Contour People: A Parameterized Model of 2D Articulated Human Shape

[1]Oren Freifeld [2]Alexander Weiss [2,3]Silvia Zuffi [2]Michael J. Black
[1]Division of Applied Mathematics and [2]Department of Computer Science,
Brown University, Providence, RI 02912, USA
[3]ITC - Consiglio Nazionale delle Ricerche, Milan, Italy

Abstract

We define a new "contour person" model of the human body that has the expressive power of a detailed 3D model and the computational benefits of a simple 2D part-based model. The contour person (CP) model is learned from a 3D SCAPE model of the human body that captures natural shape and pose variations; the projected contours of this model, along with their segmentation into parts forms the training set. The CP model factors deformations of the body into three components: shape variation, viewpoint change and part rotation. This latter model also incorporates a learned non-rigid deformation model. The result is a 2D articulated model that is compact to represent, simple to compute with and more expressive than previous models. We demonstrate the value of such a model in 2D pose estimation and segmentation. Given an initial pose from a standard pictorial-structures method, we refine the pose and shape using an objective function that segments the scene into foreground and background regions. The result is a parametric, human-specific, image segmentation.

1. Introduction

The detection of people and the analysis of their pose in images or video has many applications and has drawn significant attention. In the case of uncalibrated monocular images and video, 2D models dominate while in calibrated or multi-camera settings, 3D models are popular. In recent years, 3D models of the human body have become sophisticated and highly detailed, with the ability to accurately model human shape and pose [5] (Fig. 1(c)). In contrast, 2D models typically treat the body as a collection of polygonal regions that only crudely capture body shape (Fig. 1(a)) [6, 14, 15, 21]. Two-dimensional models are popular because they are relatively low dimensional, do not require camera calibration, and admit computationally attractive inference methods (e.g. with belief propagation [1, 8, 17, 23]). For many problems such as pedestrian detection, full 3D reasoning many not be needed. While such 2D models predominate, they have changed little in 20 or more years [14, 15].

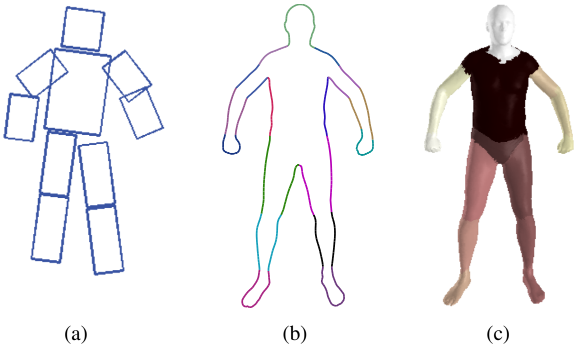

Figure 1. **Contour Person Model.** Most 2D body models (left) are simple articulated collections of geometric primitives while 3D models have become increasingly detailed and realistic (right). The contour person model (middle) has the realism of modern 3D models but with the computational benefits of a 2D model.

In this paper we describe a new 2D model of the body that has many of the benefits of the more sophisticated 3D models while retaining the computational advantages of 2D. This *Contour Person* (CP) model (Fig. 1(b)) provides a detailed 2D representation of natural body shape and captures how it varies across a population. It retains, however, the part-based representation of current 2D models as illustrated by the different colors in Fig. 1(b) and the banner. An articulated, part-based, model is required for pose estimation using inference methods such as belief propagation. Importantly, the CP model also captures the non-rigid deformation of the body that occurs with articulation. This allows the contour model to accurately represent a wide range of human shapes and poses. Like other 2D body models, the approach is inherently view-based with 2D models constructed for a range of viewing directions.

The contour person model factors changes in 2D body shape into a number of causes, with each cause represented using a low-dimensional model. These include shape changes due to 1) viewing direction; 2) identity or body shape; 3) rigid articulation; and 4) non-rigid deformation due to articulation. This is similar to recent work on 3D body shape representation using the SCAPE model [2]. In fact our 2D model is created from a 3D SCAPE model of the human body. However, rather than model deformations of triangles on a 3D mesh, we model deformations of line segments in 2D; this results in a simpler and lower-dimensional body model.

We envision many applications of the contour-person model. This paper focuses on the development of the model and the issues involved with representing an inherently 3D shape model in 2D while maintaining realism and accuracy. To illustrate the application of the model we present initial results in pose estimation and segmentation. To do so, we build on an existing state of the art person detector that uses a pictorial structures (PS) model [1]. This existing technique is used to initialize our model and then both the pose and shape of the CP model are refined using a parametric form of GrabCut [22]. Results of pose estimation and segmentation are shown on a variety of images and compared with pictorial structures for pose estimation and with traditional GrabCut for segmentation.

