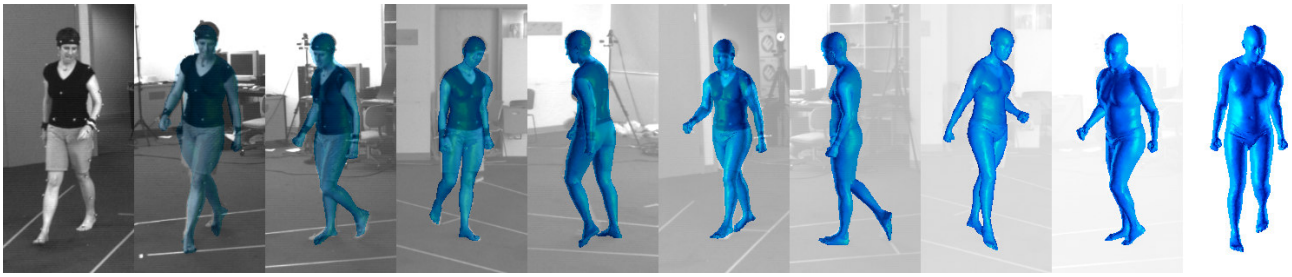
# Detailed Human Shape and Pose from Images



[1]Alexandru O. Bălan [1]Leonid Sigal [1]Michael J. Black [2]James E. Davis [3]Horst W. Haussecker
[1]Department of Computer Science, Brown University, Providence, RI 02912, USA
[2]Computer Science Department, UC Santa Cruz, Santa Cruz, CA 95064, USA
[3]Intel Research, Santa Clara, CA 95054, USA
{alb, ls, black}@cs.brown.edu davis@cs.ucsc.edu horst.haussecker@intel.com

## Abstract

*Much of the research on video-based human motion capture assumes the body shape is known* a priori *and is represented coarsely (e.g. using cylinders or superquadrics to model limbs). These body models stand in sharp contrast to the richly detailed 3D body models used by the graphics community. Here we propose a method for recovering such models directly from images. Specifically, we represent the body using a recently proposed triangulated mesh model called SCAPE which employs a low-dimensional, but detailed, parametric model of shape and pose-dependent deformations that is learned from a database of range scans of human bodies. Previous work showed that the parameters of the SCAPE model could be estimated from marker-based motion capture data. Here we go further to estimate the parameters directly from image data. We define a cost function between image observations and a hypothesized mesh and formulate the problem as optimization over the body shape and pose parameters using stochastic search. Our results show that such rich generative models enable the automatic recovery of detailed human shape and pose from images.*

## 1. Introduction

We address the problem of markerless human shape and pose capture from multi-camera video sequences using a richly detailed graphics model of 3D human shape (Figure 1). Much of the recent work on human pose estimation and tracking exploits Bayesian methods which require generative models of image structure. Most of these models, however, are quite crude and, for example, model the human body as an articulated tree of simple geometric primitives such as truncated cones [8]. Arguably these generative models are a poor representation of human shape.

As an alternative, we propose the use of a graphics model of human shape that is learned from a database of detailed 3D range scans of multiple people. Specifically we use the SCAPE (Shape Completion and Animation of PEople) model [1] which represents both articulated and non-rigid deformations of the human body. SCAPE can be thought of as having two components. The *pose deformation model* captures how the body shape of a person varies as a function of their pose. For example, this can model the bulging of a bicep or calf muscle as the elbow or knee joint varies. The second component is a *shape deformation model* which captures the variability in body shape across people using a low-dimensional linear representation. These two models are learned from examples and consequently capture a rich and natural range of body shapes, and provide a more detailed 3D triangulated mesh model of the human body than previous models used in video-based pose estimation.

The model has many advantages over previous deformable body models used in computer vision. In particular, since it is learned from a database of human shapes it captures the correlations between the sizes and shapes of different body parts. It also captures a wide range of human forms and shape deformations due to pose. Modeling how the shape varies with pose reduces problems of other approaches associated with modeling the body shape at the joints between parts.

Figure 1. **SCAPE from images.** Detailed 3D shape and pose of a human body is directly estimated from multi-camera image data. Several recovered poses from an image sequence of a walking subject are shown.

While recent work in the machine vision community has focused on recovering human kinematics from video, we argue that there are many motivations for recovering shape simultaneously. For example, anthropomorphic measurements can be taken directly from the recovered body model and may be useful for surveillance and medical applications. For some graphics applications, having direct access to the shape model for a particular subject removes an additional step of mapping kinematic motions to 3D models.

Our current implementation estimates the parameters of the body model using image silhouettes computed from multiple calibrated cameras (typically 3-4). The learned model provides strong constraints on the possible recovered shape of the body which means that pose/shape estimation is robust to errors in the recovered silhouettes. Our generative model predicts silhouettes in each camera view given the pose/shape parameters of the model. A fairly standard Chamfer distance measure is used to define an image likelihood and optimization of the pose/shape parameters is performed using a stochastic search technique related to annealed particle filtering [7, 8]. Our results show that the SCAPE model better explains the image evidence than does a more traditional coarse body model.

