

Estimation of Human Body Shape in Motion with Wide Clothing

Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, Stefanie Wuhrer

▶ **To cite this version:**

Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, Stefanie Wuhrer. Estimation of Human Body Shape in Motion with Wide Clothing. ECCV 2016 - European Conference on Computer Vision, Oct 2016, Amsterdam, Netherlands. hal-01344795v2

HAL Id: hal-01344795

https://inria.hal.science/hal-01344795v2

Submitted on 27 Jul 2016 (v2), last revised 31 Aug 2016 (v4)

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.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives| 4.0 International License

Estimation of Human Body Shape in Motion with Wide Clothing

Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer

Inria Grenoble Rhône-Alpes, France
Laboratoire Jean Kuntzmann, Université Grenoble Alpes, France
{jinlong.yang,jean-sebastien.franco,franck.hetroy,stefanie.wuhrer}inria.fr

Abstract. Estimating 3D human body shape in motion from a sequence of unstructured oriented 3D point clouds is important for many applications. We propose the first automatic method to solve this problem that works in the presence of loose clothing. The problem is formulated as an optimization problem that solves for identity and posture parameters in a shape space capturing likely body shape variations. The automation is achieved by leveraging a recent robust pose detection method [1]. To account for clothing, we take advantage of motion cues by encouraging the estimated body shape to be inside the observations. The method is evaluated on a new benchmark containing different subjects, motions, and clothing styles that allows to quantitatively measure the accuracy of body shape estimates. Furthermore, we compare our results to existing methods that require manual input and demonstrate that results of similar visual quality can be obtained.

Keywords: human body modeling, shape and motion estimation, statistical shape space

1 Introduction

Estimating 3D human body shape in motion is important for applications ranging from virtual change rooms to security. While it is currently possible to effectively track the surface of the clothing of dressed humans in motion [2] or to accurately track body shape and posture of humans dressed in tight clothing [3], it remains impossible to automatically estimate the 3D body shape in motion for humans captured in loose clothing.

Given an input motion sequence of raw 3D meshes or oriented point clouds (with unknown correspondence information) showing a dressed person, the goal of this work is to estimate the body shape and motion of this person. Existing techniques to solve this problem are either not designed to work in the presence of loose clothing [4,5] or require manual initialization for the pose [6,7], which limits their use in general scenarios. The reason is that wide clothing leads to strong variations of the acquired surface that is challenging to handle automatically. We propose an *automatic* framework that allows to estimate the human body shape and motion that is robust to the presence of *loose clothing*.

Existing methods that estimate human body shape based on an input motion sequence of 3D meshes or oriented point clouds use a shape space that models human body shape variations caused by different identities and postures as prior. Such a prior allows to reduce the search space to likely body shapes and postures. Prior works fall into two lines of work. On the one hand, there are human body shape estimation methods specifically designed to work in the presence of loose clothing [6,7]. These techniques take advantage of the fact that observations of a dressed human in motion provides important cues about the underlying body shape as different parts of the clothing are close to the body shape in different frames. However, these methods require manually placed markers to initialize the posture. On the other hand, there are human body shape estimation methods designed to robustly and automatically compute the shape and posture estimate over time [4,5]. However, these methods use strong priors of the true human body shape to track the posture over time and to fit the shape to the input point cloud, and may therefore fail in the presence of loose clothing.

In this work, we combine the advantages of these two lines of work by proposing an automatic framework that is designed for body shape estimation under loose clothing. Like previous works, our method restricts the shape estimate to likely body shapes and postures, as defined by a shape space. We use a shape space that models variations caused by different identities and variations caused by different postures as linear factors [8]. This simple model allows for the development of an efficient fitting approach. To develop an automatic method, we employ a robust pose detection method that accounts for different identities [1] and use the detected pose to guide our model fitting. To account for clothing, we take advantage of motion cues by encouraging the estimated body shape to be located inside the acquired observation at each frame. This constraint, which is expressed as a simple energy that is optimized over all input frames jointly, allows to account for clothing without the need to explicitly detect skin regions on all frames as is the case for previous methods [7,9].

To the best of our knowledge, existing datasets in this research area do not provide 3D sequences of both body shape as ground truth and dressed scans for estimation. Therefore, visual quality is the only evaluation choice. To quantitatively evaluate our framework and allow for future comparisons, we propose the first dataset consisting of synchronized acquisitions of dense unstructured geometric motion data and sparse motion capture data of 6 subjects with 3 clothing styles (tight, layered, wide) under 3 representative motions, where the capture in tight clothing serves as ground truth body shape.

The main contributions of this work are the following.

- An automatic approach to estimate 3D human body shape in motion in the presence of loose clothing.
- A new benchmark consisting of 6 subjects captured in 3 motions and 3 clothing styles each that allows to quantitatively compare human body shape estimates.

2 Related work

Many works estimate human posture without aiming to estimate body shape, or track a known body shape over time. As our goal is to simultaneously estimate body shape and motion automatically and in the presence of loose clothing, we will focus our discussion on this scenario.

Statistical shape spaces. To model human body shape variations caused by different identities, postures, and motions, statistical shape spaces are commonly used. These shape spaces represent a single frame of a motion sequence using a low-dimensional parameter space that typically models shape variations caused by different identities and caused by different postures using separate sets of parameters. Such shape spaces can be used as prior when the goal is to predict a likely body shape under loose clothing.

