Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Three-Dimensional Human Shape Inference from Silhouettes : Reconstruction and Validation

Jonathan Boisvert · Chang Shu · Stefanie Wuhrer · Pengcheng Xi

Received: date / Accepted: date

Abstract Silhouettes are robust image features that provide considerable evidence about the three-dimensional (3D) shape of a human body. The information they provide is, however, incomplete and prior knowledge has to be integrated to reconstruction algorithms in order to obtain realistic body models. This paper presents a method that integrates both geometric and statistical priors to reconstruct the shape of a subject assuming a standardized posture from a frontal and a lateral silhouette. The method is comprised of three successive steps. First, a non-linear function that connects the silhouettes appearances and the body shapes is learnt and used to create a first approximation. Then, the body shape is deformed globally along the principal directions of the population (obtained by performing principal component analysis over 359 subjects) to follow the contours of the silhouettes. Finally, the body shape is deformed locally to ensure it fits the input silhouettes as well as possible. Experimental results showed a mean absolute 3D error of 8mm with ideal silhouettes extraction. Furthermore, experiments on body measurements (circumferences or distances between two points on the body) resulted in a mean error of 11mm.

Keywords Human models · statistical prior · shape-from-silhouettes · three-dimensional reconstruction

J. Boisvert, C. Shu, S. Wuhrer and P. Xi
National Research Council Canada
Institute for Information Technology
1200 Montreal Road, Ottawa, Canada

Corresponding author : J. Boisvert
Tel.: +1-613-991-3388
Fax: +1-613-952-0215
E-mail: jonathan.boisvert@nrc-cnrc.gc.ca

1 Introduction

Three-dimensional human shape models are instrumental to a large number of applications. Special effects, video games, ergonomic design and biomedical engineering are just a few examples of industries that rely heavily on those models. Depending on the application at hand, the requirements (accuracy, cost, speed, *etc.*) vary tremendously and, consequently, so must the technology used to perform the reconstruction.

Active vision systems (laser or structured light scanners, for example) can produce highly accurate models when the use of specialized hardware is acceptable and stereo-reconstruction methods [26] can be used to produce dense and accurate models without specialized hardware when image quality and experimental conditions are thoroughly controlled. However, there are applications (garment fitting being one example) where tolerance to imperfect experimental conditions and cost make traditional approaches unacceptable.

Silhouettes-based reconstruction methods, on the other hand, are much more resilient to imperfect experimental conditions. Silhouettes are generally simpler to extract than other image features (especially when background subtraction can be used) and they provide significant cues about the 3D shape of the imaged object. Moreover, extracting silhouettes does not require special equipment and can be performed using low-cost digital cameras.

Unfortunately, different objects can cast the same silhouettes and, thus, reconstructing an object's 3D shape from its silhouette(s) is an ill-defined problem. The inherent uncertainty linked to the reconstruction of 3D models from their silhouettes can, however, be alleviated by the use of prior knowledge. Instead of computing the shape of an object based on its silhouette(s),

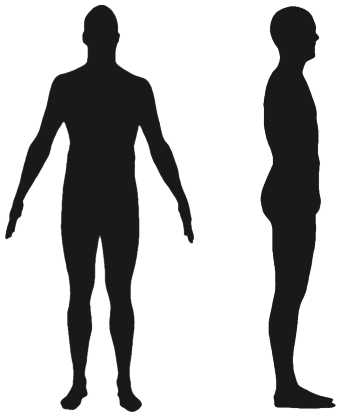


Fig. 1: Typical silhouettes casted by a human subject. Left: Front silhouette. Right: Lateral silhouette.

we can compute the shape of the most likely object given the observed silhouettes if a statistical model of the object's shape is known beforehand.

Human body shapes represents a good example of a family of shapes where great variability can be observed, but where the set of admissible shapes is also highly constrained. Height, weight, musculature, sex, *etc.* contribute to the great diversity of shapes that can be observed, but basic anatomy greatly limits the variations that do occur. In this context, a statistical model of the human body shape is likely to constrain the reconstruction from silhouettes enough to achieve good accuracy, ease of use and speed.

In this paper, we tackle the specific problem of reconstructing human bodies in a standardize posture using two silhouettes. Figure 1 presents an example of these inputs. We choose a single standardized posture to minimize the potential causes of error in the body shape reconstruction. This does mean, however, that subjects need to be compliant for the reconstruction algorithm to be applicable. Human pose estimation and body shape reconstruction in spite of varying postures are very interesting problems, but are different topics.

The contributions of this paper are two-fold. First, we present a new reconstruction method that combines three ideas that were used in isolation in the past. In the proposed approach, an initial estimation is generated using a mapping learnt from examples of associations between 3D models and their silhouettes. The shape is then refined by searching within typical variations of the human body which one is the most likely candidate. Then, since every individual is unique, the final model is optimized locally to ensure the final model explains as much of the silhouette contour as possible. This combination has several advantages:

- it is largely insensitive to convergence problems since the functional mapping ensure the optimization processes start close to the final solution,
- it minimizes an intuitive and image-based error metric, and
- it is not limited to shape variations seen in a predefined databases.

Second, we propose an extensive validation of the method using both synthetic experiments where a ground truth was directly available and real-world experiments where ground truth was only available for derived measurements. To our knowledge it is the most extensive validation of a method that reconstructs 3D model of the human body based on two silhouettes.

2 State of the art: Reconstructing human body shape from silhouettes

The appeal of shape-from-silhouettes methods is simple to understand. They only require geometrically calibrated cameras and a way to separate the silhouettes from the background (see [16,12,31,21,25] and reference therein). No assumptions have to be made about the object's reflectance (i.e. the visible surfaces do not have to be Lambertian), lights positions, color balance, *etc.* However, it is impossible to reconstruct from silhouette(s) with absolute certainty, since different objects can cast the same silhouettes.

One solution is to simply seek the largest 3D shape that can be explained by a set of silhouettes. The resulting shape is then called the *visual hull*; a concept defined by Laurentini [20]. Unfortunately, the *visual hull* will not in general tend asymptotically to the object's shape when the number of viewpoints is increased. It is bounded by the convex hull of the object in most cases (that is when no cameras are installed inside the convex hull of the object).