We provide an automated method for recovering pose throughout an image sequence by using body models with various levels of complexity and abstraction. Here we exploit previous work on 3D human tracking using simplified body models. In particular, we take the approach of Deutscher and Reid [8] which uses anneal particle filtering to track an articulated body model in which the limbs are approximated by simple cylinders or truncated cones. This automated tracking method provides an initialization for the full SCAPE model optimization. By providing a reasonable starting pose, it makes optimization of the fairly high-dimensional shape and pose space practical.

Results are presented for multiple subjects (none present in the SCAPE training data) in various poses.

## 2. Related Work

We exploit the SCAPE model of human shape and pose deformation [1] but go beyond previous work to estimate the parameters of the model directly from image data. Previous work [1] estimated the parameters of the model from a sparse set of 56 markers attached to the body. The 3D locations of the markers were determined using a commercial motion capture system and provided constraints on the body shape. Pose and shape parameters were estimated such that the reconstructed body was constrained to lie inside the measured marker locations. This prior work assumed that a 3D scan of the body was available. This scan was used to place the markers in correspondence with the surface model of the subject.

We go beyond the original SCAPE paper to estimate the pose and shape of a person directly from image measurements. This has several advantages. In particular, video-based shape and pose capture does not require markers to be placed on the body. Additionally, images provide a richer source of information than a sparse set of markers and hence provide stronger constraints on the recovered model. Furthermore, we show shape recovery from multi-camera images for subjects not present in the shape training set.

Previous methods have established the feasibility of estimating 3D human shape and pose directly from image data but have all suffered from limited realism in the 3D body models employed. A variety of simplified body models have been used for articulated human body pose estimation and tracking including cylinders or truncated cones (e.g. [8]) and various deformable models such as superquadrics [9, 14, 20] and free-form surface patches [17]. These models do not fit the body shape well, particularly at the joints and were typically built by hand [14] or estimated in a calibration phase prior to tracking [9, 17, 20]. Detailed but fixed, person-specific, body models have been acquired from range scans and used for tracking [6] by fitting them to voxel representations; this approach did not model the body at the joints.

Kakadiaris and Metaxas used generic deformable models to estimate 3D human shape from silhouette contours taken from multiple camera views [11] and tracked these shapes over multiple frames [12]. Their approach involved a 2-stage process of first fitting the 3D shape and then tracking it. In contrast, pose and shape estimation are performed simultaneously in our method. Their experiments focused on upper-body tracking in simplified imaging environments in which near-perfect background subtraction results could be obtained.

In related work Plänkers and Fua [15] defined a "soft" body model using 3D Gaussian blobs arranged along an articulated skeletal body structure. The relative shapes of these "metaballs" were defined *a priori* and were then scaled for each limb based on an estimated length and width

parameter for that limb. Left and right limbs were constrained to have the same measurements. The surface of the body model was then defined implicitly as a level surface and an iterative optimization method was proposed to fit each limb segment to silhouette and stereo data. Most experiments used only upper body motion with simplified imaging environments, though some limited results on full body tracking were reported in [16].

Also closely related to the above methods is the work of Hilton *et al.* [10] who used a VRML body model. Their approach required the subject to stand in a known pose for the purpose of extracting key features from their silhouette contour which allowed alignment with the 3D model. Their model has a similar complexity to ours (∼20K polygons) but lacks the detail of the learned SCAPE model.

In these previous models the limb shapes were modeled independently as separate parts. This causes a number of problems. First, this makes it difficult to properly model the shape of the body where limbs join. Second, the decoupling of limbs means that these methods do not model pose dependent shape deformations (such as the bulging of the biceps during arm flexion). Additionally none of these previous method automatically estimated 3D body shape using learned models. Learning human body models has many advantages in that there are strong correlations between the size and shape of different body parts; the SCAPE model captures these correlations in a relatively low-dimensional body model. The result is a significantly more realistic body model which both better constrains and explains image measurements and is more tolerant of noise. In previous work, generic shape models could deform to explain erroneous image measurements (e.g. one leg could be made fatter than the other to explain errors in silhouette extraction). With the full, learned, body model, information from the entire body is combined to best explain the image data, reducing the effect of errors in any one part of the body; the resulting estimated shape always faithfully represents a natural human body. The SCAPE representation generalizes (linearly) to new body shapes not present in the training set.

Finally, there have been several non-parametric methods for estimating detailed 3D body information using voxel representations and space carving [3, 4, 5, 13]. While flexible, such non-parametric representations require further processing for many applications such as joint angle extraction or graphics animation. The lack of a parametric shape model means that it is difficult to enforce global shape properties across frames (e.g. related to the height, weight and gender of the subject). Voxel representations are typically seen as an intermediate representation from which one can fit other models [6, 21]. Here we show that a detailed parametric model can be estimated directly from the image data.