2. Related Work

Two-dimensional models of the human body are popular due to their representational and computational simplicity. Existing models include articulated pictorial structures models, active shape models (or point distribution models), parametrized non-rigid templates, and silhouette models. We focus here on models that explicitly represent shape with contours and, furthermore, those that have been used to represent non-rigid human shape and pose. There is an extensive literature on general contour models for object representation and recognition that we do not consider here.

2D articulated person models. Most 2D articulated person models have focused on estimating human pose and have ignored body shape. We argue that a good body shape representation can improve pose estimation by improving the fitting of the model to image evidence.

The first use of a human "puppet" model was due to Hinton [14] and there have been many related models since. The classic 2D model is the "cardboard person" [15], defined by a kinematic tree of polygonal regions, where each limb may be rotated or scaled in 2D. Similarly the scaled-prismatic model (SPM) treats the limbs as rigid templates that can be scaled in length [6]. Both the cardboard person and SPM approximate foreshortening caused by motion of the limbs in depth.

More restricted models, with only rotation at the joints (and a global scale), form the basis of most of the current pictorial structures (PS) models [10] used for detecting and tracking people in monocular imagery [1, 8, 9, 17, 21]. These models admit efficient search with belief propagation (BP) due to the simplification of the representation. Sigal and Black [23] use a 2D model that includes foreshortening and do inference with BP. The advantage of the richer model is that it allows better prediction of 3D pose from the estimated 2D model [23].

Our work falls solidly in the PS camp but increases the realism beyond previous methods by modeling shape variation across bodies as well as non-rigid deformation due to articulated pose changes.

Active shape and contour models. Active shape models (ASMs) capture the statistics of contour deformations from a mean shape using principal component analysis (PCA) [7]. PCA can be performed on points, control points or spline parameters. These models have been used extensively, particularly for representing human faces and their deformations [7]. Note that facial features deform in such models but there is no explicit representation of part rotation, they have little depth variation relative to each other, and there is no self occlusion. The articulated human body has all these issues.

Baumberg and Hogg were the first to use ASMs for representing the full human body [3]. Given a training set of pedestrians segmented from the background, they define contours around each with the same number of points and roughly the same starting locations. They analyze the modes of variation in this contour using PCA and use this model to track pedestrians.

In such a model, changes in body shape and pose are combined in one PCA representation. Furthermore, with no notion of body parts, the alignment between training body contours is difficult to establish. This results in principal components that capture the non-informative sliding of points along the contour. Finally this simple PCA model does not directly encode articulated body pose, limiting its use for human motion and gesture analysis.

Ong and Gong [20] extend these point distribution models to deal with articulated 3D human motion of the upper body. They construct a training vector that includes the contour points of the upper body, 2D points corresponding to the locations of the hands and head, and the 3D joint angles of an underlying articulated body model. To deal with the non-linearity of the contour with respect to pose, they use a hierarchical PCA method that finds linear clusters in the non-linear space. In contrast, we explicitly model the parts of the body and do not use PCA to capture articulations. Rather we use it to capture body shape variations (and camera view changes). This provides a blend between the PS models and the active contour methods.

Grauman *et al.* [12] map multi-view silhouettes to contours and learn a low dimensional shape representation in conjunction with 3D body pose. Like other methods they model shape in terms of the contour points. Our model differs in that it models *deformations* of 2D contours and this representation is important for explicitly modeling articulation and for factoring different types of deformation.

Human models and segmentation. We evaluate our model on the problem of segmentation; the contour of the body defines the region inside (and outside) the body. In early work, Kervrann and Heitz [16] define a non-rigid model of the hand and estimate both its pose and segmentation using motion and edge cues. The model is not part-based, the deformations are not learned, and it has a limited range of motion. Alternative formulations have explored template-based models of the body [11, 18] that are not fully articulated and do not factor shape and pose.

Of particular relevance is the recent work of Ferrari and Zisserman [9] that uses a weak detector to obtain a crude estimate of human pose in an image. This pose is then used to initialize GrabCut [22] segmentation. Given an initial segmentation of the scene into a foreground person and a background, they fit a more detailed PS body model.

We use this idea of an initial guess followed by GrabCut but with a much more detailed model. Rather than end with a PS model, we begin with one. We use the method in [1] to fit a PS body model to the image. This 2D body model is used to initialize the pose and scale of our contour-person model. We then refine the parameters of the model (pose, view and shape) to improve the segmentation using a form of parametric GrabCut.