It is possible to go beyond the theoretically limited accuracy of the *visual hull* by using more image-based information. Silhouettes can be combined with color information. For example, it is possible to check for color consistency between the cameras to further refine reconstructed shape beyond the visual hull, which is the basic idea behind space carving [19]. The same idea can be integrated in methods that aim at reconstructing human shapes. Cheung et al. [8,7,9] use what they called colored surface points (which are photo-consistent points at the surface of the visual hull) to reconstruct moving objects and tracking human beings. Using more image information is a perfectly valid choice, one has, however, to be very conscious about it since it generally

means making more assumptions about the object's reflectance and the experimental conditions.

2.1 Human Body Models

It is entirely possible to obtain more accurate estimates of human shapes than their *visual hull* without using other image features. It is, however, necessary to use a smarter criterion than selecting the largest possible shape. The general idea is to integrate more prior knowledge about what makes a 3D shape look human.

One possibility is to consider the human body as an articulated model or a kinematic chain. The general shape of the model can then be expressed by the angles (and/or positions) of the different articulations. Delamarre and Faugeras [10], for instance, used forces to attract an articulated made of parallelepipeds, spheres and truncated cones to silhouettes contours. Mikić et al. [23] used an analogous approach where an articulated model made of ellipsoids and cylinders was drawn to the *visual hull*. Similarly, Kakadiaris and Metaxas [17] used a part decomposition algorithm to parameterize the human body with superellipsoids.

Those articulated models compactly represents the general shape of the subjects (especially moving subjects). Furthermore, basic measurements such as height or shoulder width can be done on them. However, finer details such as musculature or body fat are generally filtered out by this kind of representation. It is thus necessary to adopt a richer body shape description when various body shape measurements are expected to be important, especially when a standardized posture already mitigate the usefulness of articulated modeling.

A dense mesh can be used to describe the body shape. The number of parameters to estimate becomes, however, very large unless it is deformed only via the principal modes obtained by applying principal component analysis over a sufficiently large database of subjects. This idea was applied by Allen et al. [2] to model synthesis. Moreover, the same idea can be useful to reconstruct 3D models from silhouettes. Seo et al. [27] used it with two orthogonal silhouettes while Xi et al. [33] worked with a single frontal silhouette.

Both ideas of articulated modeling and dense shape representation based on a statistical model can also be combined. That was the idea of the SCAPE model published by Anguelov et al. [3] and later applied to reconstruction of 3D human body model from silhouettes images [5,30]. Hasler et al. [15] also presented recently a method to combining both shape and pose using a bilinear model. The added functionality of such a model comes, however, with a higher number of degrees of freedom (which in turn makes the model estimation problem more difficult).

2.2 Model estimation

Selecting a proper model for the body shape is of great importance, but another critical matter is the estimation method. Indeed, a great model is useless if its estimation is intractable in practice. Among the methods previously published in the literature, we distinguish two large families of estimation methods.

First, iterative methods that optimize the similarity between the observed silhouettes and simulated silhouettes of the reconstructed body model. The similarity measure can be defined in various ways. For instance, Delamarre and Faugeras [10] defined it using the silhouettes' contours, Balan et al. [5] used the overlap between observed and simulated silhouetttes while Mikić et al. [23] used a voxel representation of observed silhouettes to achieve the same purpose. The non-linear optimization method used to maximize similarity also varies considerably: Mikić et al. [23] used extended Kalman filtering, Balan et al. [5] stochastic optimization and Seo et al. [27] a direction set method.

Second, instead of proceeding by iteratively improving an initial estimation, some methods learn a multivariate function that associates silhouettes features and 3D shapes using a set of training examples. For example, Agarwal and Triggs [1] used relevance vector machines (RVMs) to recover the pose of a human body from a single photograph. The approach uses the histogram of shape context as feature space for the silhouettes. Gond et al. [13] also used RVMs to predict the pose of a human body, but operates on a set of n silhouettes. The feature space in this case is given by a voxel distribution in a cylinder centered on the center of the body mass.

Xi et al. [33] aim to estimate the human body shape in a fixed pose based on a given silhouette. Starting from a parameterized database of human meshes, the approach performs PCA of the 3D body shape and the 2D silhouette contour. The approach then computes a linear mapping from the PCA space of the silhouette data to the PCA space of the 3D data. Chen and Cipolla [6] later proposed a similar method where the functional mapping between the silhouettes and the 3D shapes was learnt using a Shared Gaussian Process Latent Variable Model (SGPLVM) [28]. Given a new silhouette, these approaches map the silhouette into silhouette PCA space and use the SGPLVM to map to the PCA space of the 3D meshes. This allows to compute the new body shape. Ek et al. [11] use a similar

approach to estimate the pose of a human body based on a given silhouette.

Sigal et al. [30] predict human pose and body shape from a single image. They encode the prior model using the SCAPE model [3]. They proceed by encoding the image silhouette using a histogram of shape context descriptor and by learning a mapping from this descriptor to the SCAPE parameters of the models using kernel linear regression.

2.3 Validation

It is essential that the precision and accuracy of 3D human models reconstructed using a shape-from-silhouettes approach be studied, if those models are going to be used in any practical applications. However, most of the shape-from-silhouettes methods applicable to human bodies were proof-of-concepts and thus were not extensively validated. For instance, Seo et al. [27] demonstrated only their method on one example. Xi et al. [33] used 24 subjects in a synthetic experiment to determine the mean error associated with their method and demonstrated the results on two individuals with real images. Balan et al. [5] reported the overlap between the real silhouettes and silhouettes simulated from reconstructed models in one sequence. They also validated one measurement (the height) on one subject. Sigal et al. [30] reported validation for two measurements (height and arm span) on two subjects.

In this paper, we propose a more extensive validation on real and synthetic experiments. Synthetic experiments will allow us to characterize the proposed method in ways that would be extremely difficult with real subjects (response to image noise, for instance). Validation experiments with a set of 14 common body measurements will also be presented, which is important since most practical applications rely primarily on measurements.

