Estimating Human Shape and Pose from a Single Image

Peng Guan   Alexander Weiss   Alexandru O. Bălan   Michael J. Black
Department of Computer Science, Brown University, Providence, RI 02912, USA
{pguan, aweiss, alb, black}@cs.brown.edu

Abstract

We describe a solution to the challenging problem of estimating human body shape *from a single photograph or painting. Our approach computes shape and pose parameters of a 3D human body model directly from monocular image cues and advances the state of the art in several directions. First, given a user-supplied estimate of the subject's height and a few clicked points on the body we estimate an initial 3D articulated body pose and shape. Second, using this initial guess we generate a tri-map of regions inside, outside and on the boundary of the human, which is used to segment the image using graph cuts. Third, we learn a low-dimensional linear model of human shape in which variations due to height are concentrated along a single dimension, enabling height-constrained estimation of body shape. Fourth, we formulate the problem of* parametric human shape from shading. *We estimate the body pose, shape and reflectance as well as the scene lighting that produces a synthesized body that robustly matches the image evidence. Quantitative experiments demonstrate how smooth shading provides powerful constraints on human shape. We further demonstrate a novel application in which we extract 3D human models from archival photographs and paintings.*

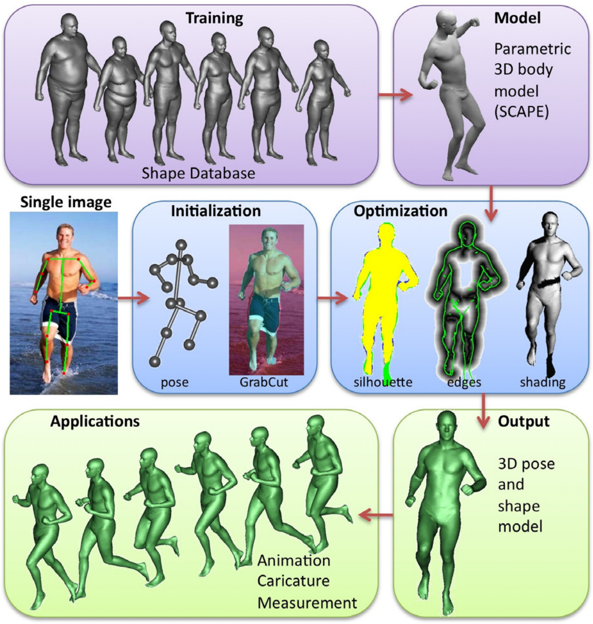


Figure 1. **Overview.** Given a single image and minimal user input, we compute an initial pose, light direction, shape and segmentation. Our method optimizes 3D body shape using a variety of image cues including silhouette overlap, edge distance, and smooth shading. The recovered body model can be used in many ways; animation using motion capture data is illustrated.

1. Introduction

While the estimation of 3D human pose in uncalibrated monocular imagery has received a great deal of attention, there has been almost no research on estimating human body shape. The articulated and non-rigid nature of the human form makes shape estimation challenging yet its recovery has many applications ranging from graphics to surveillance. Here we describe the first complete solution to the problem of human shape estimation from monocular imagery. In contrast to the standard multi-camera setting, we observe that a single image silhouette is generally insufficient to constrain 3D body shape. To address this we propose the use of additional monocular cues including smooth shading. Given an initial guess of the body pose, we optimize the pose, shape and reflectance properties of a 3D body model such that it robustly matches image measurements. The resulting body model can be measured, posed, animated, and texture-mapped for a variety of applications. The method is summarized in Figure 1.

Most work on human pose or shape estimation assumes the existence of a known background to enable the extraction of an accurate foreground silhouette. With a monocular image, however, no known background can be assumed. Still, the outline of the body provides a strong constraint on body shape. Given an initial pose, obtained by manual clicking on a few image locations corresponding to the major joints of the body [15, 24], and the mean body shape, we create an initial foreground region from which we derive a tri-map for GrabCut segmentation [18]. This produces an accurate foreground region (cf. [10]).

Our parametric representation of the body is based on the SCAPE model [2] and our optimization follows that of Bălan *et al.* [4, 5] but extends it to monocular imagery. Body pose in a monocular image is inherently ambiguous and its estimation from a single silhouette is poorly constrained. If the body limb lengths are not known, then multiple poses can equally well explain the same foreground silhouette [23]. To deal with this we constrain the height of the subject during optimization. Previous SCAPE formulations, however, represent variation in human shapes using linear shape deformation bases computed with principal component analysis. Since height is correlated with other shape variations, height variation is spread across many bases. We address this by rotating the learned SCAPE basis so that height variation is concentrated along a single shape basis direction. The height can then be held fixed during optimization, significantly improving monocular shape and pose estimation.

One of the key contributions of this work is the formulation of *body shape from shading.* Unlike the generic shape from shading problem, our goal is to estimate the body shape parameters, the pose of the body, its reflectance properties and the lighting that best explains the shading and shadows observed in the image (similar in spirit to [6] but with a more complex model). We assume a single point light source but our experiments suggest that the method is quite robust to violations of this assumption. Since skin has a significant specular component, we approximate its reflectance with a Blinn-Phong model [7] and an assumption of piecewise smoothness. Given a body shape, body pose, light direction and skin reflectance we robustly match a synthesized image of the person with the observed image. Note that exploiting shading cues requires accurate surface normals, which are provided by our learned body shape model. Shading information provides strong constraints on surface shape that improve the estimated body shape when combined with other cues.