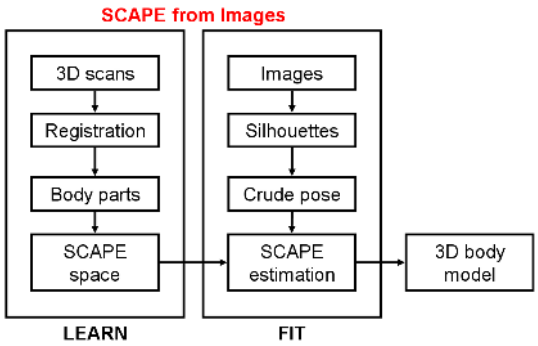



Figure 2. **Algorithm Overview.** A learning phase is used to build the 3D body model from range scans and follows the approach proposed in [1]. Our contribution provides a method for fitting the pose and shape parameters of the model to image data.

## 3. SCAPE Body Model

We briefly introduce our implementation of the SCAPE body model and point the reader to [1] for details. Our approach to 3D human shape and pose estimation has two main phases (Figure 2): A learning phase in which the human shape space is modeled, and a fitting phase in which body model parameters are estimated to match the observed shape in images.

The first phase involves learning the SCAPE model from 3D scans acquired using a Cyberware whole body scanner and merged into triangular meshes. The meshes are divided into two sets (Figure 3): A *pose set* containing the same subject in 70 diverse poses, and a *body shape set* containing 10 people with distinctive body shape characteristics standing in roughly the same standard pose. The former is used to model pose-induced variations of shape, while the latter is used to model shape variation between different people.

We define a *template mesh* in a canonical standing pose that is present in both data sets. The template mesh is hole-filled and subsampled to contain 25,000 triangles with 12,500 vertices. The remaining *instance meshes* are brought into full correspondence with the template mesh using a non-rigid mesh registration technique [1]. A skeleton reconstruction algorithm [1] is applied to the *pose set* to segment the template mesh into 15 body parts and to estimate joint locations.

**SCAPE Overview.** The *template mesh* acts as a reference mesh that is morphed into other poses and body shapes to establish correspondence between all meshes. Let $(x_1, x_2, x_3)$ be a triangle belonging to the template mesh and $(y_1, y_2, y_3)$ be a triangle from an instance mesh. We define the two edges of a triangle starting at $x_1$ as $\Delta x_j = x_j - x_1,\ j = 2, 3$.

The deformation of one mesh to another is modeled as a sequence of linear transformations applied to the triangle edges of the template mesh:

$$\Delta y = RDQ\Delta x \tag{1}$$

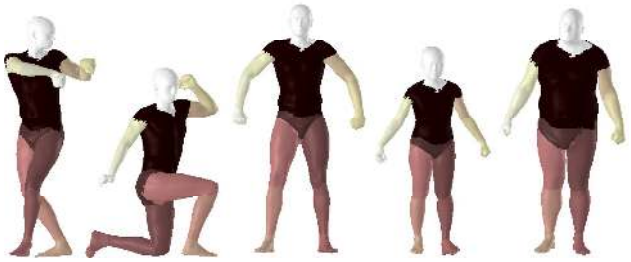


Figure 3. **3D Body Meshes.** Two example meshes from the pose set, the template mesh, and two example meshes from the body shape set (left to right).

where $Q$ is a $3 \times 3$ linear transformation matrix specific for this triangle corresponding to non-rigid pose-induced deformations such as muscle bulging. $D$ is a linear transformation matrix corresponding to body shape deformations and is also triangle specific. Finally, $R$ is a rigid rotation matrix applied to the articulated skeleton and specific to the body part containing the triangle.

**Rigid deformations.** Given an instance mesh $y$, the rigid alignment $R$ for each body part $b$ can be easily computed in closed form given the known point correspondences between $y$ and the template mesh [1].

**Non-rigid pose-dependent deformations.** Since the 70 meshes in the *pose set* belong to the same person as the template, their body shape deformation transformations $D$ are simply $3\times3$ identity matrices. Given the rigid alignment between meshes, the residual transformation $Q$ can be solved for by optimizing the deformation registering the template edges $\Delta x$ with the instance mesh edges $\Delta y$:

$$Q = \arg\min_{Q} \sum ||RQ\Delta x - \Delta y||^2 \qquad (2)$$

where the summation is over the edges of all triangles in the mesh (with some abuse of notation).

During video-based tracking, we will encounter new body poses not present in the training database and need to predict the pose-dependent deformation of the mesh. Consequently we use the 70 training examples to learn the coefficients $\alpha$ of a linear mapping from rigid body poses represented by $R$ to pose-dependent deformations $Q_\alpha(R)$. Then for any new pose we can predict the associated deformation.

