driven approach to human-body cloning using a segmented body database. In *PG'07*.
[37] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. *TG'10*.
[38] C. Zhu, R. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B Fortran subroutines for large-scale bound-constrained optimization. *TOMS'97*.

Leonid Pishchulin received his Masters degree in Computer Science from RWTH Aachen University, Germany, in 2010. Since 2010 he is a PhD student in Computer Vision and Multimodal Computing group, Max Planck Institute for Informatics, Germany. His research interests are articulated human pose estimation, articulated people detection and 3D human body shape modeling.

Stefanie Wuhrer received her M.Sc. and Ph.D. degrees in Computer Science from Carleton University in Canada in 2006 and 2009, respectively. She is a research scientist in the Morpheo team at INRIA Grenoble Rhône-Alpes. Her research interests include geometry processing, shape analysis, and human body and face modeling.

Thomas Helten , after receiving his M.Sc. (Diploma) from Bonn University in 2008, conducted research at Max Planck Institute for Informatics in analysing and reconstructing human shape and motion from multi-modal sensor input. Receiving a scholarship from JSPS, he did an internship at Keio University, Japan in 2012. He finished his Ph.D. in 2013 and took a position in industry in 2014.

Christian Theobalt is a Professor of Computer Science and the head of the research group "Graphics, Vision, & Video" at the Max-Planck-Institute for Informatics, Saarbruecken, Germany. He is also a professor at Saarland University. From 2007 until 2009 he was a Visiting Assistant Professor at Stanford University. His research lies on the boundary between Computer Vision and Computer Graphics. For instance, he works on 4D scene reconstruction, marker-less motion capture, machine learning for graphics and vision, and new sensors for 3D acquisition. Christian received several awards, for instance the Otto Hahn Medal of the Max-Planck Society (2007), the EUROGRAPHICS Young Researcher Award (2009), the German Pattern Recognition Award (2012), and an ERC Starting Grant (2013). He is also a co-founder of theCaptury (www.thecaptury.com).

Bernt Schiele received the masters degree in computer science from the University of Karl- sruhe and INP Grenoble in 1994 and the PhD degree from INP Grenoble in computer vision in 1997. He was a postdoctoral associate and visiting assistant professor with MIT between 1997 and 2000. From 1999 until 2004, he was an assistant professor with ETH Zurich and, from 2004 to 2010, he was a full professor of computer science with TU Darmstadt. In 2010, he was appointed a scientific member of the Max Planck Society and the director at the Max Planck Institute for Informat- ics. Since 2010, he has also been a professor at Saarland University. His main interests are computer vision, perceptual computing, statistical learning methods, wearable computers, and integration of multimodal sensor data. He is particularly interested in developing methods which work under realworld conditions.