We quantitatively evaluate the method in a laboratory environment with ground truth 3D shape. We also show a novel application where we compute 3D human shape from photographs and paintings. Here our assumptions about the illumination are only approximate, yet the method is able to recover plausible models of the human body.

2. Related Work

3D body pose from monocular images. There are many automated methods for extracting 2D human pose from a monocular image that could be used to initialize our method; a full review is beyond our scope. Instead we use a standard method for manual initialization. Taylor [24] recovers a set of body poses consistent with clicked 2D points corresponding to major joints. The method uses an articulated skeleton with known limb lengths, assumes an orthographic camera and finds only relative depths. A perspective solution [15] predates Taylor and forms the basis of our initialization method.

Body shape from images. Body *shape* is a pose-independent representation that characterizes the fixed skeletal structure (length of the bones) and the distribution of soft tissue (muscle and fat). There are several methods for representing body shape with varying levels of specificity: 1) non-parametric models such as visual hulls, point clouds and voxel representations (not considered further here); 2) part-based models using generic shape primitives such as cylinders or cones [9], superquadrics [14, 22] or "metaballs" [17]; 3) humanoid models controlled by a set of pre-specified parameters such as limb lengths that are used to vary shape [12, 13, 16]; 4) data driven models where human body shape variation is learned from a training set of 3D body shapes [2, 5, 20, 21].

Machine vision algorithms for estimating body shape typically rely on structured light, photometric stereo, or multiple calibrated camera views in carefully controlled settings where the use of low specificity models such as visual hulls is possible. As the image evidence decreases, more human-specific models are needed to recover shape. Several methods fit a humanoid model to multiple video frames, or multiple snapshots from a single camera [12, 22]. These methods estimate limited aspects of body shape such as scaling parameters or joint locations yet fail to capture the range of natural body shapes.

More realism is possible with data-driven methods that encode the statistics of human body shape. Seo *et al.* [20] use a learned deformable body model for estimating body shape from multiple photos in a controlled environment with the subject seen in a predefined pose. To estimate a consistent body shape in arbitrary pose it is desirable to have a body model that factors changes in shape due to pose and identity. Consequently, we use the SCAPE model [2], which is derived from laser scans and captures realistic articulated and non-rigid pose-dependent deformations, as well as shape variations between individuals. Bălan *et al.* [5] show that such a model allows for the shape and pose to be estimated directly from multi-camera silhouettes.

A single monocular image presents challenges beyond the capabilities of all the methods above. The only work we know to directly estimate body shape from a single image is that of Sigal *et al.* [21]. They train a mixture of experts model to predict 3D body pose and shape directly from various 2D shape features computed from an image silhouette. They estimate body shape in photos taken from the Internet, but require manual foreground segmentation and do not accurately estimate pose. While silhouettes constrain the surface normals at the object boundary, non-rigid deformation, articulation and self occlusion make the silhouette boundary insufficient to recover accurate shape from a single view.

Body shape from shading. Shape from shading has a long history in computer vision yet typically focuses on recovering the shape of unknown surfaces. Here we have a different problem in which we know that the object is a human but the pose and shape are unknown. For a given set of body shape and pose parameters we can compute the surface normals at each point on the body mesh. We then formulate and optimize a robust *shape from shading* objective function in which the normals are a function of the shape parameters. Similar to this is the work of Samaras and Metaxas [19], which constrains a 3D shape using shading information. We go beyond their work to deal with a learned shape deformation model and articulation.

The majority of work related to shading and the human body focuses on carefully calibrated laboratory environments. Theobalt *et al.* [25] recover human body shape and detailed reflectance properties but do so in a multi-camera calibrated environment with careful lighting. Bălan *et al.* [4] recover the albedo of the body using multiple known poses and a Lambertian reflectance model but do not use this to estimate shape. These methods are not applicable to the monocular, uncalibrated case studied here.

In recent work, de La Gorce *et al.* [8] use an accurate hand shape model and shading information for monocular tracking. Our work goes beyond this to estimate a parametric shape model for the whole body in arbitrary poses with piecewise smooth albedo and unknown background.

3. Body Model and Fitting

SCAPE is a deformable, triangulated mesh model of the human body that accounts for different body shapes, different poses, and non-rigid deformations due to articulation [2]. For vision applications, it offers realism while remaining relatively low dimensional. We use a mesh with $m = 12,500$ vertices [5].

Articulated pose is parametrized by a set of rigid body part rotations $\vec{\theta}$, while changes in body shape between individuals are captured by shape deformations gradients $\vec{d}$ between a reference mesh and a new mesh in the same pose. A low-dimensional statistical model of body shape deformations is learned using principal component analysis (PCA). We learn two gender-specific models from laser scans of over 1000 men and 1000 women, respectively. For a given mesh, the shape deformation gradients are concatenated into a single column vector and approximated as $\vec{d} = \mathbf{U}\vec{\beta} + \vec{\mu}$ where $\vec{\mu}$ is the mean body shape, $\mathbf{U}$ are the first n eigenvectors given by PCA and $\vec{\beta}$ is a vector of linear coefficients that characterizes a given shape; $n = 20$ in our experiments. In Section 4 we extend this formulation to model deformations that preserve height.

