The Naked Truth: Estimating Body Shape Under Clothing

Alexandru O. Bălan and Michael J. Black

Department of Computer Science, Brown University, Providence, RI 02912, USA
{alb,black}@cs.brown.edu

Abstract. We propose a method to estimate the detailed 3D shape of a person from images of that person wearing clothing. The approach exploits a model of human body shapes that is learned from a database of over 2000 range scans. We show that the parameters of this shape model can be recovered independently of body pose. We further propose a generalization of the visual hull to account for the fact that observed silhouettes of clothed people do not provide a tight bound on the true 3D shape. With clothed subjects, different poses provide different constraints on the possible underlying 3D body shape. We consequently combine constraints across pose to more accurately estimate 3D body shape in the presence of occluding clothing. Finally we use the recovered 3D shape to estimate the gender of subjects and then employ gender-specific body models to refine our shape estimates. Results on a novel database of thousands of images of clothed and "naked" subjects, as well as sequences from the HumanEva dataset, suggest the method may be accurate enough for biometric shape analysis in video.

1 Introduction

We address the problem of reliably estimating a person's body shape from images of that person wearing clothing. Estimation of body shape has numerous applications particularly in the areas of tracking, graphics, surveillance and forensic video analysis. To be practical any method for estimating body shape must recover a representation that is invariant to changes in body pose. To that end, we exploit a parametric 3D body model and show that the 3D body shape parameters are largely invariant to body pose. Additionally, such a model must be robust to clothing which obscures the true body shape. Here we build on the concept of the visual hull, which represents a bound on the underlying shape. In the case of a clothed person, the visual hull may only provide a loose bound on body shape. To gain tighter bounds we exploit the pose-invariance of the shape model to combine evidence from multiple poses. As a person moves, the constraints provided by the visual hull change as the clothing becomes looser or tighter on different body parts. We combine these constraints to estimate a *maximal silhouette-consistent parametric 3D shape*. Using a unique dataset of subjects with both minimal and normal clothing we demonstrate that a person's body shape can be recovered from several images of them wearing clothes.

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 15–29, 2008.

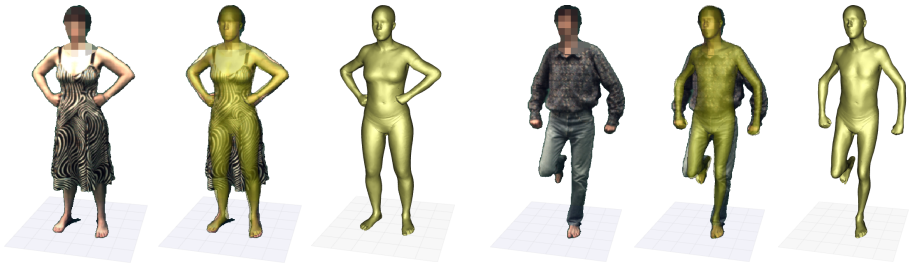

Fig. 1. Shape under clothing. Body shape recovery for two clothed subjects. For each subject we show (left to right) one of four input images, the 3D body model superimposed on the image (giving the impression of "X-ray" vision), and the estimated body model.

To our knowledge this is the first work to attempt to recover a detailed estimate of a person's 3D body shape from natural (ie. standard CCD) images when the body is obscured by clothing. The approach is illustrated in Figure 1. This work is based on a previously proposed parametric 3D body model called SCAPE [1]. The model characterizes human body shape using a low-dimensional shape subspace learned from a database of over 2000 range scans of humans. Recent work has shown that the parameters of this model can be directly estimated from image silhouettes [2,3] or from visual hulls [4] but has been restricted to people wearing tight-fitting clothing.

Our method rests on two key hypotheses. First: human body shape can be recovered independently of body pose. We test this hypothesis using a unique dataset of "naked" subjects in several poses captured with 4 calibrated cameras. We estimate their body shape in each pose both independently and in a batch fashion that combines information from multiple poses. We find that the variability in shape parameters across pose is small relative to the variability across subjects. We exploit this relative pose-independence of shape to combine multiple poses to more accurately estimate a single 3D body shape.