Anguelov et al. [10] proposed a statistical shape space called SCAPE that combines an identity model computed by performing principal component analysis (PCA) on a population of 3D models in standard posture with a posture model computed by analyzing near-rigid body parts corresponding to bones. This model performs statistics on triangle transformations, which allows to model non-rigid deformations caused by posture changes. Achieving this accuracy requires solving an optimization problem to reconstruct a 3D mesh from its representation in shape space. To improve the accuracy of the SCAPE space, Chen et al. [11] propose to combine the SCAPE model with localized multilinear models for each body part. To model the correlation of the shape changes caused by identity and posture changes, Hasler et al. [12] perform PCA on a rotation-invariant encoding of the model's triangles. These models may be used as priors when estimating human body shape in motion, but none of them allow to efficiently reconstruct a 3D human model from the shape space.

To speed up the reconstruction time from the SCAPE representation, Jain et al. [13] propose a simplified SCAPE model, denoted by S-SCAPE in the following, that computes the body shape by performing PCA on the vertex coordinates of a training set in standard posture and combines this with a linear blend skinning (LBS) to model posture changes. Any posture variations present in the training data cause posture variation to be modeled in identity space, which is known to cause counter-intuitive deformations [8]. To remedy this, recently proposed shape spaces start by normalizing the posture of the training data before performing statistics and model shape changes caused by different factors such as identity and posture as multilinear factors [14,8,15]. We use the normalized S-SCAPE model [8] in this work; however, any of these shape spaces could be used within our framework.

Recently, Pons-Moll et al. [16] proposed a statistical model that captures fine-scale dynamic shape variation of the naked body shape. We do not model dynamic geometry in this work, as detailed shape changes are typically not observable under loose clothing.

Estimation of static body shape under clothing. To estimate human body shape based on a static acquisition in loose clothing and in arbitrary posture, the following two approaches have been proposed. Balan et al. [9] use a SCAPE model to estimate the body shape under clothing based on a set of calibrated multi-view images. This work is evaluated on a static dataset of different subjects captured in different postures and clothing styles. Our evaluation on 3D motion sequences of different subjects captured in different motions and clothing styles is inspired by this work. Hasler et al. [17] use a rotation-invariant encoding to estimate the body shape under clothing based on a 3D input scan. While this method leads to accurate results, it cannot easily be extended to motion sequences, as identity and posture parameters are not separated in this encoding.

Both of these methods require manual input for posture initialization. In this work, we propose an automatic method to estimate body shape in motion.

Estimation of body shape in motion. The static techniques have been extended to motion sequences with the help of shape spaces that separate shape changes caused by identity and posture. Several methods have been proposed to fit a SCAPE or S-SCAPE model to Kinect data by fixing the parameters controlling identity over the sequence [4,5]. These methods are not designed to work with clothing, and it is assumed that only tight clothing is present.

Two more recent methods are designed to account for the presence of clothing. The key idea of these methods is to take advantage of temporal motion cues to obtain a better identity estimate than would be possible based on a single frame. Our method also takes advantage of motion cues.

Wuhrer et al. [6] use a shape space that learns local information around each vertex to estimate human body shape for a 3D motion sequence. The final identity estimate is obtained by averaging the identity estimates over all frames. While this shape space leads to results of high quality, the fitting is computationally expensive, as the reconstruction of a 3D model from shape space requires solving an optimization problem. Our method uses a simpler shape space while preserving a similar level of accuracy by using an S-SCAPE model that pre-normalizes the training shapes with the help of localized information.

Neophytou and Hilton [7] propose a faster method based on a shape space that models identity and posture as linear factors and learns shape variations on a posture-normalized training database. To constrain the estimate to reliable regions, the method detects areas that are close to the body surface. In contrast, our method constrains the estimate to be located inside the observed clothing at every input frame, which results in an optimization problem that does not require a detection.

Both of these methods require manual input for posture initialization on the first frame. Additionally, a temporal alignment is required by Neophytou and Hilton. Computing temporal alignments is a difficult problem, and manual annotation is tedious when considering larger sets of motion sequences. In contrast, our method is fully automatic and addresses both aspects.

3 S-SCAPE model

In this work, we use the S-SCAPE model as prior for human body shape changes caused by different identities and postures. While we choose this shape space, any shape space that models identity and posture as multilinear factors could be used [14,15]. Although such a simple shape space does not accurately model correlated shape changes, such as muscle bulging, it allows to effectively separate the different variations and can be fitted efficiently to input scans.

This section briefly reviews the S-SCAPE model introduced by Jain et al. [13] that allows to separate the influence of parameters controlling identity and parameters controlling posture of a human body shape. In the following, we denote by $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ the parameter vectors that influence shape changes caused by identity and posture changes, respectively. In this work, we use the publicly available posture-normalized S-SCAPE model [8], where each training shape was normalized with the help of localized coordinates [18].

In the following, let N_v denote the number of vertices on the S-SCAPE model, let $\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{\Theta}) \in \mathbb{R}^{3N_v}$ denote the vector containing the vertex coordinates of identity $\boldsymbol{\beta}$ in posture $\boldsymbol{\Theta}$, and let $\widetilde{\boldsymbol{s}}(\boldsymbol{\beta}, \boldsymbol{\Theta}) \in \mathbb{R}^{4N_v}$ denote the vector containing the corresponding homogeneous vertex coordinates. For the fixed posture $\boldsymbol{\Theta}_0$ that was used to train the identity space, S-SCAPE models the shape change caused by identity using a PCA model as

$$\widetilde{\boldsymbol{s}}(\boldsymbol{\beta}, \boldsymbol{\Theta}_0) = \widetilde{\boldsymbol{A}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\mu}}, \tag{1}$$

where $\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^{4N_v}$ contains the homogeneous coordinates of the mean body shape, $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{4N_v \times d_{id}}$ is the matrix found by PCA, and d_{id} is the dimensionality of the identity shape space. For a fixed identity $\boldsymbol{\beta}_0$, S-SCAPE models the shape change caused by posture using LBS as