Given a monocular image, our goal is to estimate the shape parameters $\vec{\beta}$ and pose parameters $\vec{\theta}$ that best explain the image evidence. The model parameters are used to produce a 3D mesh, $Y(\vec{\beta}, \vec{\theta})$, that is projected onto the image plane to obtain silhouettes, edges, or shaded appearance images (Fig. 1). We denote the body parameters by $\Theta_B = [\vec{\beta}, \vec{\theta}]$. We use standard distance functions for silhouettes, $E_{\mathrm{Si}}(\Theta_B)$, [5, 22] and edges, $E_{\mathrm{Eg}}(\Theta_B)$, [9] and introduce a novel shading term in Section 5. The objective function, which also includes an inter-penetration penalty, $E_{\mathrm{Pn}}(\Theta_B)$, is minimized using a gradient-free direct search simplex method.

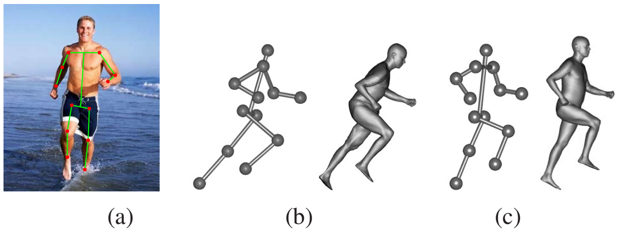

(a) (b) (c)

Figure 2. **Initialization using clicked points on the input image.** Pose estimated with orthographic (b) and perspective (c) camera models, shown from an alternate view. Mean body shape (male) is shown transformed into the pose of the initialized models.

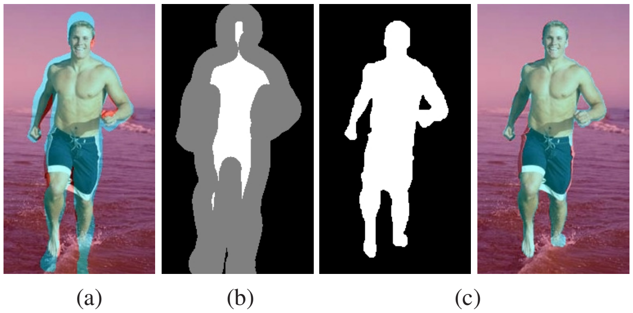
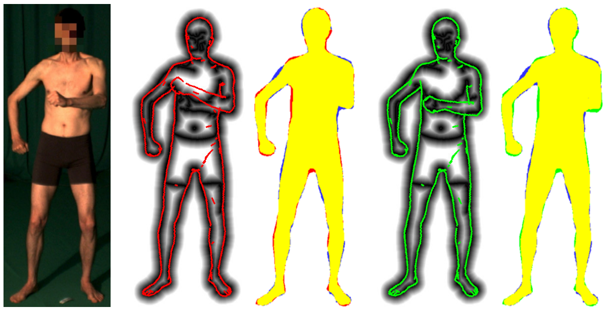

(a) (b) (c)

Figure 3. **Segmentation.** (a) Silhouette corresponding to initial pose and average shape projected into image. (b) Tri-map extracted from initial silhouette by erosion and dilation. (c) GrabCut segmentation result (silhouette and its overlay in the image).

3.1. Pose Initialization

3D body pose is initialized in the camera coordinate system using clicked 2D points corresponding to the major joints (Fig. 2) [15, 24]. We find that the orthographic method of Taylor [24] (Fig. 2b) produces poses that are inaccurate compared with the perspective method of [15] (Fig. 2c). The perspective method requires an estimate of focal length which we extract from EXIF metadata when available or which we obtain from user input. We further find that even an approximate focal length produces better initial poses than the orthographic assumption.

Unlike the orthographic case, perspective projection requires a way to position the root joint in 3D. First, the limb most parallel to the image plane is automatically identified as the one that minimizes the ratio between its image length



(a) (b) Result without edges (c) Result with edges

Figure 4. **Internal edges.** (a) Laboratory image with self occlusion. (b) Pose estimation with only the silhouette term cannot estimate the arm pose. Edges (red) of optimized model projected into the edge cost image. Yellow shows the overlap of the model and image silhouettes, blue/red represent unmatched image/model silhouette regions. (c) The estimated pose (green) with the edge term matches the true pose. Note how well the model edges align with the edge cost image.

and its actual length. If the limb is parallel to the image plane, the depth is uniquely determined using the ratio of similar triangles. If not we use a foreshortening factor similar to the scale parameter in [24].

In contrast to [24], we do not explicitly require limb lengths as input. Rather, we predict these from a database of over 2400 subjects based on user specified height and gender. We use this database to build a height constrained shape space as described in Section 4, allowing us to deform the mesh to match the mean person of the specified gender and height, and then extract limb lengths from linear combinations of specific vertices.

We extend the previous methods to also initialize head pose by solving for the neck rotation that minimizes the distance between several user-clicked 2D face feature points and the corresponding 3D vertex positions on the mesh projected into the image.

3.2. Region-based Segmentation

Given the initial mesh (Fig. 2c), we render its silhouette into the image. This provides an initial guess for foreground segmentation (Fig. 3a). Specifically, we construct a tri-map, defining each pixel as foreground, background, or uncertain by eroding and dilating the initial region by 5% of the image width (Fig. 3b). The resulting tri-map is used to initialize GrabCut [18] which is used to segment the foreground. This process is similar to that of Ferrari *et al.* [10] but with a 3D body model used for initialization.