The second hypothesis is that images of the human body in clothing provide sufficient constraints to infer the likely 3D body shape. Of course a garment or costume could be worn which completely obscures or provides false information about the body shape. In normal "street clothes" however, we argue that many constraints exist that can be combined with a learned model of 3D body shape to infer the true underlying shape. To formalize this, we define the notion of a maximal silhouette-consistent parametric shape (MSCPS) that generalizes the notion of a visual hull. A visual hull has two properties [5,6]. First, the true 3D object lies completely within the visual hull (and its projection into images lies within the image silhouettes). Second, each facet of the visual hull touches the surface of the object. In the case of clothing, property 1 holds but property 2 does not. The object itself is obscured such that the silhouette contours may, or may not, correspond to true object boundaries. Rather, the silhouettes provide a bound on the possible shape which may or may not be a tight bound. Note

also that, with clothing, in some poses the bound may be tight in some places and loose in others and these locations may change with pose.

In place of the visual hull, we define the MSCPS that optimizes the following weak constraints: 1) the shape lies inside the visual hull; 2) the volume of the shape is maximal; and 3) the shape belongs to a parametric family. In our case this family is the family of 3D human body shapes. Constraint 2 is required to avoid the trivial solution where the estimated shape is made arbitrarily small. In general, each of these constraints can be viewed as a weak constraint with the last one being a statistical prior over 3D shapes. We go a step beyond previous work to deal with time-varying constraints and non-rigid, articulated objects, by requiring constraint 1 hold over multiple poses. We also use the fact that portions of the body may actually provide tight constraints on shape. For example, when a person wears short sleeves, their bare arms provide cues not only about the arm shape, but also about the overall weight of the person. Consequently we automatically detect skin regions and exploit tight constraints in these regions.

Central to our solution is a learned human body model. We go beyond previous work to use three different models: one for men, one for women, and one gender-neutral model combining both men and women. We find that gender can be reliably inferred in most cases by fitting both gender-specific models to the image data and selecting the one that best satisfies all the constraints. Given this estimated gender, we then use a gender-specific model to produce a refined shape estimate. To our knowledge this is the first method to estimate human gender directly from images using a parametric model of body shape.

In summary, the key contributions described here include: a shape optimization method that exploits shape constancy across pose; a generalization of visual hulls to deal with clothing; a method for gender classification from body shape; and a complete system for estimating the shape of the human body under clothing. The method is evaluated on thousands of images of multiple subjects.

2 Previous Work

There are various sensing/scanning technologies that allow fairly direct access to body shape under clothing including backscatter X-ray, infra-red cameras and radio waves. While our body fitting techniques could be applied to these data, for many applications, such as forensic video analysis, body shape must be extracted from standard video images. This problem is relatively unexplored. Rosenhahn *et al.* [7] proposed a method to track lower limbs for a person wearing a skirt or shorts. Their approach uses a generative model to explicitly estimate parameters of the occluding clothing such as the cloth thickness and dynamics. In their work, they assume the shape of the body and cloth measurements are known *a priori* and do not estimate them from image evidence. There has been recent interest in generative models of cloth [8,9] but the huge variability in clothing appearance makes the use of such models today challenging.

Most human shape estimation methods attempt to estimate the shape *with* the clothing and many of these techniques are based on visual hulls [10]. Visual hull

methods (including voxel-based and geometric methods) attempt to reconstruct the *observed* 3D shape with the silhouette boundary providing an outer bound on that shape. A detailed review is beyond the scope of this paper. We focus instead on those methods that have tried to restrict the shape lying inside the visual hull. Several authors have noted that, with small numbers of images, the visual hull provides a crude bound on object shape. To address this in the case of people, Stark and Hilton [11] combine silhouettes with internal structure and stereo to refine the 3D surface. They still assume the true surface projects to match the image silhouette features.

More generally, Franco *et al.* [12] impose weak assumptions on the underlying shape. They define a notion of a set of visual shapes that are consistent with the observed silhouettes (silhouette-consistent). The key contribution of their work is the idea of adding an assumption of shape smoothness which regularizes the set of possible 3D shape solutions. The observed silhouette is always considered as providing a tight bound on the surface with the priors compensating for an impoverished set of views. Note, however, in our problem, the visual hull is not the goal. Our case is different in that the object we care about (the human body) is obscured (by clothing) meaning that observed silhouette boundaries often do not provide tight constraints. We build on the notion of a visual shape set to define a person-specific prior model of the underlying shape.

In related work Corazza *et al.* [4] fit a SCAPE body model to visual hulls extracted using eight or more cameras. They do this in a single pose and assume tight-fitting clothing. We use a more detailed body model than they did and do not explicitly reconstruct the visual hull. Instead, we fit directly to image data and this allows us to use a smaller number of cameras (4 in our case).