This parametric GrabCut idea is similar to Bray *et al.* [4], however they use a 3D articulated skeleton model and a distance transform from this to define 2D body shape. The result is a crude depiction of the body shape in 2D but the interesting element of their work is the integration of 3D pose estimation with segmentation. We also integrate parametric body shape and pose estimation with segmentation but do it in 2D with a much richer model of body shape.

3. Contour Person Model

A 2D representation of shape and pose of a 3D person presents many challenges. We seek a model that is expressive enough to represent a wide range of human bodies and poses, yet low dimensional enough to be computationally tractable for common vision problems. We build on the idea of the SCAPE model [2] and develop a *factored* representation. In particular we factor 2D body shape into: 1) a linear model characterizing shape change across the population; 2) a linear approximation to distortions caused by local camera view changes; 3) an articulation of the body parts represented by a rotation and length scaling; 4) a non-rigid deformation associated with the articulation of the parts. An example of the full model is shown in Fig. 2.

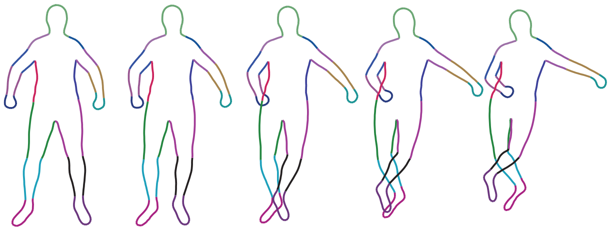

Figure 2. **The Contour Person.** Colors code the different body parts. A range of articulations is shown.

The CP model is built from training data generated from a 3D SCAPE body model [2] capturing realistic body shape variation and non-rigid pose variation. Each training body for the CP model is generated by randomly sampling a body shape, in a random pose, viewed from a random camera. The bounding contour of the 3D body is projected onto the camera plane to produce a training contour. The known segmentation of the 3D model into parts induces a similar 2D contour segmentation (Fig. 2).

3.1. Representation and synthesis

A contour C is represented discretely by N points, denoted $\{v\}_i^N$, $v_i = (x_i, y_i)$. In our experiments $N = 500$. The associated directed graph G is closed and linear, where an edge e_i, the difference vector between v_i and v_{i+1}, results from a scaled rotation, d_i, acting on l_i, the difference vector between the corresponding pair of points in a template contour T. The “deformation”, d_i, is defined by an angle θ_i and scale s_i, or equivalently, by $(s_i \cos\theta_i,\ s_i \sin\theta_i)$. Let EC be the edge representation of C, and thus,

$$EC = D(\Theta)ET \tag{1}$$

where $D(\Theta)$ is a $2N$ by $2N$ block diagonal matrix whose 2 by 2 blocks are scaled rotation matrices defined by

$$s_i R_{\theta_i} = s_i \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}. \tag{2}$$

We write D instead of $D(\Theta)$ when it is clear. E is the connectivity matrix of G (E takes values in $\{-1, 0, +1\}$ according to the usual convention of directed graphs) and C and T are vectors in $\mathbb{R}^{2N}$ of the form $(x_1, y_1, x_2, y_2, \ldots, x_N, y_N)^T$.

Note that left multiplication of ET by D can be viewed as an action of a Lie group [13]. However, applying a given D will not, in general, yield a consistent (closed) contour. Closure may be enforced by direct constraints on the deformations [13]. However, as noted in [13] this yields a linear submanifold, thus losing the group structure. Unlike

[13], we use a different approach that preserves the group structure in the generative model, and defers closure to post-processing. Specifically, given a deformation matrix D, we seek a contour C such that its deformed edges are close to the desired deformed edges in a least squares sense. In effect, we minimize the sum of squared differences between the deformed l_i's in DET and the unknown new line segments, e_i, in C:

$$\sum_{i=1}^{N} \|s_i R_{\theta_i} l_i - e_i\|^2 = \|DET - EC\|^2 . \quad (3)$$

The minimizer yields our *contour synthesis* equation:

$$C = E^{\dagger} DET \quad (4)$$

where $E^{\dagger}$, the Moore-Penrose pseudoinverse of the constant matrix E, is computed offline. The connectivity of G ensures the closure of C. This approach echos the one used in [2] for computing a consistent 3D mesh, but there the motivation was unrelated to maintaining a group structure.

Note that the minimizer is defined up to global translation. Eq. 4 shows how to synthesize C from D and the template. Conversely, given known l_i and e_i, we compute the deformations d_i by solving the invertible ($||l_i|| > 0$) linear system

$$e_i \;=\; s_i R_{\theta_i} l_i = \begin{pmatrix} l_i^{(1)} & -l_i^{(2)} \\ l_i^{(2)} & l_i^{(1)} \end{pmatrix} \begin{pmatrix} s_i \cos\theta_i \\ s_i \sin\theta_i \end{pmatrix} \quad (5)$$

where $l_i^{(k)}$ is the k^{th} element of the i^{th} line segment.