3.3. Internal Edges

It is well known that silhouettes do not provide pose constraints in regions where one body part occludes another (e.g. Fig. 4b). Numerous authors have dealt with this by combining edge information with silhouettes. We do so as well but, with the SCAPE body model, these edges provide a better fit to the image evidence than do previous models.

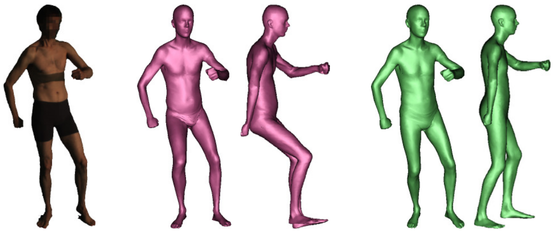


Figure 5. **Height Constrained Optimization.** Two different body shapes and poses can explain the same image silhouette. Pose and shape estimated without constraining height (magenta). When turned sideways we see it is wrong. Constraining the height during estimation produces a realistic pose (green).

We detect image edges using a standard edge detector and apply a thresholded distance transform to define an edge cost map normalized to $[0, 1]$. Model edges correspond to visible vertex edges for which there is sign change in the dot product between the corresponding triangle normals and the ray from the camera to the midpoint of the edge. We use the trapezoid rule to evaluate the line integral of the set of all visible model edges over the edge cost image. This defines an edge cost, $E_{\text{Eg}}(\Theta_B)$, that is included in the objective function, improving the accuracy of the fit (Fig. 4c).

4. Attribute Preserving Shape Spaces

The ambiguities present in inferring 3D pose and shape from a single image mean that we must constrain the search space as much as possible. Figure 5 illustrates one such ambiguity where the wrong body shape can be compensated for by a change in pose. Viewed monocularly, both models explain the image silhouette equally well. Additional information such as the height of the person can remove some of the ambiguity. Unfortunately, the SCAPE eigen-shape representation does not provide any direct control parameters corresponding to intuitive attributes like gender, height or weight that can be specified by a user. If these can be derived as functions of the linear coefficients, then they can be included as constraints during body shape estimation. Instead we take a more direct approach and construct a rotation of the original eigen-shape space such that height variation is removed from all but one of the bases. This allows us to optimize over body shapes without varying height.

In previous work, Blanz and Vetter [6] compute a direction in shape coefficient space such that any movement along this axis manipulates a certain attribute the most while keeping all the other attributes as constant as possible. This is not equivalent to saying that any movement orthogonal to this axis preserves the attribute, which is what we want. In fact, their axis is not optimized for and fails to preserve an

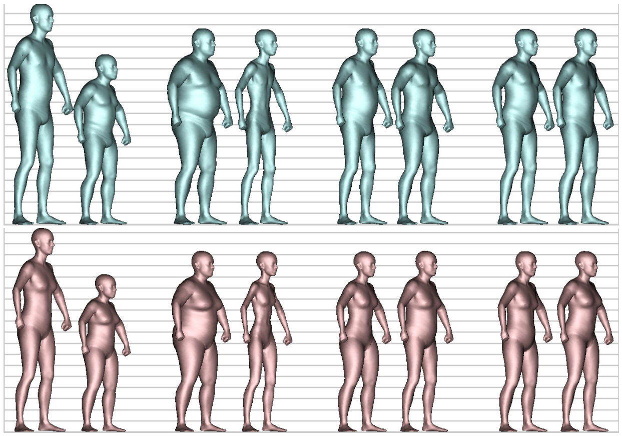
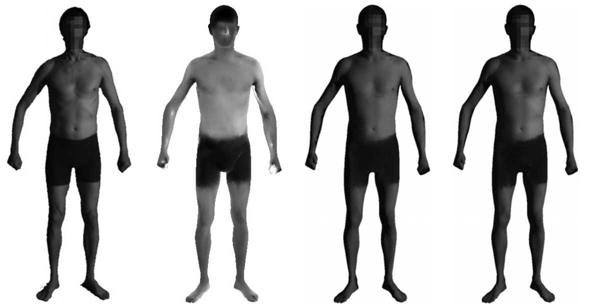


Figure 6. **Height preserving body shape space.** First pair on each row (men above, women below) shows variation (± 3 std) along the height-variation axis. The other pairs show variation (± 3 std) along the first three height-invariant axes. Note that shape varies along these axes but height varies by less than $3mm$ for each pair.



(a) Input (b) Albedo (c) Lambertian (d) Blinn-Phong

Figure 7. **Estimated reflectance.** Blinn-Phong model captures specular highlights and is more accurate than the Lambertian model. Note robust spatial term captures discontinuous albedo.

attribute value along orthogonal directions.

Allen *et al.* [1] learn a linear mapping from a fixed set of attributes to shape parameters. One could optimize body shape using these parameters instead of PCA coefficients. Preserving an attribute can then be achieved by simply keeping it fixed, but this approach reduces the modes of shape variation to the set of attributes considered.