Most visual hull reconstruction methods assume rigid objects. With non-rigid clothing we find it important to integrate information over time to constrain the underlying 3D shape. In related work, Cheung *et al.* [13] combine information over time by performing rigid alignment of visual hulls at different time instants. Knowing a rigid alignment over time effectively provides additional views. They also extend this idea to articulated body parts but focus only on recovering the bounding volume. Grauman *et al.* [14,15] estimate a 3D shape consistent with a temporal sequence of silhouettes using assumptions on the smoothness and shape transitions. They apply this method to silhouettes of humans and recover a visual hull using an example-based non-parametric model of body shapes. They do not use a parametric body model or explicitly attempt to infer the shape under clothing.

Most previous work on gender classification from images has focused on faces (e.g. [16]), but in many situations the face may be too small for reliable classification. The other large body of work is on estimating gender from gait (e.g. [17]). Surprisingly, this work typically takes silhouettes and extracts information about gait while throwing away the body shape information that can provide direct evidence about gender. We believe ours is the first method to *infer* a parametric 3D human body shape from images of clothed people and to use it for gender classification.

Fig. 2. Dataset. Example images from the clothing dataset shown here after background subtraction. *(top)* All subjects in the "naked condition" (NC); *(middle)* single subject in the "clothed condition" (CC); *(bottom)* a variety of the 11 different poses.

3 Methods

Datasets. To test our ability to infer shape under clothing we collected a dataset of 6 subjects (3 male and 3 female); a small sample of images from the dataset is shown in Figure 2. Images of the subjects were acquired in two conditions: 1) a "naked condition" (NC) where the subjects wore minimal tight fitting clothing (Figure 2 (top)) and 2) a "clothed condition" (CC) in which they wore a variety of different "street" clothes (Figure 2 (middle)). Each subject was captured in each condition in a fixed set of 11 postures, several of which are shown in Figure 2 (bottom). All postures were performed with 6 - 10 different sets of "street" clothing (trials) provided by the subjects. Overall, the dataset contains 53 trials with a total of 583 unique combinations of people, clothing and pose (a total of 2332 images). For each of these, images were acquired with four hardware synchronized color cameras with a resolution of 656 x 490 (Basler A602fc, Basler Vision Technologies). A full green-screen environment was used to remove any variability due to imprecise foreground segmentation. The cameras as well as the ground plane were calibrated using the Camera Calibration Toolbox for Matlab [18] and the images were radially undistorted. Foreground silhouette masks were obtained using a standard background subtraction method, performed in the *HSV* color space to account for background brightness variations induced by the presence of the foreground in the scene (e.g. shadows).

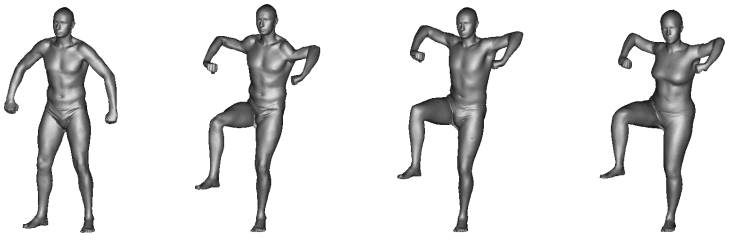

Fig. 3. SCAPE deformation process. A series of transformations are applied in succession to the reference mesh, x, including articulated rotations, R, non-rigid deformations, Q, and identity shape changes, D.

To test the generality of our methods in less than ideal circumstances, we also performed experiments on the generic HumanEva dataset [19] where the poses were "natural" and background subtraction was imperfect. This dataset only contains 2 subjects, but we tested our approach on approximately 200 frames in a wide variety of postures and with various levels of real silhouette corruption.

Parametric Body Model: SCAPE. We employ a parametric body model called SCAPE [1] that is able to capture both variability of body shapes between people, as well as articulated and non-rigid pose deformations; for more detail, the reader is referred to the original SCAPE paper. The model is derived from a large training set of human laser scans, which have been brought in full correspondence with respect to a reference mesh. Given a mesh x we deform it into a mesh y that has a different body shape and different pose. The deformation process (Figure 3) is defined as a set of transformations applied in succession to edges Δx of each triangle of the reference mesh, followed by a least-squares optimization that reconstructs a consistent mesh y:

$$y(\theta, \beta) = \arg\min_{y} \sum ||R^p(\theta) D^t_{U,\mu}(\beta) Q^t_\alpha(\theta) \Delta x - \Delta y||^2 \ . \tag{1}$$