$$\boldsymbol{s}_i(\boldsymbol{\beta}_0, \boldsymbol{\Theta}) = \sum_{j=1}^{N_b} \omega_{ij} \boldsymbol{T}_j(\boldsymbol{\Theta}) \widetilde{\boldsymbol{s}}_i(\boldsymbol{\beta}_0, \boldsymbol{\Theta}_0), \tag{2}$$

where $\boldsymbol{s}_i$ and $\widetilde{\boldsymbol{s}}_i$ denote the standard and homogenous coordinate vector of the i-th vertex of $\boldsymbol{s}$, N_b denotes the number of bones used for LBS, $\boldsymbol{T}_j(\boldsymbol{\Theta}) \in \mathbb{R}^{3\times 4}$ denotes the transformation matrix applied to the j-th bone, and ω_{ij} denotes the rigging weight binding the i-th vertex to the j-th bone.

Combining Eq. 1 and 2 in matrix notation leads to

$$\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \boldsymbol{T}(\boldsymbol{\Theta}) \widetilde{\boldsymbol{A}}\boldsymbol{\beta} + \boldsymbol{T}(\boldsymbol{\Theta}) \widetilde{\boldsymbol{\mu}}, \tag{3}$$

where $\boldsymbol{T}(\boldsymbol{\Theta}) \in \mathbb{R}^{3N_v \times 4N_v}$ is a sparse matrix containing the per-vertex transformations. Using this notation, it is easy to see that S-SCAPE is linear in both $\boldsymbol{\beta}$ and $\boldsymbol{T}(\boldsymbol{\Theta})$, which allows for a simple optimization w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$.

4 Estimating model parameters for a motion sequence

We start by providing an overview of the proposed method. Fig. 1 shows the different parts of the algorithm visually. Given as input a trained S-SCAPE

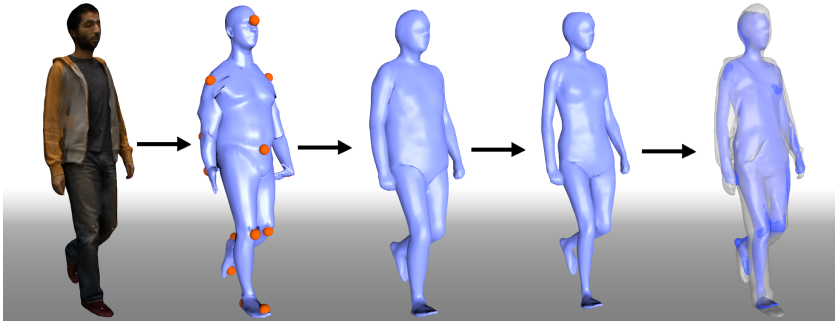


Fig. 1. Overview of the proposed pipeline. From left to right: input frame, result of Stitched Puppet [1] with annotated landmarks, result after estimation of initial identity and posture, final result, and overlay of input and final result.

model and a motion sequence consisting of N_f frames $\boldsymbol{F}_i$ represented by triangle meshes with unknown correspondence, we aim to compute a single parameter vector $\boldsymbol{\beta}$ controlling the shape of the identity (as the identity of the person is fixed during motion) along with N_f parameter vectors $\boldsymbol{\Theta}_i$ controlling the postures in each frame, such that $\boldsymbol{s}_i(\boldsymbol{\beta}, \boldsymbol{\Theta}_i)$ is close to $\boldsymbol{F}_i$.

To fit the S-SCAPE model to a single frame $\boldsymbol{F}$, we aim to minimize

$$E(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta}) = \omega_{lnd} E_{lnd}(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta}) + \omega_{data} E_{data}(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta}) + \omega_{cloth} E_{cloth}(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta}) \tag{4}$$

w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ subject to constraints that keep $\boldsymbol{\beta}$ in the learned probability distribution of parameter values. Here, ω_{lnd}, ω_{data}, and ω_{cloth} are weights that trade off the influence of the different energy terms. The energy E_{lnd} measures the distance between a sparse set of provided landmarks, which correspond to distinctive positions on the human body, to their corresponding locations on $\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{\Theta})$. The provided landmarks are computed automatically in the following. The energy E_{data} measures the distance between $\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ and $\boldsymbol{F}$ using a nearest neighbor cost. The energy E_{cloth} is designed to account for loose clothing by encouraging $\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ to be located inside the observation $\boldsymbol{F}$.

For a motion sequence of N_f frames, our goal is then to minimize

$$E\left(\boldsymbol{F}_{1:N_f}, \boldsymbol{\beta}, \boldsymbol{\Theta}_{1:N_f}\right) = \sum_{i=1}^{N_f} E\left(\boldsymbol{F}_i, \boldsymbol{\beta}, \boldsymbol{\Theta}_i\right) \tag{5}$$

w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}_{1:N_f}$ subject to constraints that keep $\boldsymbol{\beta}$ in the learned probability distribution of parameter values. Here, $\boldsymbol{F}_{1:N_f} = \{\boldsymbol{F}_1, \ldots, \boldsymbol{F}_{N_f}\}$ is the set of frames and $\boldsymbol{\Theta}_{1:N_f} = \{\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{N_f}\}$ is the set of posture parameters. The energy E_{cloth} allows to take advantage of motion cues in this formulation as it encourages the body shape to lie inside all observed frames.