**Non-rigid body shape-dependent deformations.** For each of the 10 instance meshes of different people in the *body shape set* we estimate the rigid alignment $R$ between parts and use this to predict the pose-dependent deformation $Q$ with the linear mapping from above. Then the shape-depended deformation $D$ is estimated as

$$D = \arg\min_{D} \sum ||RDQ\Delta x - \Delta y||^2. \qquad (3)$$

**Learning the SCAPE model.** Given the body shape deformations $D$ between different subjects in the *body shape set* and the template mesh, we construct a low dimensional linear model of the shape deformation using principal component analysis (PCA). Each $D$ matrix is represented as a column vector and is approximated as $D_{U,\mu}(\beta) = U\beta + \mu$ where $\mu$ is the mean deformation, $U$ are the eigenvectors given by PCA and $\beta$ is a vector of linear coefficients that characterizes a given shape. We keep the first 6 eigenvectors which account for $80\%$ of the total shape variance. Note that the shape coefficients for a specific person can be recovered by projecting the estimated deformation $D$ onto the PCA subspace.

Finally, a new mesh $y$, not present in the training set, can be synthesized given the rigid rotations $R$ and shape coefficients $\beta$ by solving

$$y(R,\beta) = \arg\min_{y} \sum ||RD_{U,\mu}(\beta)Q_\alpha(R)\Delta x - \Delta y||^2. \qquad (4)$$

This optimization problem can be expressed as a linear system that can be solved very efficiently.

## 4. Stochastic Optimization

During the fitting phase, estimating body shape and pose from image data involves optimization over the rigid limb transformations $R$, linear shape coefficients $\beta$, and global location $T$ of the body in the world coordinate system. We compactly represent the rotation matrices R using 37 Euler joint angles $r$ (after dropping some DOFs for non-spherical joints). We search for the optimal value for the state vector $s = (\beta, r, T)$ within a framework of *synthesis and evaluation*. For a predicted state $s$, a mesh is generated using Eq. 4, rendered to the image view given known camera calibration and compared to extracted image features.

State prediction is handled within an iterated importance sampling framework [7]. We represent a non-parametric state space probability distribution for state $s$ and image data $I$ as $f(s) \equiv p(I|s)p(s)$ with $N$ particles and associated normalized weights $\{s_i, \pi_i\}_{i=1}^N$. We note that we do not make any rigorous claims about our probabilistic model, rather we view the formulation here as enabling an effective method for stochastic search.

We define a Gaussian importance function $g^{(k)}(s)$ from which we draw samples at iteration $k$ of the search. This is initialized ($g^{(1)}(s)$) as a Gaussian centered on the pose determined by the initialization method (section 4.1) and the mean body shape ($\beta$ parameters zero). Particles are generated by randomly sampling from $g$ and normalizing the likelihood by the importance: $s_i \sim g(s), \pi_i = \frac{f(s_i)}{g(s_i)}$.

This process is made effective in an iterative fashion which allows $g$ to become increasingly similar to $f$. At iteration $k+1$, an importance function $g^{(k+1)}$ is obtained from the particle set at iteration $k$: $g^{(k+1)} = \sum_{i=1}^N \pi_i^{(k)} \mathcal{N}(s_i^{(k)}, \Sigma^{(k)})$.

To avoid becoming trapped in local optima, predicted particles are re-weighted using an annealed version of the likelihood function: $f^{(k)}(s) = \big(p(I|s)\big)^{t^{(k)}} p(s)$, where $t^{(k)}$ is an annealing temperature parameter optimized so that approximately half the samples get re-sampled.

### 4.1. Initialization

There exist a number of techniques that can be used to initialize the stochastic search; for example, pose prediction from silhouettes [19], voxel carving skeletonization [5], or loose-limbed body models [18]. Here we employ an existing human tracking algorithm [2] based on a cylindrical body model. The method is initialized in the first frame from marker data, and the position and joint angles of the body are automatically tracked through subsequent frames. The method uses an annealed particle filtering technique for inference, uses fairly weak priors on joint angles, enforces non-interpenetration of limbs and takes both edges and silhouettes into account. The recovered position and joint angles together with the mean body shape parameters are used to initialize the stochastic search of the SCAPE parameters.

## 5. Image Cost Function

We introduce a cost function $p(I|s)$ to measure how well a hypothesized model fits image observations. Here we rely only on image silhouettes which have been widely used in human pose estimation and tracking. The generative framework presented here, however, can be readily extended to exploit other features such as edges or optical flow.

Our cost function is a measure of similarity between two silhouettes. For a given camera view, a foreground silhouette $F^I$ is computed using standard background subtraction methods. This is then compared with the idealized silhouette $F^H$, generated by projecting a hypothesized mesh into the image plane. We penalize pixels in non-overlapping regions in one silhouette by the shortest distance to the other silhouette (cf. [19]) and vice-versa. To do so, we precompute a Chamfer distance map for each silhouette, $C^H$ for the hypothesized model and $C^I$ for the image silhouette. This process is illustrated in Figure 4.