The mesh is divided into 15 body parts, with the orientations specified by 3×3 part rotations $R^p(\theta)$ computed from a reduced set of joint angles θ of a kinematic skeleton that has 32 degrees of freedom. Non-rigid pose-dependent deformations $Q^t_\alpha(\theta)$ are linearly predicted for each triangle t from neighboring joint angles, with the linear coefficients α learned from training data of a single subject scanned in 70 different poses. Changes in 3D body shape between people are captured by shape deformations D^t which are 3×3 transformations that are applied to each triangle t in the reference mesh to deform it into an instance mesh (i.e. 3D range scan). We construct a training set of such deformations between the instance mesh and over 2000 body scans of North American adults with roughly equal gender representation (Civilian American and European Surface Anthropometry Resource (CAESAR), SAE International). We learn a low-dimensional

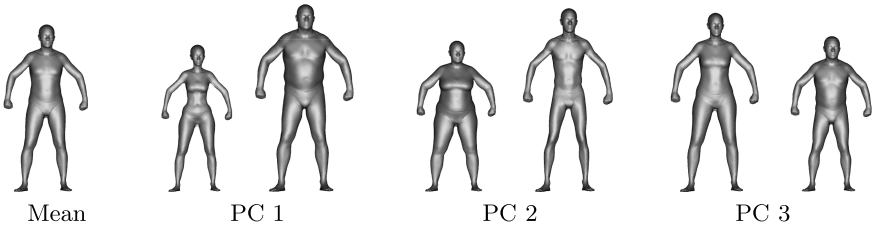

Fig. 4. 3D Body Shape Model. The mean gender-neutral body shape is shown *(left)* followed by deviations from the mean along the first three principal components (± 3 standard deviations).

linear model of human shape variability using incremental principal component analysis (iPCA) [20]. For each individual, the D^t matrices are concatenated into a single column vector that is approximated as $D_{U,\mu}(\beta) = U\beta + \mu$ where μ is the mean body shape, U are the first n eigenvectors given by iPCA and β is a vector of linear coefficients that characterizes a given shape. The variance of each shape coefficient β_j is given by the eigenvalues $\sigma^2_{\beta,j}$.

We learn separate models for male and female subjects, as well as a gender-neutral model with all the subjects. For our experiments we use the first $n = 20$ principal components which account for roughly 70% of the variance in the gender-neutral case and 65% of the variance in the gender specific cases. Figure 4 shows the mean gender-neutral shape and deviations from the mean captured by the first three principal components.

Objective Function. Our model is parameterized by a set of joint angles θ, including global position and orientation, and shape parameters β. We adopt a generative approach in which the parameters define a 3D body pose and shape model which is then projected into the camera views to produce predicted image silhouettes $S^e_{k,\beta,\theta}$. These are then compared with observed silhouettes, S^o_k, obtained by foreground segmentation in each camera view k. Specifically we define the asymmetric distance between silhouettes S and T as

$$\tilde{d}(S,T) = \sum_{i,j} \big(S_{ij} \cdot C_{ij}(T)\big) \big/ \big(\sum_{i,j} S_{ij}\big)^{3/2} \ , \tag{2}$$

where S_{ij} are the pixels inside silhouette S and $C_{ij}(T)$) is a distance function which is zero if pixel (i,j) is inside T and is the Euclidean distance to the closest point on the boundary of T for points outside. The denominator is a normalization term that gives invariance to the size of the silhouette.

We first define the objective function for the NC using the bi-directional objective used by Bălan *et al.* [2]. Later we will extend this to deal with clothing. The objective function uses a symmetric distance to match the estimated and observed silhouettes over the K camera views:

$$E(\beta,\theta) = \sum_{k=1}^{K} \tilde{d}(S^e_{k,\beta,\theta}, S^o_k) + \tilde{d}(S^o_k, S^e_{k,\beta,\theta}) \ . \tag{3}$$

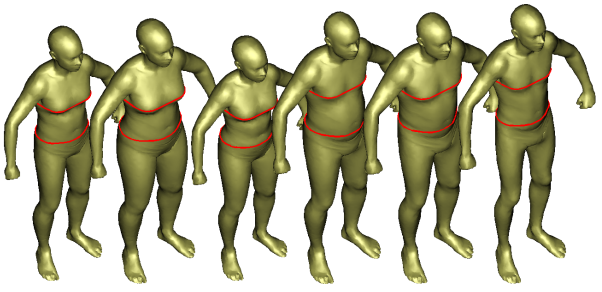

Fig. 5. Quantitative Evaluation. Ground truth body shapes are shown for the six subjects. For each subject the height, waist size and chest size are used for quantitative evaluation. The waist and chest are shown by red lines.