In the following sections, we detail the prior that is used to constrain $\boldsymbol{\beta}$ as well as the different energy terms. Optimizing Eq. 5 w.r.t. all parameters jointly results in a high-dimensional optimization problem that is inefficient to solve

and prone to get stuck in undesirable local minima. After introducing all energy terms, we discuss how this problem can be divided into smaller problems that can be solved in order, thereby allowing to find a good minimum in practice.

4.1 Prior model for $\boldsymbol{\beta}$

A prior model is used to ensure that the body shape stays within the learned shape space that represents plausible human shapes. The identity shape space is learned using PCA, and has zero mean and standard deviation σ_i along the i-th principal component. Similarly to previous work [9], we do not penalize values of $\boldsymbol{\beta}$ that stay within $3\sigma_i$ of the mean to avoid introducing a bias towards the mean shape. However, rather than penalizing a larger distance from the mean, we constrain the solution to lie inside the hyperbox $\pm 3\sigma_i$ using a constrained optimization framework. This constraint can be handled by standard constrained optimizers since the hyperbox is axis-aligned, and using this hard constraint removes the need to appropriately weigh a prior energy w.r.t. other energy terms.

4.2 Landmark energy

The landmark energy helps to guide the solution towards the desired local minimum with the help of distinctive anatomical landmarks. This energy is especially important during the early stages of the optimization as it allows to find a good initialization for the identity and posture parameters. In the following, we consider the use of N_{lnd} landmarks and assume without loss of generality that the vertices corresponding to landmarks are the first N_{lnd} vertices of $\boldsymbol{s}$. The landmark term is defined as

$$E_{lnd}\left(\boldsymbol{F}, \boldsymbol{\beta}, \boldsymbol{\Theta}\right) = \sum_{i=1}^{N_{lnd}} \left\| \boldsymbol{s}_i\left(\boldsymbol{\beta}, \boldsymbol{\Theta}\right) - \boldsymbol{l}_i(\boldsymbol{F}) \right\|^2, \tag{6}$$

where $\boldsymbol{l}_i(\boldsymbol{F})$ denotes the i-th landmark of frame $\boldsymbol{F}$, $\boldsymbol{s}_i\left(\boldsymbol{\beta}, \boldsymbol{\Theta}\right)$ denotes the vertex corresponding to the i-th landmark of $\boldsymbol{s}\left(\boldsymbol{\beta}, \boldsymbol{\Theta}\right)$, and $\|\cdot\|$ denotes the ℓ^2 norm.

The landmarks $\boldsymbol{l}_i(\boldsymbol{F})$ are computed automatically with the help of the state of the art Stitched Puppet [1], which allows to robustly fit a human body model to a single scan using a particle-based optimization. Specifically, we once manually select a set of vertex indices to be used as landmarks on the Stitched Puppet model, which is then fixed for all experiments. To fit the Stitched Puppet to a single frame, randomly distributed particles are used to avoid getting stuck in undesirable local minima. We fit the Stitched Puppet model to frame $\boldsymbol{F}$, and report the 3D positions of the pre-selected indices after fitting as landmarks $\boldsymbol{l}_i(\boldsymbol{F})$. While the Stitched Puppet aims to fit the body shape and posture of $\boldsymbol{F}$, only the coordinates $\boldsymbol{l}_i(\boldsymbol{F})$ are used by our framework. Note that our method does not require accurate $\boldsymbol{l}_i(\boldsymbol{F})$, since $\boldsymbol{l}_i(\boldsymbol{F})$ are only used to initialize the optimization.

Using many particles on each frame of a motion sequence is inefficient. Furthermore, since the Stitched Puppet is trained on a database of minimally

dressed subjects, using many particles to fit to a frame in wide clothing may lead to overfitting problems. This is illustrated in Fig. 2. To remedy this, we choose to use a relatively small number of particles which is set to 30. Starting at the second frame, we initialize the particle optimization to the result of the previous frame to guide the optimization towards the desired optimum.

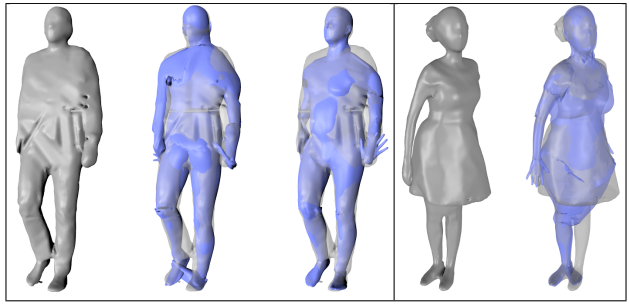


Fig. 2. Left: overfitting problem of Stitched Puppet in the presence of clothing. Input frame, Stitched Puppet result with 160 particles, and Stitched Puppet result with 30 particles are shown in order. Right: the failure case from our database caused by mismatching of Stitched Puppet.

4.3 Data energy

The data energy pulls the S-SCAPE model towards the observation $\boldsymbol{F}$ using a nearest neighbor term. This energy, which unlike the landmark energy considers all vertices of $\boldsymbol{s}$, is crucial to fit the identity and posture of $\boldsymbol{s}$ to the input $\boldsymbol{F}$ as

$$E_{data}\left(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta}\right)=\sum_{i=1}^{N_v}\delta_{\boldsymbol{NN}}\left\|\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right)-\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right),\boldsymbol{F}\right)\right\|^2, \tag{7}$$

where N_v denotes the number of vertices of $\boldsymbol{s}$ and $\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right),\boldsymbol{F}\right)$ denotes the nearest neighbour of vertex $\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right)$ on $\boldsymbol{F}$. To remove the influence of outliers and reduce the possibility of nearest neighbour mismatching, we use a binary weight $\delta_{\boldsymbol{NN}}$ that is set to one if the distance between $\boldsymbol{s}_i$ and its nearest neighbor on $\boldsymbol{F}$ is below $200mm$ and the angle between their outer normal vectors is below $60°$, and to zero otherwise.