We factor deformations of the template contour into several constituent parts: pose, shape, and camera. Each of these is described in turn below and then we show how they are composed to derive the full model.

3.2. Variation in body shape

To train the shape deformation model $D_{shape}(\Theta_{shape})$ we take the 3D SCAPE model and generate numerous realistic bodies shapes in a canonical pose and project their contours into the image. Since the segmentation of the body parts is known in 3D [2], we also know the segmentation of contour points in 2D. We use this to evenly space points along a training part. The known segmentation prevents points from "sliding" between parts. The result is 2D training contours with known alignment of the contour points.

For each contour we compute its deformation from a single template contour, T, using Eq. 5. We form a matrix of all these training deformations (subtracting the mean) and perform PCA. This gives a linear approximation to contour deformations caused by body shape variation paramterized by the PCA coefficients Θ_{shape}.

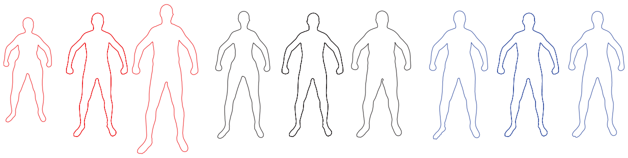

Figure 3. **Shape variation.** Gender-neutral shape model. Red: first PC. Black: second PC. Blue: third PC. In each color, from left to right: -3, 0 and +3 σ from the mean in direction of respective principal component.

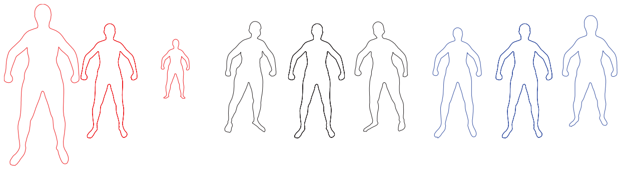

Figure 4. **Camera variation.** The first three camera principal components for the Female model. Red: first PC. Black: second PC. Blue: third PC. In each color, from left to right: -3, 0 and +3 σ from the mean in the direction of the respective PC.

Note that by maintaining the Lie group structure, we can perform PCA on the linear Lie algebra. This notion was introduced by [24] in a more complicated setting. We briefly note that in the CP case the associated computations are simpler as our Lie group is both Abelian and of finite dimension. The models learned using PCA versus those using PCA on the Lie algebra are similar but the latter requires fewer principal components to capture the same percentage of the cumulative variance.

The first three principal components (PCs) of a shape model learned from samples of both genders, can be seen in Fig. 3 (similar gender-specific models are created and used when the gender is known). The principal components clearly capture correlated properties of human shape such as variations in height, weight, girth and so on.

3.3. Variation in view

A procedure analogous to the above is used to capture contour deformation due to camera pose. Training data consists of contours generated from a single fixed body shape and posture viewed by cameras of different 3D location and tilt angle. Focal length is held fixed as it has a similar affect on the model as person-to-camera distance.

The deformations due to camera variation are well captured by PCA, with 6 components accounting for more than 90% of the variance; i.e. $\Theta_{camera} \in \mathbb{R}^6$. The first three principal components, for the female model, can be seen in Fig. 4 and roughly correspond to changes in distance between the camera and the person, rotation of the camera about the person, and foreshortening of the body caused by tilt of the camera.

Note that the view-variation is learned on the template person in a canonical pose and then is applied to other people and poses; this is an approximation.

3.4. Variation in pose

In the 3D SCAPE model, deformations due to body articulation are modeled by a two-step process. First, a rigid rotation is applied to the entire limb or body part, and then local non-rigid deformations are applied according to a learned linear model. We employ a similar approach.

For example, in Fig. 5(b), a rigid motion of the upper arm does not account for non-rigid deformations of the shoulder. This is corrected by applying a learned non-rigid deformation to the edges of the contour in the vicinity of the joint (Fig. 5(d)). Specifically, we break the deformation into $\theta_i = \theta^R + \Delta\theta_i$ and $s_i = s^R \times \Delta s_i$, where $d^R = (\theta^R, s^R)$ is the rigid deformation and $\Delta d_i = (\Delta\theta_i, \Delta s_i)$ corresponds to non-rigid deformation. d^R has the same value for all edges, e_i, in same body part (for example, the left upper arm) while Δd_i varies along the contour.