Optimization Method. Unlike previous work we minimize $E(\beta, \theta)$ using a gradient-free direct search simplex method. For the clothing dataset, 11 canonical pose parameters θ^0 were hand defined and used for initialization of each pose. To help avoid local minima we alternate between optimizing pose and shape in an incremental fashion: after initializing with the predefined pose and mean shape model, we begin by optimizing the global position and a few shape coefficients. We then continue optimizing the joint angles of individual joints along the kinematic chain (hips, shoulders, elbows), and then jointly optimize all joint angles together with a few more shape coefficients. In the last phase we optimize the full set of 20 principal shape components.

Evaluation Measures. We quantitatively evaluate the accuracy of our 3D body models using a variety of derived biometric measurements such as height, waist size, chest size, etc. These body measurements were collected directly from the subjects, from laser range scans of four of the subjects, and from the results of fitting the body in the NC. We verified the accuracy of the body fit in the NC using the direct measurements and then took this to be the ground truth for the remainder of the paper. Figure 5 shows the recovered 3D body shape models used for ground truth. Displayed on the models in red are the locations used to compute derived measurements for chest and waist size.

4 Shape Constancy

Previous work [2] only optimized the body shape at a particular time instant. Here we take a different approach and integrate information about body shape over multiple poses. Our first hypothesis is that the SCAPE model provides a representation of body shape that is invariant to body pose. To test this hypothesis we optimize body shape and pose for each posture independently in the NC. Figure 6 (left) shows three examples of the 3D body shape recovered for one of the subjects in this fashion. Figure 6 (right) plots the variance in the recovered shape coefficients across pose for all subjects (yellow) versus the

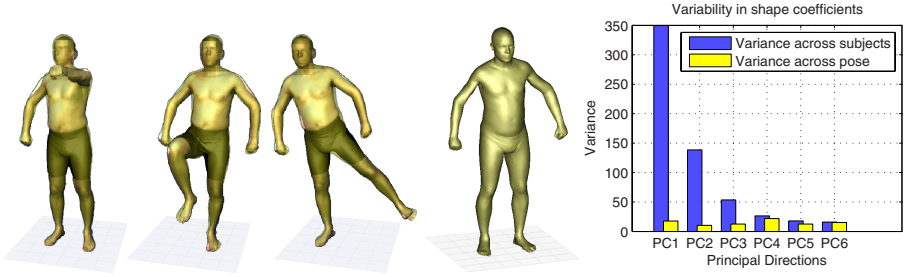

Fig. 6. Invariance of body shape to pose. *(left)* Reconstructions of 3D body shape for three individual poses using four camera views. *(middle)* Model recovered by combining across 11 different poses. *(right)* Variance in shape coefficients across subjects and poses.

variability in these shape coefficients across subjects. The variation of the major 3D shape parameters with pose is small relative to the variation across people.

To exploit this constancy we define a "batch" optimization that extends the objective function to include P different poses:

$$E(\beta, \theta_1, \ldots, \theta_P) = \sum_{p=1}^{P} \sum_{k=1}^{K} \tilde{d}(S^e_{k,\beta,\theta_p}, S^o_{k,p}) + \tilde{d}(S^o_{k,p}, S^e_{k,\beta,\theta_p}) \ . \tag{4}$$

We alternate between optimizing a single set of shape parameters β applicable to all postures, and optimizing the pose parameters θ_p independently for each posture. Figure 6 (middle) shows the body shape recovered by integrating across pose. Note that establishing this property of shape invariance with respect to pose is useful for tracking applications and biometric shape analysis.

5 Clothing

Above we established the basic model, its optimization and its application to shape estimation in the absence of loose clothing. We now turn to our main contribution which is a method for body shape recovery with clothing.

Maximal Silhouette-Consistent Parametric Shape. Laurentini [6] introduced the concept of *visual hull* to represent the maximal 3D object shape consistent with observed image silhouettes. Our assumption is that the true object shape belongs to a parametric family and lies completely within the visual hull. We also generalize the concept to cases where the visual hull only provides a loose bound to the object's surface. Since the problem becomes ill-posed in this case, we regularize the problem by making use of prior knowledge of the object shape and integrate information from time-series data. Specifically, we introduce the concept of a *maximal silhouette-consistent parametric shape* that weakly satisfies three constraints: 1. the projected model falls completely inside

the foreground silhouettes; 2. the volume of the 3D model is maximal; and 3. the shape of the object belongs to a parametric family of shapes (in our case human bodies).