3 Reconstruction of Human Body Models from Silhouettes

In light of the literature review and of the boundaries of our exact problem, we designed a new method that reconstructs 3D human body models from silhouettes. This method does not use any form of articulated model, since our subjects are compliant and their posture normalized. The method proceeds by learning a relationship between a set of silhouettes and the body shapes using a data base of n 3D models modeled as triangular manifold meshes. Let X_n denote the 3D models used for training and let S_n^k denote the corresponding silhouettes, where k is the number of silhouettes used for the reconstruction. In this work, we use two sets of silhouettes ($k = 2$), namely the front and side views of the models. The method is, however, general enough to handle any combination of silhouettes. Given a set of silhouettes $S^0, \ldots, S^{k-1}$ of a human shape that correspond to the same views used in the training data, the method uses this statistical model to predict a new shape X. Then, this shape X is refined iteratively by minimizing an image-based cost function along the principal components of a 3D statistical model. Finally, fine-grained adjustment are made to the model X to ensure its silhouettes are as close as possible to the input silhouettes.

3.1 Learning the distribution of 3D human shapes and 2D human silhouettes

In order to uncover the statistical relationship between human bodies and human body silhouettes, we started with the Civilian American and European Surface Anthropometry Resource (CAESAR) dataset. CAESAR collected thousands of range scans as well as 3D anthropometric landmarks from volunteers aged from 18 to 65 in Europe and North America [24]. The anthropometric landmarks were manually collected by experts and are usually associated with bone markers that are close to the skin surface. The range scans can include holes and are not readily comparable because their parameterization can differ substantially.

They therefore need to be processed before further uses. We thus fitted a generic template to 359 range scans from this dataset using an approach proposed by Xi et al. [32] in order to obtain a consistent parameterization for all models. Xi et al.'s approach takes advantage of the anthropometric landmarks to compute an initial alignment using an RBF kernel. Then, finer alignment is obtained by minimizing a non-linear cost function that combines fitting error and transformation smoothness.

Silhouettes were then simulated by rendering the 3D models with the desired camera parameters. A consistent parameterization of the silhouettes is attained by sampling the contours and using the projection of predefined anthropometric landmarks to ensure the same number of points are sample for all subjects in a given section of the silhouette. For each training subject, the parameterized silhouettes in different views are concatenated into one single vector for training.

Then, we performed Principal Component Analysis (PCA) of the 3D models X_i. We denote the shape space of the 3D models by $\mathcal{S}^{3D}$. In PCA space, each shape X_i is represented by a vector $W_i^{(X)}$ of PCA weights.

PCA yields a mean shape $\mu^{(X)}$ and a matrix $A^{(X)}$ that can be used to compute a new shape X_{new} based on a new vector of PCA weights $W_{new}^{(X)}$ as $X_{new} = A^{(X)}W_{new}^{(X)} + \mu^{(X)}$. This same matrix can also be used to compute the PCA weights of a new shape X_{new} as $W_{new}^{(X)} = A^{(X)T}(X_{new} - \mu^{(X)})$.

Furthermore, we performed PCA of the concatenated silhouettes S_i. We denote the shape space of the silhouettes by $\mathcal{S}^{2D}$. In PCA space, each concatenated silhouette S_i is represented by a vector $W_i^{(S)}$ of PCA weights. We denote the mean and matrix corresponding to this shape space by $\mu^{(S)}$ and $A^{(S)}$.

3.2 3D Body Shape Regression

We learn a functional mapping between $\mathcal{S}^{2D}$ and $\mathcal{S}^{3D}$ that is similar to the approach by Chen and Cipolla [6]. That is, we compute a mapping from the PCA space of the silhouette data to the PCA space of the 3D data using a Shared Gaussian Process Latent Variable Model (SGPLVM) [28].

The Gaussian Process Latent Variable Model is effective to model a high dimensional data set lying on a low-dimensional manifold in a probabilistic way. The model automatically extracts a set of low-dimensional latent variables of an object from a high-dimensional observation space. SGPLVM is a variant, which extends to multiple observations that share the same underlying latent structure.

We are given n pairs of observations as training data, namely $[(W_0^{(S)}, W_0^{(X)}),$
$(W_1^{(S)}, W_1^{(X)}), \ldots, (W_{n-1}^{(S)}, W_{n-1}^{(X)})]$. SGPLVM computes a set of latent variables $L = [l_0, l_1, \ldots, l_{n-1}]$ that describe the manifold containing the observations, where l_i controls the pair $(W_i^{(S)}, W_i^{(X)})$ (we used source code by Ek et al. [11] to perform this operation). The observations in $\mathcal{S}^{2D}$ and $\mathcal{S}^{3D}$ are conditionally independent given the latent structure L. Once the mapping is known, it can be used to predict a new latent point l given a new observation $W^{(S)}$ in $\mathcal{S}^{2D}$.

The new latent point l is then used to predict the new observation $W^{(X)}$ in $\mathcal{S}^{3D}$ that corresponds to $W^{(S)}$. A first approximation X_{init} of the 3D body model can then be easily computed since $X_{init} = A^{(X)}W^{(X)} + \mu^{(X)}$. In our experiments, we used 50 dimensions for the PCA spaces and 20 shared dimensions in the latent space.

In theory, the latent space can be multi-modal and therefore multiple solutions could be generated by this method. It could happen if, for example, we had people randomly deciding to either face the camera or to turn their back to it. In a case like this, there might be two modes in latent space. One would correspond to a forward facing subject and a second to a person facing away. Selecting the wrong mode would result in larger reconstruction errors. Fortunately, our experimental setup is standardized in such a way that this type of problem never occurred.

3.3 Maximum a-posteriori shape estimation

Regression-based methods have the advantage of not requiring an initial guess to operate correctly. However, they don't maximize explicitly the agreement between the input silhouettes and the silhouettes casted by the estimated model. Their result can thus often be refined further using an iterative method. Furthermore, it is also desirable to take into account the subjects' positions and orientations, since even compliant subjects do not always align themselves perfectly with the cameras.