In contrast, our approach explicitly searches for attribute-preserving directions in the eigen-space and reorients the bases along these directions. While we focus on constraining height, our method applies to any other geometric attribute that can be measured directly from the mesh (volume, geodesic distances, etc.). Body height $H(\vec{\beta})$ can be measured by reconstructing a mesh $Y(\vec{\beta}, \vec{\theta}^H)$ in a predefined standing pose $\vec{\theta}^H$. Let $\mathbf{G}_1 = \mathbf{I}_n = [\vec{e}_1, \ldots, \vec{e}_n]$ be the identity basis for the shape coefficients ($\vec{d} = \mathbf{U}\mathbf{G}_1\vec{\beta} + \mu$). We seek a new orthonormal basis $\mathbf{G}$ such that none of its bases account for height except one, which becomes the height axis. $\mathbf{G}$ should also preserve the representational power of the original bases: the sub-space spanned by the first j bases is the same after the change of bases, absent the height axis. Our solution works in an incremental fashion and maintains orthogonality at all times by rotating pairs of bases so that one of the bases preserves height while the other moves towards the height axis. First, we start by selecting a candidate basis $\vec{e}_k$ for the height axis as the one that maximizes the absolute correlations between height and shape coefficients of the training examples. Second, we iterate over the remaining bases $\vec{e}_j$ and rotate the current $(\vec{e}_j, \vec{e}_k)$ plane to make $\vec{e}_j$ height preserving. Third, the rotation matrix is used to update, at iteration j, the orthonormal basis $\mathbf{G}_j =$

$$\mathbf{G}_{j-1}\mathbf{R}_{jk}\left(\arg\min_{\varphi}\left(H(\vec{0}_n) - H(\mathbf{G}_{j-1}\mathbf{R}_{jk}(\varphi)\vec{e}_j)\right)^2\right),$$

where $\mathbf{R}_{jk}(\varphi)$ is a $n \times n$ rotation in the $(\vec{e}_j, \vec{e}_k)$ plane:

$$\mathbf{R}_{jk}(\varphi) = \begin{array}{cc} & \begin{array}{ccccc} \phantom{\mathbf{I}} & j & \phantom{\mathbf{I}} & k & \phantom{\mathbf{I}} \end{array} \\ \begin{array}{c} \\ j \\ \\ k \\ \\ \end{array} & \left(\begin{array}{ccccc} \mathbf{I} & & & & 0 \\ & \cos(\varphi) & & -\sin(\varphi) & \\ & & \mathbf{I} & & \\ & \sin(\varphi) & & \cos(\varphi) & \\ 0 & & & & \mathbf{I} \end{array}\right) \end{array}.$$

The body shape in the new height-preserving shape space can be expressed as $\vec{d} = (\mathbf{U}\mathbf{G}_n)\vec{\beta}' + \vec{\mu}$, where $\vec{\beta}' = (\mathbf{G}_n)^{-1}\vec{\beta}$. By convention, we compute the variance along the new bases and order them in decreasing order following the height axis. Figure 6 shows deviations from the mean shape in the male and female height-preserving shape spaces.

For many subjects (e.g. celebrities), height may be known. When unknown (e.g. in paintings) we use the mean height for each gender (women=$1.65m$, men=$1.82m$).

5. Body Shape from Shading

We approximate the body's reflectance using a Blinn-Phong model with diffuse and specular components [7]. We assume a single light source and ambient illumination. Let $X(\Theta_B) = [\vec{x}_1, \vec{x}_2, ..., \vec{x}_m]$ be the positions of the m vertices of a body mesh, and $N(\Theta_B) = [\vec{n}_1, \vec{n}_2, ..., \vec{n}_m]$ be the associated unit length normals. Notice that both X and N are functions of the pose and shape parameters, allowing us to formulate a parametric *shape from shading* problem. Let $\vec{a} = [a_1, a_2, ..., a_m]$ be the albedo of each vertex and $\vec{s} = [s_1, s_2, ..., s_m]$ be the specularity of each vertex. The shading value of each surface point i is approximated by:

$$\hat{r}_i = b + a_i(\vec{\ell}_i \cdot \vec{n}_i)l + s_i(\vec{h}_i \cdot \vec{n}_i)^{\alpha}l \qquad (1)$$

where $\vec{\ell}_i$ is the direction from vertex $\vec{x}_i$ toward the light source, $\vec{h}_i$ is the halfway vector between $\vec{\ell}_i$ and the direction from vertex i toward the camera, b is ambient illumination, l is light intensity, and α is the specular exponent.

For a distant directional light source (outdoor scene) l is constant for every vertex, while for a point light source (indoor scene) we use a quadratic attenuation function for light intensity with distance from the source (as in [4]).

Optimization. The body is placed at the origin of a spherical coordinate system and the light position is parametrized as $\Theta_L = [\gamma, \phi, z]$ with respect to the body center, where γ and ϕ are azimuth and elevation respectively and z is the distance between the light source and the body. The parameters $\vec{l}_i$, $\vec{h}_i$ and l in Eq. 1 depend on Θ_L. We denote the reflectance parameters $\Theta_R = [\vec{a}, \vec{s}, b, \alpha]$. Suppose r_i is the linearly interpolated intensity in the input image where vertex i is projected, our goal is to minimize the energy function $E_{\mathrm{Sh}}(\Theta_B, \Theta_R, \Theta_L) \propto$

$$\begin{aligned}&\sum_{i\in visible}\Big\{\rho_{\eta_1}(\hat{r}_i(\Theta_B,\Theta_R,\Theta_L)-r_i) &(2)\\ &\quad+\lambda_1\sum_{j\in\mathcal{N}(i)}\frac{\rho_{\eta_2}(a_j-a_i)}{d_{j,i}}+\lambda_2\sum_{j\in\mathcal{N}(i)}\frac{\rho_{\eta_3}(s_j-s_i)}{d_{j,i}}\Big\}\end{aligned}$$

where $\mathcal{N}(i)$ are the vertices connected to vertex i, $d_{j,i}$ is $|\vec{x}_j - \vec{x}_i|$, and $\rho_\eta(x) = \frac{x^2}{\eta^2+x^2}$ is a robust error function [11] used to deal with outliers.