To learn the non-rigid deformation model we generate training contours using the 3D SCAPE model, in random poses, projected into a fixed camera view. Note that the 3D SCAPE model already captures the non-rigid deformations of the limbs, so that the generated 2D contour looks natural. The rigid 2D rotation, θ^R, and limb scaling, s^R, of each limb is computed between the template contour and the training contour. The scale is important as it captures foreshortening of the body parts and thus helps model out of plane movements.

We also compute the deformations, d_i, between the line segments of the template and training contours using Eq. 5. We then remove the rigid rotation, θ^R, and limb scaling, s^R, from d_i for all line segments i affected by this body part to derive a residual deformation Note that a rigid motion of the upper arm affects the non-rigid deformation of the upper arm as well as those of the lower arm and the shoulder. The residual is the deformation of the contour that is not accounted for by part-rotation and part-scaling.

Given many such Δd_i and d^R (of the same i, but from different training contours) we learn a linear predictor from the rigid transformation parameters to the non-rigid deformations. Such a model is defined by

$$\begin{pmatrix} \Delta\theta_i \\ \Delta s_i \end{pmatrix} = \begin{pmatrix} \alpha_1(i) & \cdots & \alpha_{2n(i)}(i) & \alpha_0(i) \\ \beta_1(i) & \cdots & \beta_{2n(i)}(i) & \beta_0(i) \end{pmatrix} p, \quad (6)$$

where $p = \left(\theta_1^R, s_1^R, \ldots, \theta_{n(i)}^R, s_{n(i)}^R, 1\right)^T \in \mathbb{R}^{2n(i)+1}$ is a vector of rigid transformations, $n(i)$ is the number of parts affecting d_i, and the α's and the β's are parameters to be learned. Once the model is learned, for every choice of d^R, we compute the associated Δd_i's. Then we can compute the full d_i's, and define D_{pose} is a similar way to D_{shape} and D_{camera}. The difference is that Θ_{pose} does not represent PCA coefficients. Instead, it represents the different scales and rotations of each body part.

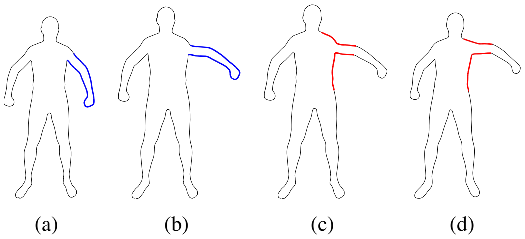

Figure 5. **Non-rigid deformation.** (a) template with left arm marked in blue. (b) rigid transformation of upper arm. (c) same as (b) but with parts which should be non-rigidly deformed due to the rigid motion marked in red. (d) final deformed contour with the non-rigidly deformed parts marked in red.

A crucial point is that the CP model utilizes 3D information when it is constructed; this is quite different from standard 2D models. This point is illustrated in the way it deals with self occlusions as well as out of the plane rotations, as depicted in Fig. 2. In a standard contour model, the ordering of the points would be poorly defined (cf. [3]). Since our contours are generated from a 3D mesh, we have known correspondence between contour points and their respective points and body parts on the 3D mesh. This provides the correct connectivity of the contour even when it crosses itself in 2D.

3.5. The full model

We train each deformation model independently above and then compose them. Because of our representation in terms of deformations this composition of factors is simply matrix multiplication. This is a key advantage over contour representations that use vertices directly. Since 2D rotation matrices commute, the composition order is immaterial. Given parameters for shape, pose, and camera view, the overall deformation is given by the *deformation synthesis* equation:

$$D(\Theta) = D_{pose} D_{shape} D_{camera}, \quad (7)$$

where $\Theta = \{\Theta_{pose}, \Theta_{shape}, \Theta_{camera}\}$. $D(\Theta)$ can be substituted into Eq. 4 to produce a new C. Here we use 24 pose parameters (12 joints $\times 2$), 10 shape coefficients and 6 camera coefficients, for a total of 40 parameters.

This factored model is an approximation, but one that works well. Example contours synthesized from the generative model are shown in Fig. 6. Note that PCA gives a Gaussian probabilistic model defined by the variance long the principal component directions. This works well for body shape and camera pose where the training samples are roughly normally distributed. The figure shows camera- and shape-variations sampled from this model. The joint angles and limb scaling are sampled uniformly over a predefined