We thus wish to find the shape X and rigid transformation T that maximizes the posterior probability $p(T(X)|S^0, \ldots, S^{k-1})$. This probability is unfortunately difficult to model directly. However, using Bayes' theorem and since the silhouettes $S^0, \ldots, S^{k-1}$ are constant for each reconstruction, we can show that maximizing the posterior probability is equivalent to maximizing $\log(p(S^0, \ldots, S^{k-1}|T(X))p(T(X)))$.

The likelihood $p(S^0, \ldots, S^{k-1}|T(X))$ is modeled using the distances from the parameterized simulated silhouettes contours of the current model estimate to the input silhouettes contours. More precisely, the current shape estimate $T(X)$ is projected to the image planes π_j and vertices p_i of $T(X)$ that lie on the silhouettes contours are identified. The indices of those points are regrouped in a set noted $Sil(\pi_j)$. Then, the distances from the projection of vertices p_i (noted $p_i^{\pi_j}$) to their nearest neighbors on the input silhouettes $NN^{S^j}(p_i^{\pi_j})$ are summed. This operation is summarized by the following equation :

$$\log(p(S^0, \ldots, S^{k-1}|T(X))) \propto \sum_{j=0}^{k-1} \sum_{i \in Sil(\pi_i)} \|p_i^{\pi_j} - NN^{S^j}(p_i^{\pi_j})\|^2.$$

The prior $p(T(X))$ was modeled as a multivariate Gaussian distribution over the 3D points of the body shape. The parameters of that Gaussian distribution are those obtained in section 3.1 using PCA on aligned 3D body shapes, which means it is given by the following equation:

$$\log(p(X)) \propto {W^{(X)}}^T W^{(X)}.$$

In order to make the problem numerically tractable, we perform the optimization by adjusting only the first n PCA weights of the vector $W^{(X)}$. Which means we have the following optimization problem :

$$\{W^{(X)}, T\} = \underset{W^{(X)},T}{\operatorname{argmax}} \gamma {W^{(X)}}^T W^{(X)} + \sum_{j=0}^{k-1} \left(\sum_{i \in Sil(\pi_i)} \|p_i^{\pi_j} - NN^{S^j}(p_i^{\pi_j})\|^2 \right),$$

where γ governs the weight of the prior with respect to the importance given to the likelihood and where $W^{(X)}$ has non-zero values only in its first n components. The rigid transformation T is implicitly part of the cost function since it is used to create the projected points $p_i^{\pi_j}$. When the optimization is completed, we can compute the current body shape $X_{MAP} = T(AW^{(X)} + \mu^{(X)})$.

3.4 Silhouettes-shape similarity optimization

This step aims to deform the current shape estimate to fit the given silhouettes $S^0, \ldots, S^{k-1}$ while allowing the human shape to leave the learned shape space $\mathcal{S}^{3D}$ if necessary. This may be required if either the shape or the posture of the person described by $S^0, \ldots, S^{k-1}$ exhibit a variability not present in the training data. This step can only be used if the current shape estimate already fits the silhouettes well overall. Hence, we start with $X_{current} = X_{MAP}$ as current shape estimate.

Since we aim to deform $X_{current}$ to fit the given silhouettes $S^0, \ldots, S^{k-1}$, we start by identifying the vertices p_i of $X_{current}$ that lie on the silhouettes. This is achieved by projecting $X_{current}$ to the image planes π_j in order and by finding the vertices that project to the silhouettes. Let $p_i^{\pi_j}$ denote the projection of p_i to π_j and $Sil(\pi_j)$ the set of indices of points that project to the silhouette in image plane π_j. In the following, we only consider the indices i in $Sil(\pi_j)$. We find the nearest neighbor of $p_i^{\pi_j}$ in S^j and we denote this point by $NN^{S^j}(p_i^{\pi_j})$. The shape $X_{current}$ fits the given silhouettes perfectly if $p_i^{\pi_j}$ and $NN^{S^j}(p_i^{\pi_j})$ are identical. Hence, we aim to move p_i such that its projection $p_i^{\pi_j}$ is close to $NN^{S^j}(p_i^{\pi_j})$.

There is a unique best direction in which to move $p_i^{\pi_j}$ to get as close as possible to $NN^{S^j}(p_i^{\pi_j})$ in π_j. However, due to the projection, there is no unique best direction in which to move p_i. Hence, we restrict the movement of p_i to be along the surface normal $\mathbf{n}_i$ of $X_{current}$ at p_i. We can now formulate the problem of moving p_i as an optimization problem, where we aim to find an offset o_i such that the projection of $p_i + o_i\mathbf{n}_i$ to π_j is as close as possible to $NN^{S^j}(p_i^{\pi_j})$.

We can solve this problem by minimizing an energy function. We project the normal vector $\mathbf{n}_i$ at p_i to π_j and denote this projection by $\mathbf{n}_i^{\pi_j}$. We now aim to minimize

$$E_{sil} = \sum_{j=0}^{k-1} \sum_{i \in Sil(\pi_j)} \|p_i^{\pi_j} + o_i\mathbf{n}_i^{\pi_j} - NN^{S^j}(p_i^{\pi_j})\|^2.$$

If we only minimize E_{sil}, then only some vertices of $X_{current}$ will move. This will lead to a non-smooth and highly non-human shape. Hence, we also consider the following smoothness term

$$E_{smooth} = \sum_{i=0}^{m-1} \sum_{j \in N_1(p_i)} (o_i - o_j)^2, \quad (1)$$

where $N_1(p_i)$ is the one-ring neighborhood of p_i in $X_{current}$.

We minimize $E = (1-\lambda)E_{sil} + \lambda E_{smooth}$ for the fixed nearest neighbors $NN^{S^j}(p_i^{\pi_j})$, where λ is a weight with $0 < \lambda < 1$ (in practice we used $\lambda = 0.5$). We then repeat the computation of the nearest neighbors and the energy minimization until the difference in E_{sil} no longer changes significantly.

3.5 Implementation details

The reconstruction method described above was designed to be accurate, efficient and intelligible. All the details needed for a general implementation of the method have therefore already been provided. However, some details of the implementation are worth describing since they contribute to the quality of the results for our specific application.