We satisfy the first constraint by penalizing the regions of the projected model silhouette, S^e, that fall outside the observed foreground silhouette S^o. That is, we use exactly the same distance function as defined above: $E_{inside}(\beta,\theta) = \tilde{d}(S^e_{k,\beta,\theta}, S^o_k)$. For the second constraint, we would like the projected model to explain as much of the foreground silhouette as possible; if the subject were not wearing clothing this would just be the second term from above $\tilde{d}(S^o_k, S^e_{k,\beta,\theta})$. In the more general setting where people wear clothing or interact with objects, the observed foreground silhouettes will be too large producing a bias in the shape estimates. To cope with this, we employ two strategies. The first is to down-weight the contribution of the second constraint, meaning it is more important for the estimated shape to project inside the image silhouette than to fully explain it. The second is to use features in the image that we are more confident accurately conform to the underlying shape. In particular, we detect skin-colored regions and, for these regions, we give the second constraint full weight. We denote by S^s the detected skin regions and by $S^o \backslash S^s$ non-skin regions of the observed foreground silhouette. The "expansion" constraint is then written as

$$E_{expand}(\beta,\theta) = \tilde{d}(S^s_k, S^e_{k,\beta,\theta}) + \lambda \tilde{d}(S^o_k \backslash S^s_k, S^e_{k,\beta,\theta}) \tag{5}$$

with $\lambda \ll 1$.

Finally to enforce domain knowledge, we define priors on shape parameters, as well as on joint angles. We use very weak priors designed to prevent wildly unnatural shapes but which do not bias the estimates for "normal" body shapes. Specifically we add a penalty on the shape coefficients β_i of the form

$$E_{shape}(\beta) = \sum_j \max\left(0, \frac{|\beta_j|}{\sigma_{\beta,j}} - \sigma_{thresh}\right)^2 . \tag{6}$$

No penalty is paid for small values of β and we only start penalizing shape coefficients when they exceed $\sigma_{thresh} = 3$ standard deviations from the mean. We also enforce a prior, $E_{pose}(\theta)$, on body pose which is uniform within joint angle limits and only penalizes poses beyond these limits.

Finally, the objective function we minimize is

$$E_{clothes}(\beta,\Theta) = \sum_{p=1}^{P}\sum_{k=1}^{K} \frac{E_{inside}(\beta,\theta_p)}{\sigma_o^2} + \frac{E_{expand}(\beta,\theta_p)}{\sigma_o^2} + E_{shape}(\beta) + E_{pose}(\theta_p)$$

where $\Theta = \theta_1, \ldots, \theta_P$ represents the different body poses. We assume a normal distribution over the silhouette distance metric $\tilde{d}$ with variance σ_o^2 that accounts for noisy image observations.

One advantage of our method is that being robust to clothing also provides some robustness to silhouettes that are too large (e.g. due to shadows). Also, in future work, we will learn a statistical model for E_{expand} that captures the deviations of clothing from the naked form observed in training data.

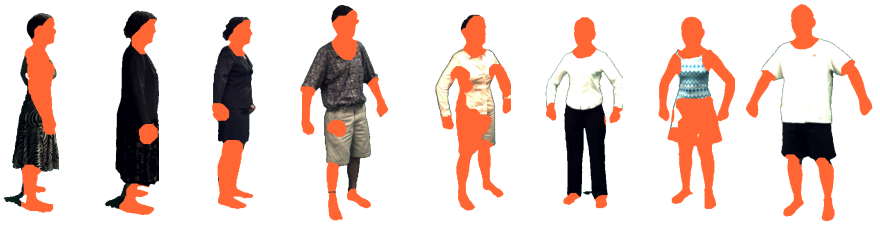

Fig. 7. Skin segmentation. Examples of segmented people and the regions identified as skin, shown in orange. Not that this segmentation need not be perfect to provide useful constraints on the shape fitting.

Skin Detection and Segmentation. In order to detect skin colored regions in an image, multiple skin detectors are trained. We use a leave-one-out cross-validation method and train one classifier for each person using all the other people in the database. Hence the skin of the left-out subject is segmented using a skin model that excludes his/her data. The skin detectors are built from training data using a simple non-parametric model of skin pixels in hue and saturation space. Figure 7 shows several examples of people in different clothing and the identified skin regions.

6 Results

Gender classification. For gender classification, we estimated the pose and the first 6 shape parameters in each *test instance* using the gender-neutral shape model. After convergence, we kept the pose parameters fixed and re-estimated the shape parameters with both gender-specific shape models. The best fitting model according to the objective function corresponded to the true gender 86% of the time when the optimization was performed on individual poses. Taking the majority classification across all poses in a trial increased the classification accuracy to 90.6%. Finally we estimated shape parameters in batch fashion over all poses in a trial, and the gender classification improved to 94.3% accuracy on the dataset of 53 trials with natural clothing.