The predicted silhouette should not exceed the image foreground silhouette (therefore minimizing $F^H \cdot C^I$), while at the same time try to explain as much as possible of it (thus minimizing $F^I \cdot C^H$). Both constraints are combined into a cost function that sums the errors over all image pixels $p$

$$-\log p(I|s) = \frac{1}{|p|} \sum_p \Big( a F_p^H \cdot C_p^I + (1-a) F_p^I \cdot C_p^H \Big), \quad (5)$$

where $a$ weighs the first term more heavily because the image silhouettes are usually wider due to the effects of clothing. When multiple views are available, the total cost is taken to be the average of the costs for the individual views.

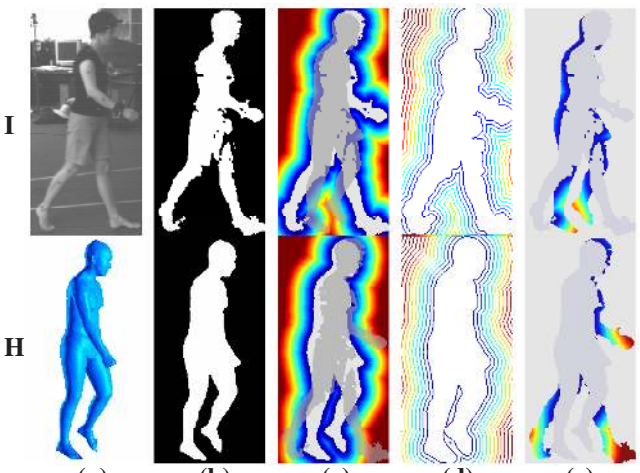



Figure 4. **Cost function.** **(a)** original image $I$ (top) and hypothesized mesh $H$ (bottom); **(b)** image foreground silhouette $F^I$ and mesh silhouette $F^H$, with 1 for foreground and 0 for background; **(c)** Chamfer distance maps $C^I$ and $C^H$, which are 0 inside the silhouette; the opposing silhouette is overlaid transparently; **(d)** contour maps for visualizing the distance maps; **(e)** per pixel silhouette distance from $F^H$ to $F^I$ given by $\sum_p F_p^H \cdot C_p^I$ (top), and from $F^I$ to $F^H$ given by $\sum_p F_p^I \cdot C_p^H$ (bottom).

## 6. Results

Figure 5 shows representative results obtained with our method. With 3 or 4 camera views we recover detailed mesh models of three different people in various poses and wearing sports and street clothing; none of the subjects were present in the SCAPE training set. In contrast, voxel carving techniques require many more views to reach this level of detail. The results illustrate how the SCAPE model generalizes to shapes and poses not present in the training data. While we have not performed a detailed analysis of the effects of clothing, our results appear relatively robust to changes in the silhouettes caused by clothes. As long as some parts of the body are seen un-occluded, these provide strong constraints on the body shape; this is an advantage of a learned shape model.

Results for an entire sequence are shown in Figure 6. Even though the optimization was performed in each frame independently of the others frames, the body shape remained consistent between frames. In general, our framework is capable of explicitly enforcing shape consistency between frames. We can either process several frames in a batch fashion where the shape parameters are shared across frames or employ a prior in tracking that enforces small changes in shape over time; this remains future work.

### 6.1. Comparison with the Cylindrical Body Model

Figure 7 presents the results obtained for one frame in each camera view used. First, we note that the optimization