First of all, the method relies on the ability to efficiently determine which vertices of a 3D model were projected to the contour of the silhouettes. This operation has to be performed for each evaluation of the cost functions presented in sub-sections 3.3 and 3.4, speed is thus critical. There are many valid solutions to this problem, but we observed in practice that the simplest and fastest method was to: render the 3D model, retrieve the z-buffer [29], convert the z-buffer content back to 3D positions, and finally to search for the nearest vertices using a kd-tree [4].

Second, we use the limited-memory Broyden-Fletcher-Goldfarb-Shanno scheme [22] for all the non-linear optimization procedures. This choice worked well in practice, but other minimization methods could very well be considered for other specific applications.

Third, since the resolution at the fingers of the model used in our implementation is low, the effect of the smoothing term of Equation 1 is too strong in this area. That is, fingers tend to get unrealistically fat (collapse, respectively) when some points on the silhouette move in direction of the outer normal (inner normal, respectively). Hence, we do not move the points on the fingers of the template model during the last step of the algorithm.

Finally, the position of the cameras as well as the posture of the subjects were selected to avoid seeing the arms on the side silhouettes. That decision reduces the variability of the silhouettes shapes seen on the side silhouettes and makes the functional mapping between the silhouettes and the 3D model more specific. Unfortunately, the CAESAR dataset was not designed with this constraint. Thus, the 3D statistical model created using the CAESAR dataset integrates too much variability with respect to the arms orientation as seen from the side view. We compensated for this by simply excluding the arms when the current model estimation has to be rendered from the side view to create virtual silhouettes.

4 Validation Method and Experimental Results

In order to use the proposed method for practical applications, it is important to characterize the quality of the reconstructions that it provides. Doing so is, however, more difficult than it appears. Human subjects move constantly and cannot reliably reproduce a pose (even a neutral one), thus comparison between multiple acquisitions are at best difficult to perform.

We thus decided to first use high-quality computerized models (from the CAESAR dataset) to perform synthetic experiments. That allowed us to design tests where the ground truth was known with absolute certainty. Furthermore, it also enabled us to isolate factors that are intrinsic to the reconstruction procedure presented in this paper and not factors that are primarily linked to the experimental setup used to collect the silhouettes.

Applications have different needs which cannot be quantified using a single standardized error metric. We therefore validated the proposed method using three different strategies: measuring the three-dimensional error between a known 3D model and a 3D reconstruction, measuring the influence of silhouettes extraction errors on the quality of the reconstructed model, and, finally, collating measurements realized on the reconstructed models (from real images) and measurements obtained using a different source.

4.1 Three-dimensional comparisons

We selected a total of 220 human subjects from the CEASAR database [24] for validation purposes. These models are high-accuracy surface models of human subjects assuming a natural, but standardized posture. Moreover, none of these subjects were used to create the statistical model used in the proposed method.

Then, we created front and side silhouettes for all the selected models. To do so, we simply rendered the models with ambient light only. Those silhouettes images were then used as inputs to reconstruct 3D models.

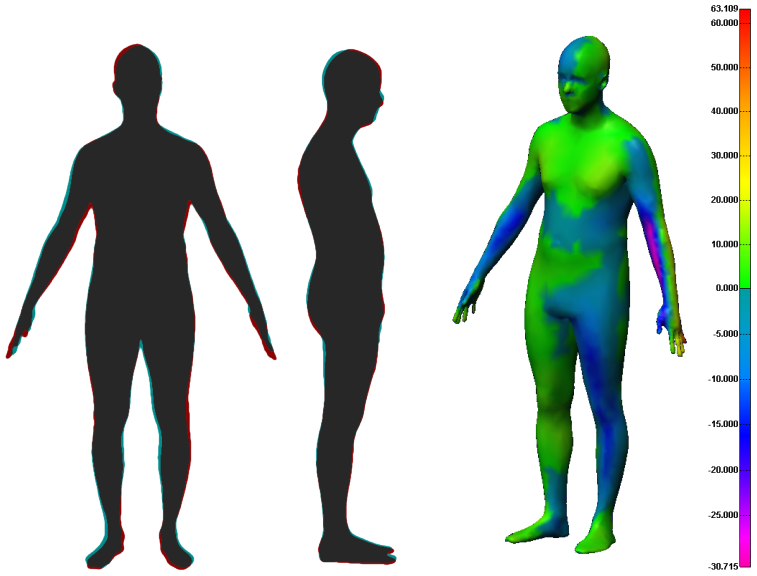
Because the original 3D models are known with high precision, it is possible to compare both the original and its corresponding reconstruction to analyze reconstruction errors. The final mean absolute error for all models was 8mm (it was 15mm after 3D body shape regression and 9mm after MAP estimation). Figure 2 presents an example of the difference between the original 3D model and its reconstruction. The absolute three-dimensional error is generally below 10mm, but higher values are observed on the front and back edge of both arms as well as in the hands regions. Arms and hands are not visible on the side view thus larger errors were indeed expected.

The reconstructed models were all comprised of 60,000 triangles and, generally, about a thousand vertices were selected by the nearest neighbor searches at each iteration to compute the cost functions associated with the MAP estimation and the silhouettes-shape similarity optimization.

The processing time required to obtain the final models did not vary considerably. On average, the 3D body shape regression took 6 seconds, the MAP estimation 30 seconds and the silhouettes-shape similarity optimization 3 minutes (on a Intel Core i7 CPU cadenced at 3GHz). It is important to stress that those processing times are from single-threaded implementation, which could be greatly greatly optimized.

As the example of Figure 2 suggests, some regions of the anatomy are better reconstructed than others. This is very important since some applications are more tolerant to errors in certain areas. Thus, even if the global error measure seems acceptable, it is crucial to make sure the distribution of errors on the body is also acceptable.

To investigate this matter more closely, we reconstructed the 220 subjects already selected for validation using the proposed method. We computed the reconstruction error at each location for all subject and averaged it. In this context, the reconstruction error is defined as the distance from a given vertex in the reconstructed model to its nearest neighbor in the original



(a) (b) (c)

Fig. 2: Comparison of an original model with the corresponding reconstruction. Input silhouettes are printed in red, silhouettes of the reconstructed model in cyan and superpositions of both in grey. (a) Front view. (b) Side view. (c) 3D Reconstruction with color-coded error (in millimeters).