4.4 Clothing energy

The clothing energy is designed to encourage the predicted body shape $\boldsymbol{s}$ to be located entirely inside the observation $\boldsymbol{F}$. This energy is particularly important when considering motion sequences acquired with loose clothing. In such cases, merely using E_{lnd} and E_{data} leads to results that overestimate the circumferences

of the body shape because $\boldsymbol{\beta}$ is estimated to fit to $\boldsymbol{F}$ rather than to fit inside of $\boldsymbol{F}$, see Fig. 3. To remedy this, we define the clothing energy as

$$E_{cloth}(\boldsymbol{F},\boldsymbol{\beta},\boldsymbol{\Theta})=\sum_{i=1}^{N_v}\delta_{out}\delta_{\boldsymbol{NN}}\left\|\boldsymbol{s}_i(\boldsymbol{\beta},\boldsymbol{\Theta})-\boldsymbol{NN}\left(\boldsymbol{s}_i(\boldsymbol{\beta},\boldsymbol{\Theta}),\boldsymbol{F}\right)\right\|^2+\omega_r\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\right\|^2, \quad (8)$$

where δ_{out} is used to identify vertices of $\boldsymbol{s}$ located outside of $\boldsymbol{F}$. This is achieved by setting δ_{out} to one if the angle between the outer normal of $\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right),\boldsymbol{F}\right)$ and the vector $\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right)-\boldsymbol{NN}\left(\boldsymbol{s}_i\left(\boldsymbol{\beta},\boldsymbol{\Theta}\right),\boldsymbol{F}\right)$ is below 90°, and to zero otherwise. Furthermore, ω_r is a weight used for the regularization term, and $\boldsymbol{\beta}_0$ is an initialization of the identity parameters used to constrain $\boldsymbol{\beta}$.

When observing a human body dressed in loose clothing in motion, different frames can provide valuable cues about the true body shape. The energy E_{cloth} is designed to exploit motion cues when optimizing E_{cloth} w.r.t. all available observations $\boldsymbol{F}_i$. This allows to account for clothing using a simple optimization without the need to find skin and non-skin regions as in previous work [9,19,7]. The regularization $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2$ used in Eq. 8 is required to avoid excessive thinning of limbs due to small misalignments in posture.

Fig. 3 shows the influence of E_{cloth} on the result of a walking sequence in layered clothing. The left side shows overlays of the input and the result for $\omega_{cloth}=0$ and $\omega_{cloth}=1$. Note that while circumferences are overestimated when $\omega_{cloth}=0$, a body shape located inside the input frame is found for $\omega_{cloth}=1$. The comparison to the ground truth body shape computed as discussed in Sec. 6 is visualized in the middle and the right of Fig. 3, and shows that E_{cloth} leads to a significant improvement of the accuracy of $\boldsymbol{\beta}$.

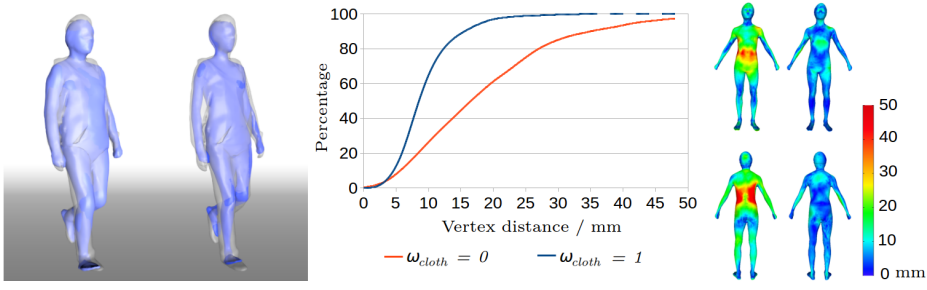


Fig. 3. Influence of E_{cloth} on walking sequence. Left: input data overlayed with result with $\omega_{cloth}=0$ (left) and $\omega_{cloth}=1$ (right). Middle: cumulative per-vertex error of estimated body shape with $\omega_{cloth}=0$ and $\omega_{cloth}=1$. Right: color-coded per-vertex error with $\omega_{cloth}=0$ (left) and $\omega_{cloth}=1$ (right).

4.5 Optimization schedule

Minimizing $E\left(\boldsymbol{F}_{1:N_f},\boldsymbol{\beta},\boldsymbol{\Theta}_{1:N_f}\right)$ defined in Eq. 5 over all N_f frames w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}_i$ jointly is not feasible when considering motion sequences containing

hundreds of frames as this is a high-dimensional optimization problem. To solve this problem without getting stuck in undesirable local minima, we optimize three smaller problems in order.