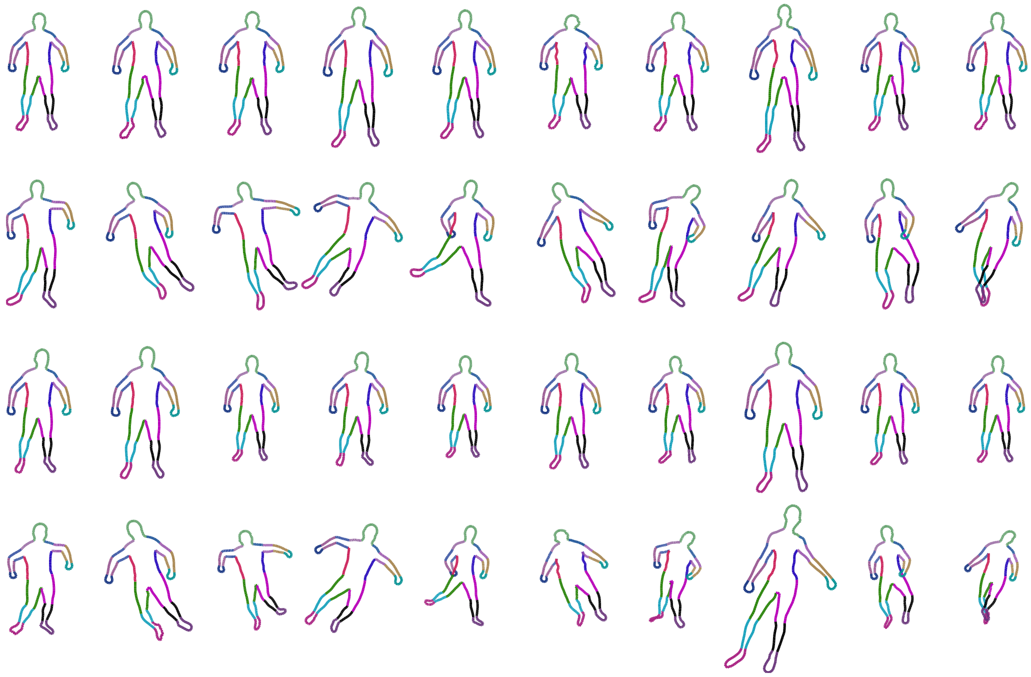

Figure 6. **Contour people sampled from the model.** Large deviations from the mean body are shown for shape, pose, and camera. Row 1: variations in body shape. Row 2: variations in pose. Row 3: variations in camera view. Row 4: all variations together.

range. Note that, because the 2D model is generated from 3D, there are correlations in 2D joint angles and scaling that could be modeled; this is future work.

4. Segmentation

As an example application of this model, we considered the problem of segmenting images of humans. The CP model provides a strong prior over human body shape that can be used to constrain more general segmentation algorithms such as GrabCut [22]. Specifically we search over the CP parameters that optimally segment the image into two regions (person and non-person) using a cost function that 1) compares image statistics inside the contour with those outside; 2) favors contours that align with image edges; 3) enforces our prior model over shape, pose and camera parameters.

Initialization. We initialize the CP model using the output of a standard pictorial structures algorithm [1]. The PS model is lower dimensional than the full CP model and hence provides a more efficient initialization. We simply set the rigid deformation parameters (rotation and scale) in the CP model to be equal to those of the PS model. While the PS model defines a segmentation of the image, it is a crude depiction of the human form. Consequently we refine the segmentation using the CP model.

Region term. The region term of the segmentation objective compares intensity and color histograms inside and outside the body contour. We take the pixel mask $m(\Theta)$ consisting of all pixels of the image plane I_c within the contour and compare the normalized histograms $H_c^{\text{in}}(I,m) = hist(I_c(m))$ and $H_c^{\text{out}}(I,m) = hist(I_c(\bar{m}))$ using the χ^2 histogram distance:

$$d_c(I,m) = 2 - \sum_i \frac{\left(H_c^{\text{in}}(i) - H_c^{\text{out}}(i)\right)^2}{H_c^{\text{in}}(i) + H_c^{\text{out}}(i)}. \tag{8}$$

We follow Martin, *et al.* [19] in treating intensity histograms and color histograms as separate features (we use the YCbCr colorspace)

$$E_{\text{St}}(I,m) = \lambda_1 d_Y(I,m) + \lambda_2 d_{Cb}(I,m) + \lambda_3 d_{Cr}(I,m).$$

Edge term. The segmented contour should also follow image edges. We detect image edges using a standard edge detector and apply a thresholded distance transform to define an edge cost map normalized to $[0,1]$. We use the trapezoid rule to evaluate the line integral of the set of all model edges over the edge cost image. This defines an edge cost, $E_{\text{Eg}}(I,\Theta)$, that is included in the objective function.

Prior. We use a loose prior, $E_{\text{Pr}}(\Theta)$, on shape, pose and camera only to prevent values significantly outside what the model is trained on. This prior remains zero until the parameters are three standard deviations from the mean and then rises linearly from there.