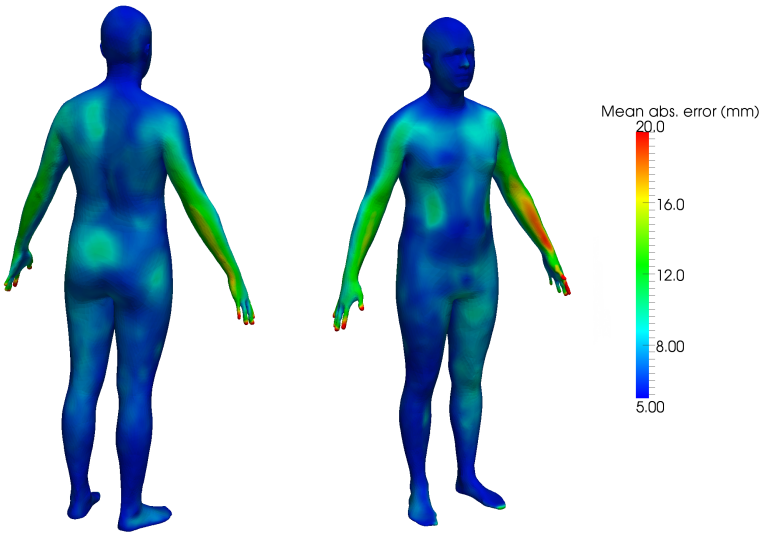


Fig. 3: Localization on the human body of the reconstruction errors. The color at each point of this model illustrates the mean absolute error at this anatomical location for all the subjects used in the validation process.

model. Figure 3 presents the distribution of reconstruction errors on a template model.

Once again, we can observe that error is concentrated on the back and front edges of the arms and around the hands. We can also observe slightly higher errors on the left side of the subjects, which is often occluded since the lateral silhouettes are captured from a camera located on the right side of the subjects.

4.2 Silhouettes extraction errors

An obvious factor that influences the quality of the reconstructions obtained by any shape-from-silhouette method is the quality of the aforementioned silhouettes. If the silhouettes have to be absolutely perfect, then the algorithm is of little use.

To demonstrate the robustness of the shape-from-silhouette reconstruction, we added first order auto-

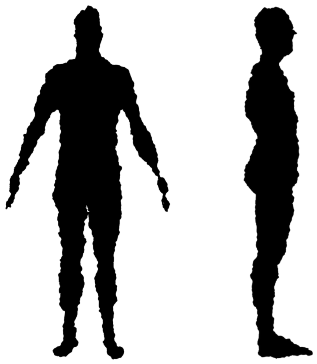


Fig. 4: Example of noise corrupted silhouette ($\sigma_\epsilon = 3$)

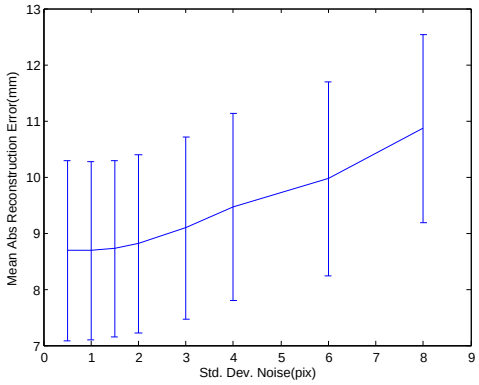


Fig. 5: Influence of an auto-regressive noise added to input silhouettes on the accuracy of the reconstruction body shape.

regressive Gaussian noise to the parameterize silhouettes of the validation models. That is, if we parameterize with parameter t the pixels coordinates of the silhouette contour $(C_x(t), C_y(t))$, then the noise corrupted points are given by:

$$C'_x(t) = C_x(t) + \epsilon_x(t)$$
$$C'_y(t) = C_y(t) + \epsilon_y(t)$$

with $$\epsilon_x(t) = \phi\epsilon_x(t-1) + \psi n_{0,\sigma}$$
$$\epsilon_y(t) = \phi\epsilon_y(t-1) + \psi n_{0,\sigma}.$$

In this experiment $\phi = 0.9512$ (which means the time constant of the noise process τ is 20 pixels) and $n_{0,\sigma}$ represents zero-mean Gaussian noises. The standard deviation of the gaussian noise was varied so that the standard deviation of ϵ_x and ϵ_y went from 0.5 pixels to 16 pixels. Figure 4 illustrates the effect of such noise process on a human silhouette and Figure 5 demonstrates the effect of an increasing noise on the silhouettes on the reconstruction accuracy for a large number of test subjects.

Even if a first order auto-regressive model is a simplistic noise model for segmented silhouettes, the results presented in Figure 5 demonstrate that three-dimensional reconstruction of human body model from two silhouettes can be robust even with large extraction errors in the silhouettes. However, in practical scenarios, the departure of the extracted from the true silhouette may not be governed by a zero mean process and outliers may also be present.

4.3 Synthetic Measurements

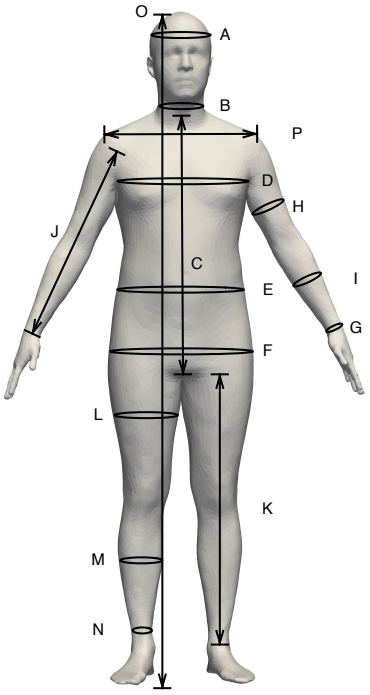
We designed a set of measures that was manageable (number of measurements), diverse (covering different portions of the body), and commonly used in garment fitting (since this is the industry we are primarily concerned about) to further validate the reconstructions obtained from silhouettes. Figure 6 introduces the 16 measurements that were selected for our experiments.