Initial identity estimation. We start by computing an initial estimate $\boldsymbol{\beta}_0$ based on the first N_k frames of the sequence by optimizing $E\left(\boldsymbol{F}_{1:N_k}, \boldsymbol{\beta}, \boldsymbol{\Theta}_{1:N_k}\right)$ w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}_i$. For increased efficiency, we start by computing optimal $\boldsymbol{\beta}_i$ and $\boldsymbol{\Theta}_i$ for each frame using Eq. 4 by alternating the optimization of $\boldsymbol{\Theta}_i$ for fixed $\boldsymbol{\beta}_i$ with the optimization of $\boldsymbol{\beta}_i$ for fixed $\boldsymbol{\Theta}_i$. This is repeated for N_{it} iterations. Temporal consistency is achieved by initializing $\boldsymbol{\Theta}_{i+1}$ as $\boldsymbol{\Theta}_i$ and $\boldsymbol{\beta}_{i+1}$ as $\boldsymbol{\beta}_i$ starting at the second frame. As it suffices for the identity parameters to roughly estimate the true body shape at this stage, we set $\omega_{cloth} = 0$. In the first iterations, E_{lnd} is essential to guide the fitting towards the correct local optimum, while in later iterations E_{data} gains in importance. We therefore set $\omega_{data} = 1 - \omega_{lnd}$ and initialize ω_{lnd} to one. We linearly reduce ω_{lnd} to zero in the last two iterations. We then initialize the posture parameters to the computed $\boldsymbol{\Theta}_i$, and the identity parameters to the mean of the computed $\boldsymbol{\beta}_i$ and iteratively minimize $E\left(\boldsymbol{F}_{1:N_k}, \boldsymbol{\beta}, \boldsymbol{\Theta}_{1:N_k}\right)$ w.r.t. $\boldsymbol{\Theta}_{1:N_k}$ and $\boldsymbol{\beta}$. This leads to stable estimates for $\boldsymbol{\Theta}_{1:N_k}$ and an initial estimate of the identity parameter, which we denote by $\boldsymbol{\beta}_0$ in the following.

Posture estimation. During the next stage of our framework, we compute the posture parameters $\boldsymbol{\Theta}_{N_k+1:N_f}$ for all remaining frames by sequentially minimizing Eq. 4 w.r.t. $\boldsymbol{\Theta}_i$. As before, $\boldsymbol{\Theta}_{i+1}$ is initialized to the result of $\boldsymbol{\Theta}_i$. As the identity parameters are not accurate at this stage, we set $\omega_{cloth} = 0$. For each frame, the energy is optimized N_{it} times while reducing the influence of ω_{lnd} in each iteration, using the same weight schedule as before. This results in posture parameters $\boldsymbol{\Theta}_i$ for each frame.

Identity refinement. In a final step, we refine the identity parameters to be located inside all observed frames $\boldsymbol{F}_{1:N_f}$. To this end, we initialize the identity parameters to $\boldsymbol{\beta}_0$, fix all posture parameters to the computed $\boldsymbol{\Theta}_i$, and minimize $E\left(\boldsymbol{F}_{1:N_f}, \boldsymbol{\beta}, \boldsymbol{\Theta}_{1:N_f}\right)$ w.r.t. $\boldsymbol{\beta}$. As the landmarks and observations are already fitted adequately, we set $\omega_{lnd} = \omega_{data} = 0$ at this stage of the optimization.

5 Implementation details

The S-SCAPE model used in this work consists of $N_v = 6449$ vertices, and uses $d_{id} = 100$ parameters to control identity and $d_{pose} = 30$ parameters to control posture by rotating the $N_b = 15$ bones. The bones, posture parameters, and rigging weights are set as in the published model [8].

For the Stitched Puppet, we use 60 particles for the first frame, and 30 particles for subsequent frames. We use a total of $N_{lnd} = 14$ landmarks that have been shown sufficient for the initialization of posture fitting [6], and are located at forehead, shoulders, elbows, wrists, knees, toes, heels, and abdomen. Fig. 1 shows the chosen landmarks on the Stitched Puppet model. During the optimization, we set $N_{it} = 6$ and $N_k = 25$. The optimization w.r.t. $\boldsymbol{\beta}$ uses analytic gradients, and we use Matlab L-BFGS-B to optimize the energy. The setting of the regularization weight ω_r depends on the clothing style. The looser the clothing, the smaller ω_r, as this allows for more corrections of the identity parameters. In our experiments, we use $\omega_r = 1$ for all the sequences with layered and wide clothing in our dataset.

6 Evaluation

6.1 Dataset

This section introduces the new dataset we acquired to allow quantitative evaluation of human body shape estimation from dynamic data. The dataset consists of synchronized acquisitions of dense unstructured geometric motion data and sparse motion capture (MoCap) data of 6 subjects (3 female and 3 male) captured in 3 different motions and 3 clothing styles each. The geometric motion data are sequences of meshes obtained by applying a visual hull reconstruction to a 68-color-camera (4M pixels) system at 30FPS. The basic motions that were captured are walk, rotating the body, and pulling the knees up. The captured clothing styles are very tight, layered (long-sleeved layered clothing on upper body), and wide (wide pants for men and dress for women). The body shapes of 6 subjects vary significantly. Fig. 4 shows some frames of the database.

To evaluate algorithms using this dataset, we can compare the body shapes estimated under loose clothing with the tight clothing baseline. The comparison is done per vertex on the two body shapes under the same normalized posture. Cumulative plots are used to show the results.



Fig. 4. Six representative examples of frames of our motion database. From left to right, a female and male subject is shown for tight, layered, and wide clothing each.

6.2 Evaluation of posture and shape fitting

We applied our method to all sequences in the database. For one sequence of a female subject captured while rotating the body in wide clothing, Stitched Puppet fails to find the correct posture, which leads to a failure case of our method (see Fig. 2). We exclude this sequence from the following evaluation.

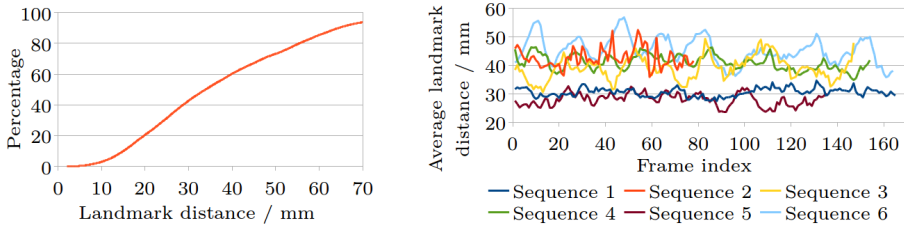


Fig. 5. Accuracy of posture estimation over the walking sequences in tight clothing. Left: cumulative landmark errors. Right: average landmark error for each sequence.