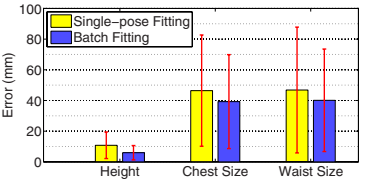

Fig. 8. Quantitative evaluation of shape. Accuracy of estimated body measurements (height, waist, chest) relative to ground truth. Batch estimation across pose decreases the errors.

Shape under Clothing. We recovered the body pose and shape for all 583 independent poses and the 53 batch trials in the CC. Fitting takes approximately 40min for a single model on a 2GHz processor. A few representative results[1] are shown in Figure 9. We quantitatively evaluate the accuracy with respect

[1] See http://www.cs.brown.edu/research/vision/scapeClothing for additional results.

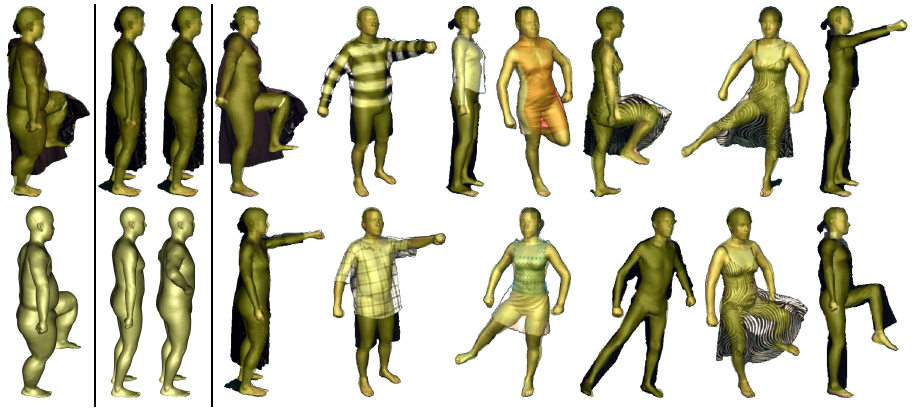

Fig. 9. Example shapes estimated by 3 different methods. *(left)* using the clothing-oblivious method of [2] for a single pose and without enforcing priors on shape leads to unrealistic overestimates of size/shape; *(middle)* estimated shapes using the proposed clothing-robust objective function formulation (7) applied only to individual poses; *(right)* shapes estimated using the multi-pose shape consistency constraint.

Fig. 10. Example body shapes from the HumanEva-II dataset. (*left*) One segmented frame; (*middle*) several frames with the estimated model overlaid; (*right*) estimated body shape.

to biometric measurements derived from the ground truth shape computed in batch fashion from the NC data. Figure 8 shows how errors in height, waist and chest size decrease by combining information across pose.

We also tested our method on the publicly available HumanEva-II dataset which consists of two sequences with the subjects S2 and S4 walking and jogging in a circle, followed by a leg-balancing action. A kinematic tree tracking algorithm using a coarse cylindrical body model was used to obtain rough initial pose estimates at each frame which were subsequently refined during shape estimation using our framework. We used a subset of the frames spanning a wide

variety of postures to estimate the body shape (Figure 10); these results suggest the method is relatively robust to errors in foreground segmentation. The optimization converged in all test cases we considered.

7 Conclusions

We defined a new problem of inferring 3D human shape under clothing and presented a solution that leverages a learned model of body shape. The method estimates the body shape that is consistent with an extended definition of the visual hull that recognizes that shape bounds provided by the visual hull may not be tight. Specifically, the recovered shape must come from the parametric family of human shapes, it should lie completely within the visual hull, and it should explain as much of the image evidence as possible. We observed that by watching people move, we could obtain more constraints on the underlying 3D body shape than in a single pose. Consequently, we exploited the relative pose-independence of our body shape model to integrate constraints from multiple poses by solving for the body pose at each time instant and a single 3D shape across all time instants. We integrated a skin detector to provide tight constraints on 3D shape when parts of the body are seen unclothed. We also showed that gender could be reliably classified based on body shape and defined gender-specific shape models to provide stronger priors on the unobserved shape. The method was tested on a laboratory dataset of people in a fixed set of poses and two clothing conditions: with and "without" clothes. The latter gave us "ground truth" with which to evaluate the method. We also demonstrated the method for more natural sequences of clothed people from the HumanEva dataset [19].