The first term in Eq. 2 penalizes the difference between measured intensities in the observed image, r_i, and the predicted brightness of corresponding model vertices, $\hat{r}_i(\cdot)$. The second term ensures that neighboring vertices have similar albedo. The robust formulation provides a piecewise smooth prior on albedo that allows spatial variations due to clothing, hair, variation in skin color, etc. The third term provides a piecewise smooth prior over specularity. λ_1 and λ_2 weight the relative faithfulness to the observed data and the spatial smoothness assumptions.

The user coarsely initializes Θ_L and then the energy function is minimized in an alternating fashion. First, Θ_L is optimized given fixed Θ_B and Θ_R. (Note that in the first iteration, Θ_B is the initial guess of pose and shape; the albedo and specularity in Θ_R are considered uniform.) Second, we optimize Θ_R with fixed Θ_L and Θ_B. Given the robust formulation in Eq. 2 no closed form solution is possible so we minimize using gradient descent. Third, we fix Θ_L, Θ_R and optimize Θ_B but here the optimization is more difficult since changing Θ_B affects the predicted brightness through changes in the vertex normals. Consequently a gradient-free simplex method is employed to solve step 3. We alternate between the three steps until a convergence criterion is met. We vary the λ values during optimization, starting with larger values and gradually decreasing them, so that the shape is forced to change in order to make the predicted brightness match the image observations. We find that initial pose needs to be fairly accurate, but illumination direction is relatively insensitive to the initialization. Figure 7 shows the estimated reflectance for one input image.

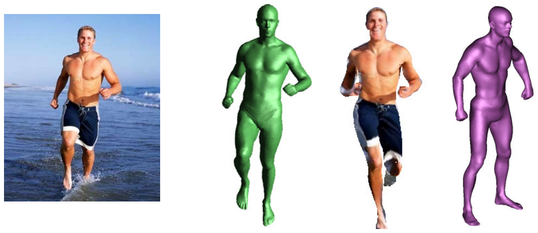


Figure 9. **Applications.** Shape and pose recovered from a single image; texture-mapped in new pose; caricature.

6. Results

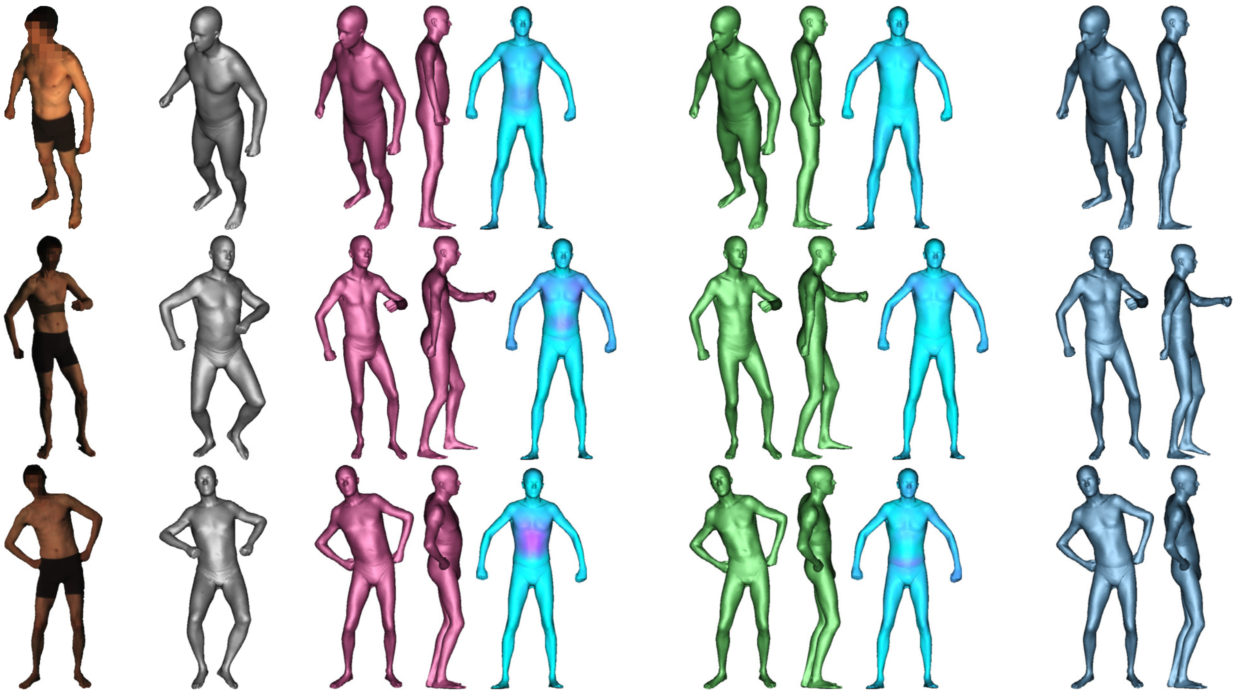
For quantitative analysis, we captured the pose and shape of a subject using eight synchronized and calibrated cameras with a single "point light source" and a green screen background. We fit the SCAPE model to the eight silhouettes and treat the resulting shape as ground truth.