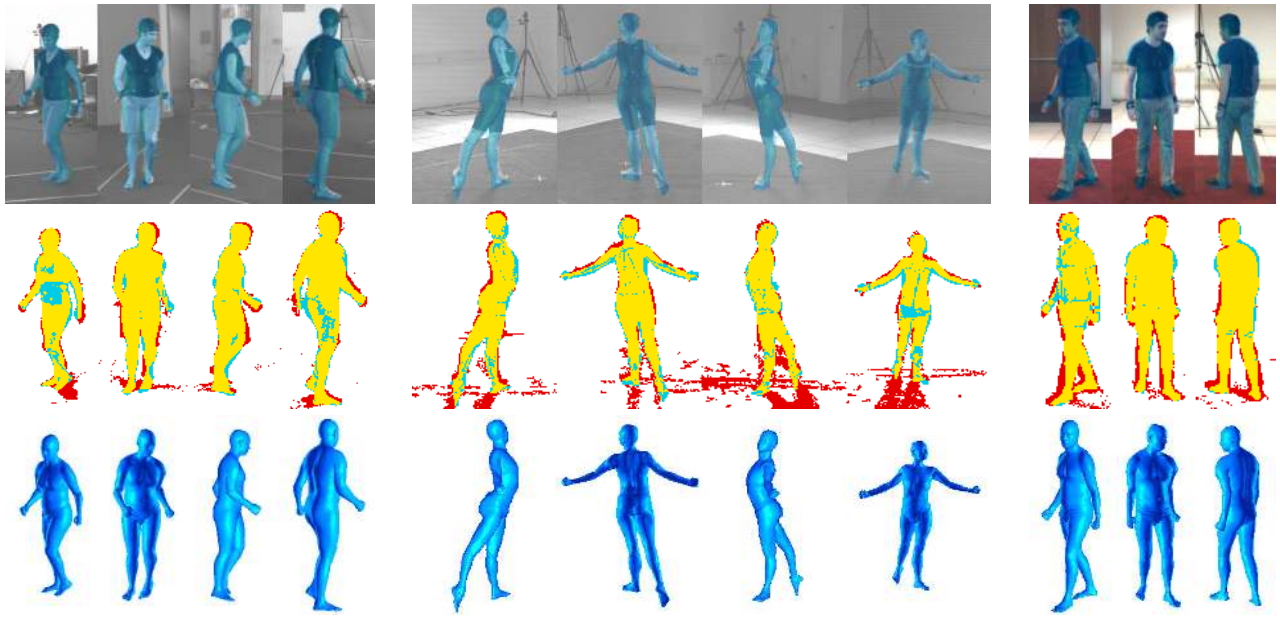


Figure 5. **SCAPE-from-image results.** Reconstruction results based on the views shown for one male and two female subjects, in walking and ballet poses, wearing tight fitting as well as baggy clothes. *(top)* Input images overlaid with estimated body model. *(middle)* Overlap (yellow) between silhouette (red) and estimated model (blue). *(bottom)* Recovered model from each camera view.

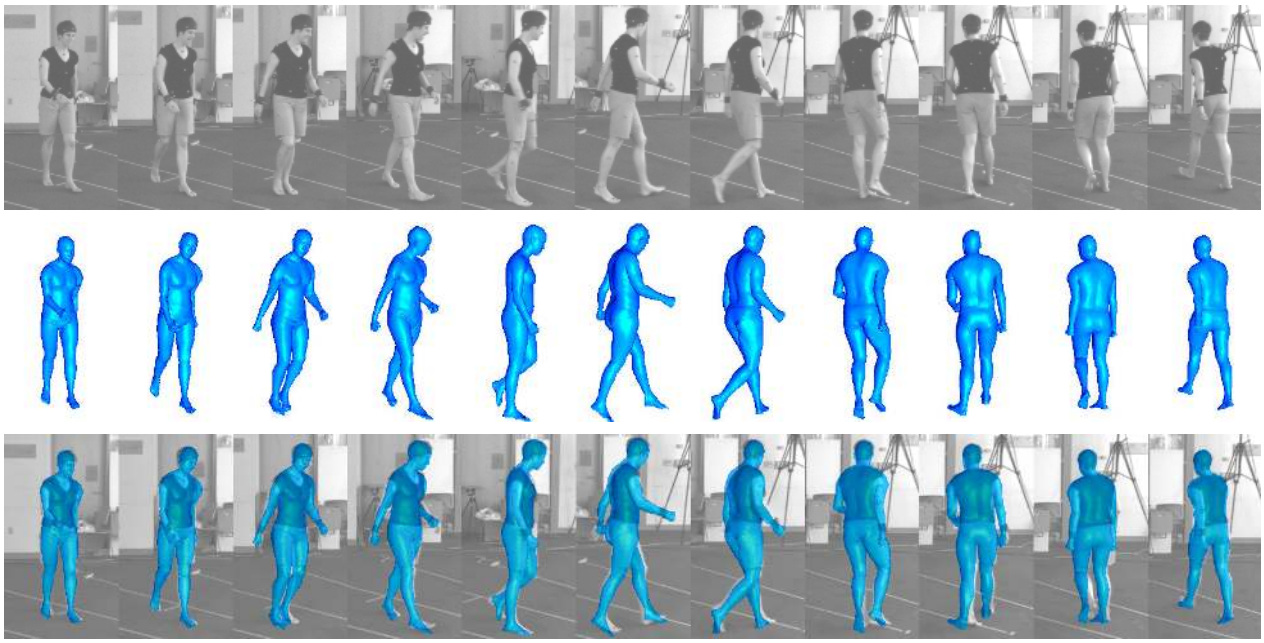


Figure 6. **First row:** Input images. **Second row:** Estimated mesh models. **Third row:** Meshes overlaid over input images. By applying the shape parameters recovered from 33 frames to the template mesh placed in a canonical pose, we obtained a shape deviation per vertex of $8.8 \pm 5.3mm$, computed as the mean deviation from the average location of each surface vertex.