Privacy concerns must be addressed for any technology that purports to "see" what someone looks like under their clothes. Unlike backscatter X-ray and infrared sensors, our approach does not *see through* clothing. It does not have any information about the person's body that is not available essentially to the naked eye; in this sense it is not intrusive. The unwanted production of a likeness or facsimile of a person's unclothed shape might still be considered a violation of privacy. It is important to maintain a distinction between body shape measurements and the graphical representation of those measurements as a realistic 3D model; it is the latter which has the greatest potential for concern but this may not be needed for many vision applications.

There are many applications and future directions for this new line of research. For human tracking in video it may be useful to estimate limb lengths, body shape parameters and body mass as these could be used in the inference of dynamics. Body shape parameters could be used in visual tracking applications to identify and re-acquire subjects who come in and out of the field view. For forensic video applications, the extraction of body shape parameters could be useful in identifying suspects. There are also many applications of these methods in personal fitness, retail apparel and computer games. In future work we will extend these methods to extract body shape from monocular image sequences.

Acknowledgments. This work was supported in part by the Office of Naval Research (N00014-07-1-0803), NSF (IIS-0535075), by an award from the Rhode Island Economic Development Corporation (STAC), and by a gift from Intel Corp. The authors thank L. Reiss for assistance with image segmentation and biometric measurements.

References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: Shape completion and animation of people. SIGGRAPH 24, 408–416 (2005)
2. Bălan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: CVPR (2007)
3. Bălan, A.O., Black, M.J., Sigal, L., Haussecker, H.W.: Shining a light on human pose: On shadows, shading and the estimation of pose and shape. In: ICCV (2007)
4. Mündermann, L., Corazza, S., Andriacchi, T.: Accurately measuring human movement using articulated ICP with soft-joint constraints and a repository of articulated models. In: CVPR (2007)
5. Cheung, K.M., Baker, S., Kanade, T.: Visual hull alignment and refinement across time: A 3D reconstruction algorithm combining shape-from-silhouette with stereo. In: CVPR, vol. 2, pp. 375–382 (2003)
6. Laurentini, A.: The visual hull concept for silhouette-based image understanding. PAMI 16, 150–162 (1994)
7. Rosenhahn, B., Kersting, U., Powell, K., Klette, R., Klette, G., Seidel, H.P.: A system for articulated tracking incorporating a clothing model. Machine Vision and Applications (MVA) 18, 25–40 (2007)
8. Salzmann, M., Pilet, J., Ilic, S., Fua, P.: Surface deformation models for nonrigid 3D shape recovery. PAMI 29, 1481–1487 (2007)
9. White, R., Crane, K., Forsyth, D.: Capturing and animating occluded cloth. In: ACM Transactions on Graphics (TOG), SIGGRAPH (2007)
10. de Aguiar, E., Theobalt, C., Stoll, C., Seidel, H.P.: Marker-less deformable mesh tracking for human shape and motion capture. In: CVPR, pp. 2502–2509 (2007)
11. Stark, J., Hilton, A.: Surface capture for performance-based animation. IEEE Comp. Graphics and Applications 27, 21–31 (2007)
12. Franco, J.S., Lapierre, M., Boyer, E.: Visual shapes of silhouette sets. In: Proc. 3rd Int. Symp. on 3D Data Processing, Visualization and Transmission (2006)
13. Cheung, K.M., Baker, S., Kanade, T.: Shape-from-silhouette across time: Part II: Applications to human modeling and markerless motion tracking. IJCV 63, 225–245 (2005)
14. Grauman, K., Shakhnarovich, G., Darrell, T.: A Bayesian approach to image-based visual hull reconstruction. In: CVPR, vol. 1, pp. 187–194 (2003)
15. Grauman, K., Shakhnarovich, G., Darrell, T.: Inferring 3D structure with a statistical image-based shape model. In: ICCV, pp. 641–648 (2003)
16. Moghaddam, B., Yang, M.: Learning gender with support faces. PAMI 24, 707–711 (2002)
17. Li, X., Maybank, S., Yan, S., Tao, D., Xu, D.: Gait components and their application to gender recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 38, 145–155 (2008)

18. Bouguet, J.Y.: Camera calibration toolbox for Matlab. Technical report (CalTech)
19. Sigal, L., Black, M.J.: HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University (2006)
20. Brand, M.: Incremental singular value decomposition of uncertain data with missing values. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 707–720. Springer, Heidelberg (2002)