We then quantify the extent to which shading cues improve monocular shape estimation by comparing the shape estimated with two formulations. In the "Silhouettes and Edges" (***SE***) formulation, we fit the pose and shape of the SCAPE model in the height preserving space by optimizing the cost function $E_1 = E_{\mathrm{Si}}(\Theta_B) + E_{\mathrm{Eg}}(\Theta_B) + E_{\mathrm{Pn}}(\Theta_B)$. The "Silhouettes, Edges, and Shading" (***SES***) formulation extends the first by incorporating shading cues; that is, $E_2 = E_1 + E_{\mathrm{Sh}}(\Theta_B, \Theta_R, \Theta_L)$.

Figure 8 illustrates how smooth shading improves shape recovery. Silhouettes, even with internal edges, are not sufficient to capture accurate body shape from monocular images. Incorrect estimates happen in areas where surface normals are oriented towards the camera, such as the abdomen in frontal images. In these regions shading provides a strong cue that constrains the body shape.

Anthropometric measurements of chest size, waist size, and weight are provided in Table 1. Waist and chest circumference are computed by transforming the body to a canonical pose, slicing the mesh on a fixed plane and computing the convex hull of the contour. Weight is estimated from the body volume by assuming it has the constant density of water. ***SES*** shows substantial improvement over ***SE***.

Figure 9 shows an image from the Internet with recovered pose and shape. Note that reflections off the water clearly violate our simple lighting model. Despite that the shape is well recovered. We animate the figure by preserving shape and generating meshes in novel poses from motion capture data. The model can be texture mapped with the image texture or some new texture. Large pose changes may require the texture synthesis of missing data. We can also vary the recovered shape to produce a caricature (Fig. 9 right). We do so by finding the shape coefficient with the



(a) Image (b) Init. (c) Silhouette and Edges (d) Silhouette, Edges and Shading (e) Ground Truth

Figure 8. **Comparison between *SE* (red) and *SES* (green).** Comparisons are performed on three different poses taken from different viewing angles. The initialization (b) is shown in the camera view. Results and ground truth are shown in both the camera view and in profile. For each result we also show an error map in a canonical pose, indicating per vertex displacement from ground truth; blue indicates low error, while purple indicates high error. Note the lower error for the ***SES*** model.

	Chest Size (cm)			Waist Size (cm)			Body Weight (kg)		
	SE	***SES***	***GT***	***SE***	***SES***	***GT***	***SE***	***SES***	***GT***
Pose 1	95.7 (+3.1)	92.7 (+0.1)	92.6	86.4 (+6.2)	79.6 (-0.6)	80.2	72.0 (+8.2)	65.4 (+1.6)	63.8
Pose 2	84.3 (-7.3)	87.1 (-4.5)	91.6	83.7 (+4.3)	78.5 (-0.9)	79.4	62.5 (-0.7)	62.4 (-0.8)	63.2
Pose 3	95.4 (+4.0)	91.9 (+0.5)	91.4	88.0 (+7.7)	76.9 (-3.4)	80.3	70.8 (+8.2)	63.5 (+0.9)	62.6

Table 1. **Anthropometric Measurements. *GT*** stands for ground truth size. The value in the parenthesis is the deviation from GT size. (Note that the ground truth sizes for each frame vary a little bit, since non-rigid deformations caused by articulations of body will result in variations of shape details.)

most significant deviation from the mean and exaggerate it, moving the shape further from the mean in that direction. Here it produces a more muscular physique.

Although paintings rarely conform to a physical lighting model, we find that shading cues are often significant. Using the same robust formulation as for photographs we recover body pose and shape from two paintings in Fig. 10.

7. Conclusions and Future Work

We have described a complete solution for reconstructing a model of the human body from a single image with only minimal user intervention. The main insight is that even a single image contains a range of cues that can constrain the interpretation of 3D body shape. While the bounding contour of the body alone is not sufficient, smooth shading can provide a powerful additional cue. Consequently we developed a new robust method for computing parametric body shape from shading. We also developed a new linear model of body shape deformation in which height variation is removed. The ability to extract body shape from a single image makes several new applications possible. For example, a character from a painting or photograph can be "brought to life" and animated in new poses.

The method as described has several limitations. We assume a single point light source and a simplified model of surface reflectance. None of our experiments actually conform to this model, and yet it still provides a useful approximation. Future work should consider expanding this to more general lighting conditions. We also plan to study more qualitative models of shading. Even in art which is not physically realistic, there are still strong local cues that we should be able to exploit to constrain body shape.