To evaluate the accuracy of the posture parameters $\boldsymbol{\Theta}$, we compare the 3D locations of a sparse set of landmarks captured using a MoCap system with the corresponding model vertices of our estimate. This evaluation is performed in very tight clothing, as no accurate MoCap markers are available for the remaining clothing styles. Fig. 5 summarizes the per-marker errors over the walking sequences of all subjects. The results show that most of the estimated landmarks are within $35mm$ of the ground truth and that our method does not suffer from drift for long sequences. As the markers on the Stitched Puppet and the MoCap markers were placed by non-experts, the landmark placement is not fully repeatable, and errors of up to $35mm$ are considered fairly accurate.

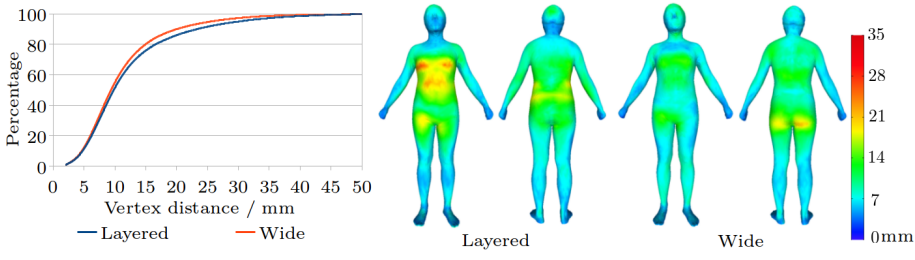


Fig. 6. Summary of shape accuracy computed over the frames of all motion sequences of all subjects captured in layered and wide clothing. Left: cumulative plots showing the per-vertex error. Right: mean per-vertex error color-coded from blue to red.

To evaluate the accuracy of the identity parameters $\boldsymbol{\beta}$, we use for each subject the walking sequence captured in very tight clothing to establish a ground truth

identity $\boldsymbol{\beta}_0$ by applying our shape estimation method. Applying our method to sequences in looser clothing styles of the same subject leads identity parameters $\boldsymbol{\beta}$, whose accuracy can be evaluated by comparing the 3D geometry of $\boldsymbol{s}\left(\boldsymbol{\beta}_0, \boldsymbol{\Theta}_0\right)$ and $\boldsymbol{s}\left(\boldsymbol{\beta}, \boldsymbol{\Theta}_0\right)$ for a standard posture $\boldsymbol{\Theta}_0$.

Fig. 6 summarizes the per-vertex errors over all motion sequences captured in layered and wide clothing, respectively. The left side shows the cumulative errors, and the right side shows the color-coded mean per-vertex error. The color coding is visualized on the mean identity of the training data. The result shows that our method is robust to loose clothing with more than 50% of all the vertices having less than $10mm$ error for both layered and wide clothing. The right side shows that as expected, larger errors occur in areas where the shape variability across different identities is high.

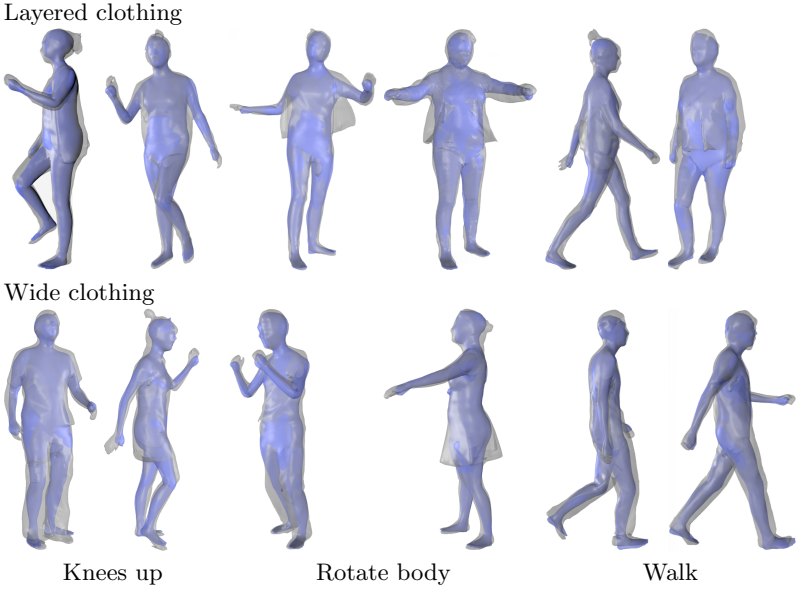


Fig. 7. Overlay of input data and our result.

Fig. 7 shows some qualitative results for all three types of motions and two clothing styles. Note that accurate body shape estimates are obtained for all frames. Consider the frame that shows a female subject performing a rotating motion in layered clothing. Computing a posture or shape estimate based on this frame is extremely challenging as the geometry of the layered cloth locally resembles the geometry of an arm, and as large portions of the body shape are occluded. Our method successfully leverages temporal consistency and motion cues to find reliable posture and body shape estimates.