can tolerate a significant amount of noise in the silhouettes due to shadows, clothing and foreground mis-segmentation. Second, the figure illustrates how the fitted SCAPE body model is capable of explaining more of the image foreground silhouettes than the cylindrical model. This can potentially make the likelihood function better behaved for the SCAPE model. To quantify this, we have computed how much the predicted silhouette overlapped the actual foreground (*precision*) and how much of the foreground was explained by the model (*recall*).

| 33 frames | Precision | Recall |
|---|---|---|
| Cylinder Model | 91.07% | 75.12% |
| SCAPE Model | 88.13% | 85.09% |

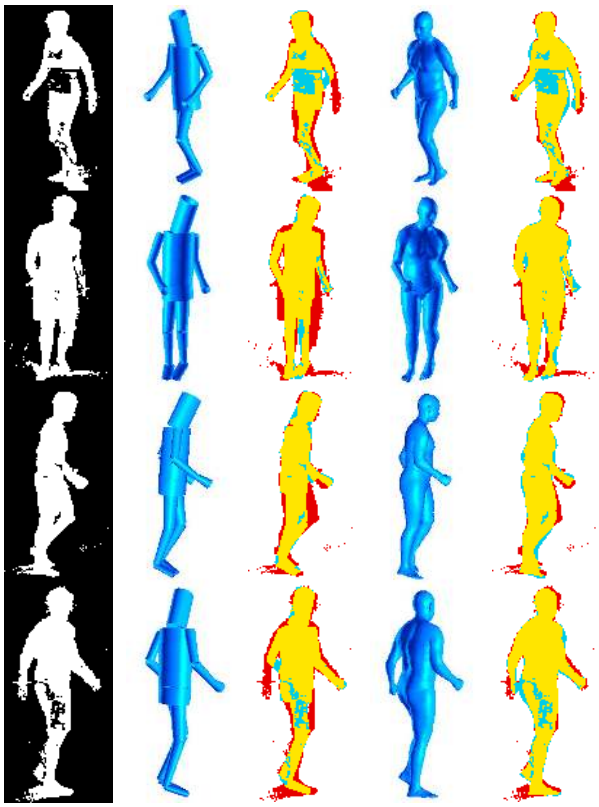


Figure 7. **Same pose, different camera views.** Each row is a different camera view. $1^{st}$ column: image silhouettes. $2^{nd}$ column: 3D cylindrical model. $3^{rd}$ column: overlap between image silhouettes and cylindrical model. $4^{th}$ column: 3D shape model. $5^{th}$ column: overlap between image silhouettes and SCAPE model.

The cylindrical model has 3% better precision because it is smaller and consequently more able to overlap the image silhouettes. On the other hand, the SCAPE model has 10% better recall because it is able to modify the shape to better explain the image silhouettes.

### 6.2. Convergence

We illustrate the process of convergence in Figure 8 in two different scenarios. The top row contains a real example of converging from the mean PCA shape and the pose estimated by the cylindrical tracker to the final fit of pose and shape to silhouettes. The bottom row shows synthetically generated silhouettes using a SCAPE model with shape parameters close to the initialized shape but with a distant pose. Except for the right leg which was trapped in local optimum, the likelihood formulation was able to attract the body and the right arm into place.

### 6.3. Anthropometric Measurements

Once the shape parameters have been estimated in each frame, we can then place the mesh with the corresponding

Figure 8. **Top:** Convergence from coarse tracking. **Bottom:** Convergence from a distant initial pose. In both cases the optimization is based on 4 views.

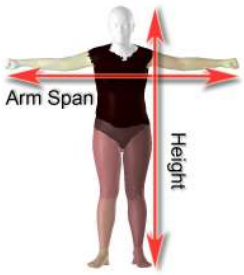



Figure 9. **T-pose.** Pose useful for extracting anthropometric measurements once shape was recovered from images.

shape in an appropriate pose for extracting anthropometric measurements. From the T-pose in Figure 9 we can easily measure the height and arm span for each shape.

| 33 frames | Actual | Mean | StDev |
|---|---|---|---|
| Height (mm) | 1667 | 1672 | 15 |
| Arm Span (mm) | 1499 | 1492 | 16 |

The actual values for the height and arm span are within half a standard deviation from the estimated values, with a deviation of less than $7mm$. For reference, one pixel in our images corresponds to about $6mm$.

Other measurements that could also be estimated are leg length, abdomen and chest depths, shoulder breadth etc. by measuring distances between relevant landmark positions on the template mesh, or mass and weight by computing the mesh volume. This suggests the potential use of the method for surveillance and medical applications.