Objective. The full cost function is then $E(I,\Theta) = E_{\text{St}}(I,m) + \lambda_4 E_{\text{Eg}}(I,\Theta) + \lambda_5 E_{\text{Pr}}(\Theta)$, which we optimize using a gradient-free direct search simplex method.

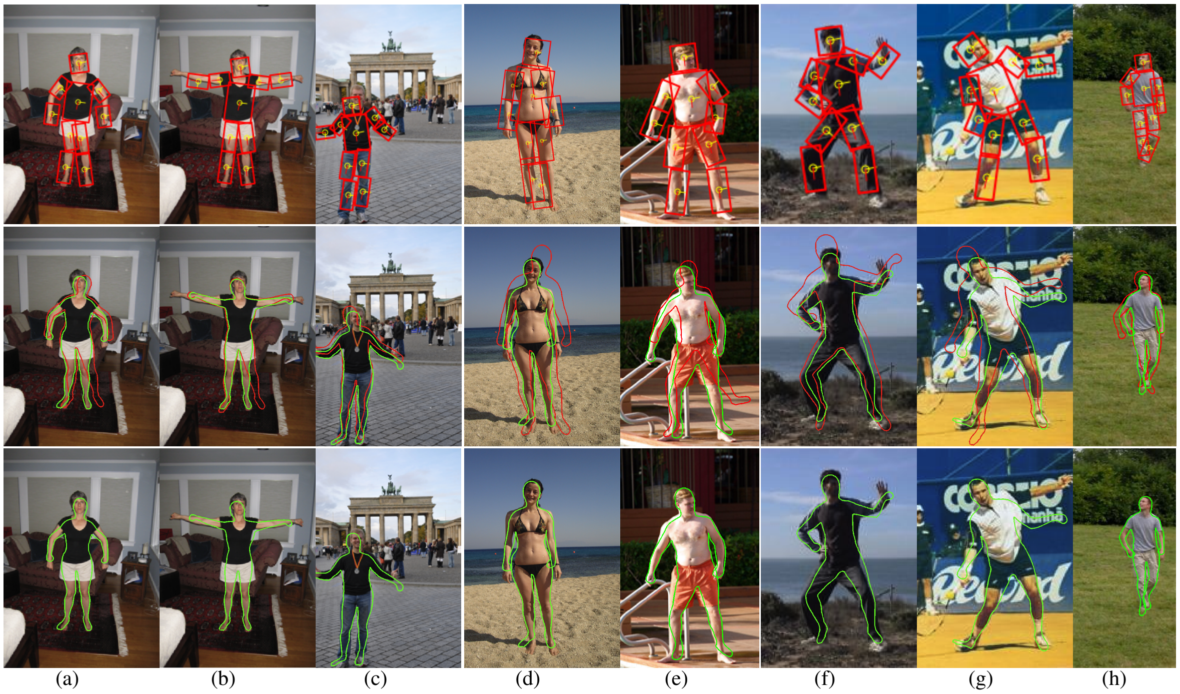

Figure 7. **Results.** Row 1: Pictorial Structures result. Row 2: CP initialization from PS (red) and CP result (green). Row 3: CP result.

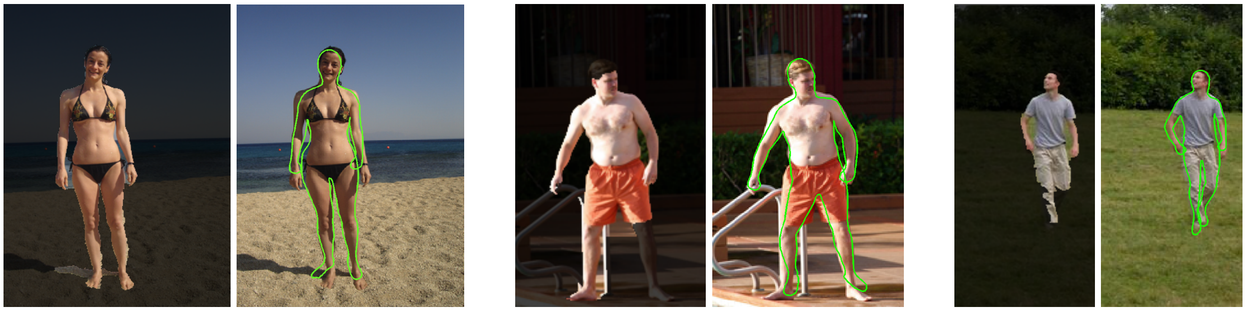

Figure 8. **Comparison to GrabCut.** GrabCut with a manual initialization step and no manual cleanup, compared to fully automatic CP segmentation.