Measurements represented with straight lines are Euclidean distances between vertices of the reconstructed models and measurements represented by an ellipse are circumferences that are measured on the body surface using geodesic distances [18] between a few points on the desired contour. Since the reconstructed model have consistent parameterization (*i.e.* vertices with the same indices in different models are positioned on the same anatomical structures), indices of the relevant vertices were identified on a template model and are used to automatically measure the reconstructed models.

The accuracy of the measurements was tested by reconstructing the 220 subjects selected for validation using their frontal and lateral silhouettes, measuring the resulting models, and comparing with measurements performed on the original models from the CAESAR dataset. Also, we compared our reconstruction approach to two other possible methods while using the same silhouette parameterization (sGPLVM mapping [6] which is the first step of our algorithm and linear mapping [33]). The results are compiled in Table 1. It should be noted that the proposed method performs better on all measurements, although the differences between two best performing methods can be large (as in the case of the vertical distance between the should-blade and the crotch) or small (as in the case of the pelvis circumference).

4.4 Live Subjects Measurements

In addition to synthetic experiments, we also performed an experiment where four subjects' bodies were reconstructed from silhouettes extracted from photographs



	Measurement
A	Head circumference
B	Neck circumference
C	Shoulder-blade/crotch length
D	Chest circumference
E	Waist circumference
F	Pelvis circumference
G	Wrist circumference
H	Bicep circumference
I	Forearm circumference
J	Arm length
K	Inside leg length
L	Thigh circumference
M	Calf circumference
N	Ankle circumference
O	Overall height
P	Shoulder breadth

Fig. 6: Three-dimensional measurements used in the validation process.

Measurement	Proposed Mean error±Std. Dev.	sGPLVM Mapping Mean error±Std. Dev.	Linear Mapping Mean error±Std. Dev.
A	**10±12**	23±27	50±60
B	**11±13**	27±34	59±72
C	**4±5**	52±65	119±150
D	**10±12**	18±22	36±45
E	**22±23**	37±39	55±62
F	**11±12**	15±19	23±28
G	**9±12**	24±30	56±70
H	**17±22**	59±76	146±177
I	**16±20**	76±100	182±230
J	**15±21**	53±73	109±141
K	**6±7**	9±12	19±24
L	**9±12**	19±25	35±44
M	**6±7**	16±21	33±42
N	**14±16**	28±35	61±78
O	**9±12**	21±27	49±62
P	**6±7**	12±15	24±31

Table 1: Comparison between measurements made on the ground truth models to the same measurements made automatically on the reconstructed 3D models. Errors are expressed in millimeters. See Figure 6 for measurements illustration.

acquired using two Canon EOS 5D cameras. Then, we measured those individuals with the tools that would normally be used to create a custom-fitted garment (*i.e.* measuring tape and ruler).

The differences between the measurements performed manually on the subjects and the automatic measurements obtained from the reconstructed 3D are presented in Table 2. The differences are compatible with the results obtained from the synthetic experiments (see Table 1), but usually slightly higher.

Three main reasons explains this slight increase. First, the experimenter had little experience in acquiring manual body shape measurements. Gordon et al. [14] reported repeatability close to 1cm for similar measurements after extensive training and regular controls. Second, some of the proposed measurements involve

bony landmarks that are difficult to locate without palpation. Thus, the two procedures (the manual and the automated one) were perhaps not measuring from the exact same locations. Third, the consistency in the parameterization may not be perfect. The same vertices in different reconstructed models may not correspond exactly to the same anatomical location, which means that the manual and automatic procedure once again may refer to slightly different measures. The 3D models are reconstructed using silhouettes without manually pre-established 3D correspondences across the body. Therefore, there are areas of the body where little information is available from the silhouettes. The parameterization in those areas is thus primarily determined by statistics and not by image features.

5 Discussion and Conclusion

In this paper, we choose to limit ourselves to one standardize posture. This transpires both in the reconstruction method itself and in the validation experiments that we performed. This choice reduced the number of parameters that had to be accounted for, which made the experiments tractable. However, this choice might also be responsible for a portion of the reconstruction errors. Even though the subjects are directed to assume a standardize posture, there was always some variations in the subjects' stance. That means some posture variability is integrated in the shape variability model which decreases its predictive power. As a future work, it would therefore be interesting to test whether the addition of a skeleton (to represent the postures) would result in more accurate reconstructions or if the additional parameters would just create more problems with the optimization procedures.

The differences between the synthetic experiments and the real experiments highlighted to fact that certain measurements are more difficult to perform reliably both on computerized models and on actual subjects. Studying and/or designing a new set of body measurements that can be performed easily without resorting to palpation would be useful for further validation studies. Moreover, it would also be interesting to investigate the quality of the parameterization of the resulting models, since automated measurements relies on correct parameterization to perform properly.

In summary, we demonstrated in this paper that shape-from-silhouettes can be applied to the reconstruction of the human body from a lateral and a frontal silhouette. Moreover, we showed that with the integration of a statistical prior, the resulting 3D models are realistic and accurate. The proposed method is comprised of three steps. First, a non-linear mapping that goes from silhouette appearance space to the space of body shape models is used to generate an initial 3D model. That mapping uses a shared gaussian process latent variable model (sGPLVM) to link the principal components of the silhouettes to the principal components of the body shape model. Then, *maximum a posteriori* estimation of the body shape is performed using the first step's results as initial approximation. Finally, the body shape model is refined to best fit the input silhouettes.

We also demonstrated through real and synthetic experiments that the 3D models obtained with the proposed method are robust to perturbations applied to the input silhouettes. These experiments led us to believe that the proposed method may be adequate for applications such as garment fitting. More importantly, the method is also fast and completely repeatable whereas measurements performed manually are slower and subject to large inter-observer variations. To our knowledge, it was the first time the accuracy of reconstructions performed using a frontal and a lateral silhouettes were extensively analyzed.