Our experiments have focused on naked or minimally

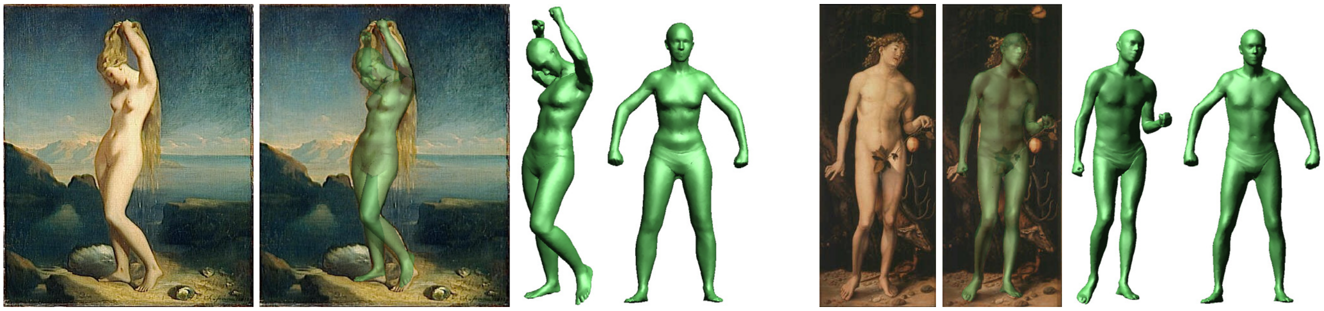


Figure 10. **Body shape and pose from paintings.** (left) Venus Anadyomene, Théodore Chasseriau, 1838. (right) Adam and Eve (detail), Hans Baldung Grien,1507. Images (left to right): painting, model overlay, recovered shape and pose, shape in new pose.

clothed people. Previous work has shown that body shape can be recovered even when people are wearing clothing if multiple poses and camera views are available [3]. Extending this to the monocular case is challenging as shading cues would need to be extended to model the complex shading variation caused by clothing.

Other future work will consider automating the initialization stage using a bottom-up 2D person detector and integrating body segmentation with the 3D model fitting process. Since our body shape representation is independent of pose we can also combine constraints from multiple snapshots of the same person. Each image may contain only weak cues but together they could constrain body shape.

Acknowledgements. We thank D. Hirshberg for his advice and model-edge projection code. Financial support: NSF IIS-0812364 and the RI Economic Development Corp.

References

[1] B. Allen, B. Curless, and Z. Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *SIGGRAPH*, pp. 587–594, 2003.

[2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape completion and animation of people. *SIGGRAPH*, 24(3):408–416, 2005.

[3] A. Bălan and M. Black. The naked truth: Estimating body shape under clothing. *ECCV*, LNCS 5303:15–29, 2008.

[4] A. Bălan, M. Black, H. Haussecker, and L. Sigal. Shining a light on human pose: On shadows, shading and the estimation of pose and shape. *ICCV*, pp. 1–8, 2007.

[5] A. Bălan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. *CVPR*, 2007.

[6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *SIGGRAPH*, pp. 187–194, 1999.

[7] J. Blinn. Models of light reflection for computer synthesized pictures. *SIGGRAPH*, pp. 192–198, 1977.

[8] M. de La Gorce, N Paragios and D. Fleet, Model-based hand tracking with texture, shading and self-occlusions. *CVPR*, pp. 1–8, 2008.

[9] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2005.

[10] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. *CVPR*, 2008.

[11] S. Geman and D. McClure, Statistical methods for tomographic image reconstruction. *Bulletin Int. Statistical Institute*. LII-4:5–21, 1987.

[12] D. Grest and R. Koch. Human model fitting from monocular posture images. *Proc. VMV*, pp. 665—1344, 2005.

[13] A. Hilton, D. Beresford, T. Gentils, R. Smith, W. Sun and J. Illingworth, Whole-body modelling of people from multiview images to populate virtual worlds, *The Visual Computer*, 16(7):411–436, 2000.

[14] I. Kakadiaris and D. Metaxas, Three-dimensional human body model acquisition from multiple views, *IJCV*, 30(3):191–218, 1998.

[15] H. Lee and Z. Chen. Determination of 3D human body postures from a single view. *CVGIP*, 30(2):148–168, 1985.

[16] W. Lee., J. Gu, N. Magnenat-Thalmann. Generating animatable 3D virtual humans from photographs. *Eurographics*, 19(3):1–10, 2000.

[17] R. Plänkers and P. Fua. Articulated soft objects for multiview shape and motion capture. *PAMI*, 25(10):63–83, 2003.

[18] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": Interactive foreground extraction using iterated graph cuts. *SIGGRAPH*, 23(3):309–314, 2004.

[19] D. Samaras D. Metaxas. Incorporating illumination constraints in deformable models for shape from shading and light direction estimation. *CVPR*, pp. 322–329, 1998.

[20] H. Seo, Y.I. Yeo and K. Wohn, 3D Body reconstruction from photos based on range scan, *Tech. for E-Learning and Digital Entertainment*, v. 3942, pp. 849–860, 2006.

[21] L. Sigal, A. Bălan, and M. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. *NIPS 20*, pp. 1337—1344, 2008.

[22] C. Sminchisescu and A. Telea. Human pose estimation from silhouettes, a consistent approach using distance level sets. *WSCG Int. Conf. C.G. Vis. Comp. Vision*, pp. 413–420, 2002.

[23] C. Sminchisescu and B. Triggs. Building roadmaps of local minima of visual models. *ECCV*, v. 1, pp. 566–582, 2002.

[24] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *CVIU*, 80(10):349–363, 2000.

[25] C. Theobalt, N. Ahmed, H. Lensch, M. Magnor, and H.-P. Seidel. Seeing people in different light – Joint shape, motion, and reflectance capture. *IEEE Trans. Visual. Comp. Graph.*, 13(4):663–674, 2007.