### 6.4. Computational Cost

Most of the computing time is taken by the likelihood evaluations. Our stochastic search is over a 40-D pose space plus a 6-D shape space and we perform as many as 1,500 likelihood evaluations for one frame to obtain a good fit. Our implementation in Matlab takes almost a second per hypothesis. Half of that time is taken by a linear system solver for reconstructing the 3D mesh, and half is taken by

rendering it to a Z-Buffer to extract silhouettes in 4 views. Hardware acceleration together with partitioned sampling and a lower resolution mesh for early iterations would reduce the computing time.

## 7. Discussion and Conclusions

We have presented a method for estimating 3D human pose and shape from images. The approach leverages a learned model of pose and shape deformation previously used for graphics applications. The richness of the model provides a much closer match to image data than more common kinematic tree body models. The learned representation is significantly more detailed than previous non-rigid body models and captures both the global covariation in body shape and deformations due to pose. We have shown how a standard body tracker can be used to initialize a stochastic search over shape and pose parameters of this SCAPE model. Using the best available models from the graphics community we are better able to explain image observations and make the most of generative vision approaches. Additionally, the model can be used to extract relevant biometric information about the subject.

Here we worked with a small set of body scans from only ten subjects. We are currently working on using scans of over a thousand people to achieve a much richer model of human body shapes. Recovering a richer model will mean estimating more linear shape coefficients. To make this computationally feasible, we are developing a deterministic optimization method to replace the stochastic search used here. Currently we have not exploited graphics hardware for the projection of 3D meshes and the computation of the cost function; such hardware will greatly reduce the computation time required.

Here we did not impose constraints on the shape variation over time. In future work, we will explore the extraction of a single consistent shape model from a sequence of poses. Additionally, we will add interpenetration constraints while estimating the SCAPE parameters.

Our long term goal is to exceed the level of accuracy available from current commercial marker-based systems by using images which theoretically provide a richer source of information. We expect that, with additional cameras and improved background subtraction, the level of detailed shape recovery from video will eventually exceed that of marker-based systems.

## References


[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape completion and animation of people. *ACM Trans. Graphics*, 24(3):408–416, 2005. 1, 2, 3, 4

[2] A. Balan, L. Sigal and M. Black. A quantitative evaluation of video-based 3D person tracking. *VS-PETS*, pp. 349–356, 2005. 5

[3] G. Cheung, T. Kanade, J. Bouquet, and M. Holler. A real time system for robust 3D voxel reconstruction of human motions. *CVPR*, 2:714–720, 2000. 3

[4] K. M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time: Part II: Applications to human modeling and markerless motion tracking. *IJCV*, 63(3):225–245, 2005. 3

[5] C. Chu, O. Jenkins, and M. Matarić. Markerless kinematic model and motion capture from volume sequences. *CVPR*, II:475–482, 2003. 3, 5

[6] S. Corazza, L. Mündermann, A. Chaudhari, T. Demattio, C. Cobelli and T. P. Andriacchi. A markerless motion capture system to study musculoskeletal biomechanics: Visual hull and simulated annealing approach, *Annals Biomed. Eng.*, 34(6):1019–1029, 2006. 2, 3

[7] J. Deutscher, M. Isard, and J. MacCormick. Automatic camera calibration from a single Manhattan image. *ECCV*, pp. 175–205, 2002. 2, 4

[8] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2004. 1, 2

[9] D. M. Gavrila and L. S. Davis. 3D model-based tracking of humans in action: A multi-view approach. *CVPR*, pp. 73–80, 1996. 2

[10] A. Hilton, D. Beresford, T. Gentils, R. J. Smith, W. Sun, and J. Illingworth. Whole-body modelling of people from multi-view images to populate virtual worlds. *The Visual Computer*, 16(7):411–436, 2000. 3

[11] I. Kakadiaris and D. Metaxas. 3D human model acquisition from multiple views. *IJCV*, 30(3):191–218, 1998. 2

[12] I. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion. *PAMI*, 22(12):1453–1459, 2000. 2

[13] I. Mikić, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data. *IJCV*, 53(3):199–233, 2003. 3

[14] A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *PAMI*, 13(7):730–742, 1991. 2

[15] R. Plänkers and P. Fua. Articulated soft objects for video-based body modeling. *ICCV*, 1:394–401, 2001. 2

[16] R. Plänkers and P. Fua. Tracking and modeling people in video sequences. *CVIU*, 81(3):285–302, 2001. 3

[17] B. Rosenhahn, U. Kersting, K. Powel, and H.-P. Seidel. Cloth X-ray: Mocap of people wearing textiles. *DAGM*, pp. 495–504, 2006. 2

[18] L. Sigal, B. Sidharth, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. *CVPR*, I:421–428, 2004. 5

[19] C. Sminchisescu and A. Telea. Human pose estimation from silhouettes a consistent approach using distance level sets. *WSCG Int. Conf. Computer Graphics, Visualization and Computer Vision*, 2002. 5

[20] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *Int. J. Robotics Research*, 22(6):371–393, 2003. 2

[21] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. *CVPR*, 2:915–922, 2003. 3