References

1. A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, 2006.
2. B. Allen, B. Curless, and Z. Popovic. Exploring the space of human body shapes: data-driven synthesis under anthropometric control. In *Digital Human Modeling for Design and Engineering Conference*. SAE International, 2004.
3. D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):416, 2005.
4. S. Arya and D. M. Mount. Approximate nearest neighbor queries in fixed dimensions. In *Symposium on Discrete algorithms*, pages 271–280, 1993.
5. A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
6. Y. Chen and R. Cipolla. Learning shape priors for single view reconstruction. In *IEEE International Workshop on 3-D Digital Imaging and Modeling (3DIM'09)*, pages 1425–1432, 2009.
7. K.-M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 77–84, 2003.
8. K.-M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time part i: Theory and algorithms. *International Journal of Computer Vision*, 62(3):221–247, 2005.
9. K.-M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time part ii: Applications to human modeling and markerless motion tracking. *International Journal of Computer Vision*, 63(3):225–245, 2005.
10. Q. Delamarre and O. Faugeras. 3D articulated models and multi-view tracking with silhouettes. In *International Conference on Computer Vision*, volume 2, pages 716–721, 1999.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Difference	11	27	20	21	14	42	21	23	13	20	34	33	12	14	9	30

Table 2: Differences between measurements made automatically on body shape reconstructed from silhouettes and values measured on the subjects directly (values are expressed in millimeters).

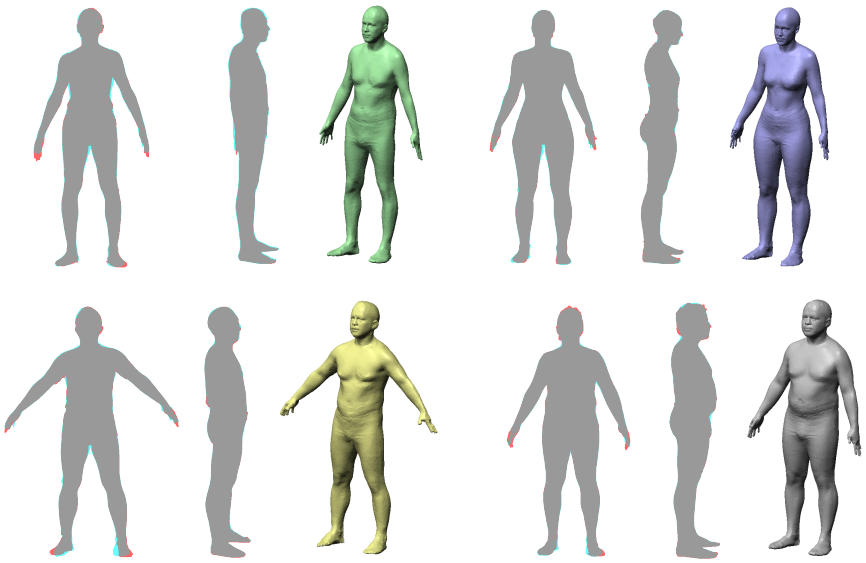


Fig. 7: Three-dimensional reconstruction of four subjects from frontal and lateral silhouettes. Silhouettes are color-coded to indicate whether the input silhouette and the silhouette casted by the reconstructed model coincide (dark grey), the input silhouette is larger (red), or the silhouette of the reconstructed model is larger (cyan).

11. C. Ek, P. Torr, and N. Lawrence. Gaussian process latent variable models for human pose estimation. In *Machine Learning for Multimodal Interaction*, pages 132–143, 2007.
12. A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. In *European Conference on Computer Vision*, pages 751–767, 2000.
13. L. Gond, P. Sayd, T. Chateau, and M. Dhome. A regression-based approach to recover human pose from voxel data. In *International Conference on Computer Vision Workshops*, pages 1012–1019, 2009.
14. C. C. Gordon, T. Churchill, C. E. Clauser, B. Bradtmiller, J. T. McConville, I. Tebbetts, and R. A. Walker. 1988 anthropometric survey of u.s. army personnel: Methods and summary statistics. Technical Report NATICK/TR-89/044, U.S. Army Natick Research, Development, and Engineering Center, 1989.
15. 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1823–1830, 2010.
16. T. Horprasert, D. Harwood, and L. Davis. A statistical approach for real time robust background subtraction and shadow detection. In *International Conference on Computer Vision - Frame Rate Workshop*, 1999.
17. I. Kakadiaris and D. Metaxas. Three-dimensional human body model acquisition from multiple views. *International Journal of Computer Vision*, 30(3):191–218, 1998.
18. R. Kimmel and J. Sethian. Computing geodesic paths on manifolds. *Proceedings of the National Academy of Sciences of the United States of America*, 95(15):8431, 1998.
19. K. Kutulakos and S. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.
20. A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:150–162, 1994.
21. L. Li, W. Huang, I. Gu, and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11):1459, 2004.
22. D. C. Liu and J. Nocedal. On the limited memory method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.
23. I. Mikić, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision*, 53(3):199–223, 2003.
24. K. Robinette and H. Daanen. The caesar project: a 3-d surface anthropometry survey. In *3D Digital Imaging and Modeling*, pages 380–386, 1999.
25. C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH*, page 314. ACM, 2004.
26. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002. cited By (since 1996) 857.
27. H. Seo, Y. Yeo, and K. Wohn. 3D Body reconstruction from photos based on range scan. *Technologies for E-Learning and Digital Entertainment*, pages 849–860, 2006.
28. A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao. Learning shared latent structure for image synthesis and

robotic imitation. In *Neural Information Processing Systems*, 2005.
29. D. Shreiner, M. Woo, J. Neider, and T. Davis. *OpenGL(R) Programming Guide : The Official Guide to Learning OpenGL(R), Version 2 (5th Edition)*. Addison-Wesley Professional, August 2005.
30. L. Sigal, A. Balan, and M. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. *Advances in neural information processing systems*, 2007.
31. Y.-P. Tsai, C.-H. Ko, Y.-P. Hung, and Z.-C. Shih. Background removal of multiview images by learning shape priors. *IEEE Transactions on Image Processing*, 16(10):2607–2616, 2007.
32. P. Xi, W.-S. Lee, and C. Shu. Analysis of segmented human body scans. In *Graphical Interface*, pages 19–26, 2007.
33. P. Xi, W.-S. Lee, and C. Shu. A data-driven approach to human-body cloning using a segmented body database. *Pacific Conference on Computer Graphics and Applications.*, pages 139–47, 2007.