6.3 Comparative evaluation

As we do not have results on motion sequences with ground truth for existing methods, this section presents visual comparisons, shown in Fig. 8. We compare to Wuhrer et al. [6] on the dancer sequence [20] presented in their work. Note that unlike the results of Wuhrer et al., our shape estimate does not suffer from unrealistic bending at the legs even in the presence of wide clothing. Furthermore, we compare to Neophytou and Hilton [7] on the swing sequence [21] presented in their work. Note that we obtain results of similar visual quality without the need for manual initializations and pre-aligned motion sequences. In summary, we present the first fully automatic method for body shape and motion estimation, and show that this method achieves state of the art results.

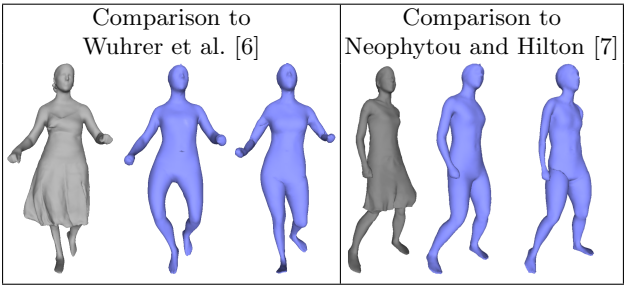


Fig. 8. Per comparison from left to right: input, result of prior works, our result.

7 Conclusion

We presented an approach to automatically estimate the human body shape under motion based on a 3D input sequence showing a dressed person in possibly loose clothing. The accuracy of our method was evaluated on a newly developed benchmark[1] containing 6 different subjects performing 3 motions in 3 different styles each. We have shown that, although being fully automatic, our posture and shape estimation achieves state of the art performance. In the future, the body shape and motion estimated by our algorithm have the potential to aid in a variety of tasks including virtual change rooms and security applications.

Acknowledgments

Funded by France National Research grant ANR-14-CE24-0030 ACHMOV. We thank Yannick Marion for help with code to efficiently fit an S-SCAPE model to a single frame, Leonid Pishchulin for helpful discussions, Alexandros Neophytou and Adrian Hilton for providing comparison data, and Mickaël Heudre, Julien Pansiot and volunteer subjects for help acquiring the database.

[1] The benchmark can be downloaded at http://dressedhuman.gforge.inria.fr/.

References

1. Zuffi, S., Black, M.: The Stitched Puppet: A graphical model of 3d human shape and pose. In: Conference on Computer Vision and Pattern Recognition. (2015) 3537–3546
2. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Conference on Computer Vision and Pattern Recognition. (2015) 343–352
3. Bogo, F., Black, M.J., Loper, M., Romero, J.: Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In: ICCV. (2015)
4. Weiss, A., Hirshberg, D., Black, M.: Home 3D body scans from noisy image and range data. In: International Conference on Computer Vision. (2011) 1951–1958
5. Helten, T., Baak, A., Bharai, G., Müller, M., Seidel, H.P., Theobalt, C.: Personalization and evaluation of a real-time depth-based full body scanner. In: International Conference on 3D Vision. (2013) 279–286
6. Wuhrer, S., Pishchulin, L., Brunton, A., Shu, C., Lang, J.: Estimation of human body shape and posture under clothing. Computer Vision and Image Understanding **127** (2014) 31–42
7. Neophytou, A., Hilton, A.: A layered model of human body and garment deformation. In: International Conference on 3D Vision. (2014) 171–178
8. Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., Schiele, B.: Building statistical shape spaces for 3d human modeling. Technical Report 1503.05860, arXiv (2015)
9. Balan, A.O., Black, M.J.: The naked truth: Estimating body shape under clothing. In: European Conference on Computer Vision. (2008) 15–29
10. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: shape completion and animation of people. ACM Transactions on Graphics **24**(3) (2005) 408–416 Proceedings of SIGGRAPH.
11. Chen, Y., Liu, Z., Zhang, Z.: Tensor-based human body modeling. In: Conference on Computer Vision and Pattern Recognition. (2013) 105–112
12. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P.: A statistical model of human pose and body shape. Computer Graphics Forum **28**(2) (2009) 337–346 Proceedings of Eurographics.
13. Jain, A., Thormählen, T., Seidel, H.P., Theobalt, C.: MovieReshape: tracking and reshaping of humans in videos. ACM Transactions on Graphics **29** (2010) 148:1–10 Proceedings of SIGGRAPH Asia.
14. Neophytou, A., Hilton, A.: Shape and pose space deformation for subject specific animation. In: International Conference on 3D Vision. (2013) 334–341
15. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.: SMPL: A skinned multi-person linear model. Transactions on Graphics **34**(6) (2015) 248:1–248:16 Proceedings of SIGGRAPH Asia.
16. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.: DYNA: a model of dynamic human shape in motion. Transactions on Graphics **34**(4) (2015) #120:1–14 Proceedings of SIGGRAPH.
17. Hasler, N., Stoll, C., Rosenhahn, B., Thormählen, T., Seidel, H.P.: Estimating body shape of dressed humans. Computers & Graphics **33**(3) (2009) 211–216 Proceedings of Shape Modeling International.
18. Wuhrer, S., Shu, C., Xi, P.: Posture-invariant statistical shape analysis using Laplace operator. Computers & Graphics **36**(5) (2012) 410–416 Proceedings of Shape Modeling International.

19. Stoll, C., Gall, J., de Aguiar, E., Thrun, S., Theobalt, C.: Video-based reconstruction of animatable human characters. ACM Transactions on Graphics **29**(6) (2010) #139:1–10 Proceedings of SIGGRAPH Asia.
20. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. ACM Transactions on Graphics **27**(3) (2008) #98,1–10 Proceedings of SIGGRAPH.
21. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. ACM Transactions on Graphics **27**(3) (2008) #97:1–10 Proceedings of SIGGRAPH.