5. Experimental Results

The CP model realistically captures a large range of real human poses in the space in which it was trained (Fig. 7). This enables it to find segmentations which, while not perfect, are guaranteed to be plausibly human, unlike more general segmentation methods (Fig. 8). This is a fairly simplistic segmentation approach which is designed only to illustrate the CP model; note that the parametric segmentation method here is similar in spirit to PoseCut [4].

Note that the model is not clothed and consequently will produce segmentations that tend to ignore clothing. While the optimization could be made explicitly robust to clothing [5], for segmentating clothed people it might be preferable to explicitly model clothing.

Given our simplistic segmentation method, the model can also make mistakes such as those in Figs. 7(a) and 7(b), where the optimization latches on to a strong edge at the hairline and finds that the hair matches the background color statistics better than the foreground statistics; this pushes the shoulders down, causes the head to be smaller than the torso and legs would otherwise indicate, so the camera gets detected as tilted upwards, which in turn causes the shoulders to narrow and the arms to shorten. In Fig. 7(g) the PS initialization is far enough off that a simple direct search optimization method cannot escape the local minimum; note though that only the left arm was poorly initialized and that only the left arm remains poorly localized and segmented. Another failure case is typified by the left hand in Fig. 7(f). We train this model without varying the angle of the wrist and our training data consists of exclusively closed fists. Consequently the model does a poor job representing open hands and bent wrists.

6. Conclusions

We have defined a new type of 2D human body model that retains the standard part-based structure of classical pictorial structures models. It goes beyond previous models in several significant ways. First, it *factors* 2D body shape into several causes. Deformations from a training template are used to describe changes in shape due to camera view, body shape, and articulated pose. The approach is similar to the 3D SCAPE model in that deformations are combined into a complete generative model. Second, the CP model captures the non-rigid deformations of the body that result from articulation. Like SCAPE, these are learned from training examples. The result is a fairly low-dimensional model that represents realistic human body contours and can be used for vision applications such as person detection and tracking.

Our 2D model is view-based and here we have have only shown examples for frontal bodies. A key next step is to extend this to other discrete views. A general solution to the human pose and shape estimation problem will require an inference method to search over the discrete set of views.

The part-based structure of the CP model makes the use of a PS-style inference method like BP appealing. There are several challenges, however: 1) the non-rigid articulated deformations mean that the body does not fully factor into independent parts; 2) the camera view and body shape are "global" properties; 3) the contour is constructed from deformations via least squares optimization. These properties mean that the simple tree-structure of the PS graphical model is lost, complicating inference.

Future work will explore the estimation of 3D body shape parameters directly from the 2D body shape parameters. We will also explore the inference of gender from the 2D shape. Finally, the 2D contour person model should be extended to model loose-fitting clothing.

Acknowledgements. This work was supported in part by NIH EUREKA 1R01NS066311–01 and NSF IIS–0812364. We thank D. Hirshberg for assistance with generating model contours. We thank the authors of [1] for their PS implementation and A. Balan for his SCAPE implementation.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. *CVPR*, pp. 1014–1021, June 2009. 1, 2, 3, 6, 8

[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. 2, 3, 4

[3] A. Baumberg and D. Hogg. Learning flexible models from image sequences. *ECCV*, vol. 1, pp. 299–308, 1994. 2, 5

[4] M. Bray, P. Kohli, and P. H. S. Torr. PoseCut: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. *ECCV*, pp. 642–655, 2006. 3, 7

[5] A. O. Bălan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. *CVPR*, 2007. 1, 7

[6] T.-J. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. *CVPR*, pp. 239–245, 1999. 1, 2

[7] T. Cootes, D. Cooper, C. Taylor, and J. Graham. Active shape models - Their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, Jan 1995. 2

[8] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 1, 2

[9] V. Ferrari, M. Marin-Jiminez, , and A. Zisserman. Progressive search space reduction for human pose estimation. *CVPR*, 2008. 2, 3

[10] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. on Computer*, 22(1):67–92, Jan 1973. 2

[11] D. Gavrila. A Bayesian, exemplar-based approach to hierarchical shape matching. *PAMI*, 29:1408–1421, 2007. 3

[12] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3D structure with a statistical image-based shape model. *ICCV*, pp. 641–647, 2003. 3

[13] U. Grenander and M. I. Miller. *Pattern theory: From representation to inference*. Oxford Univ. Press, 2007. 3, 4

[14] G. E. Hinton. Using relaxation to find a puppet. *Proc. of the A.I.S.B. Summer Conference*, pp. 148–157, 1976. 1, 2

[15] S. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. *Int. Conf. Automatic Face and Gesture Recog.*, pp. 38–44, 1996. 1, 2

[16] C. Kervrann and F. Heitz. A hierarchical Markov modeling approach for the segmentation and tracking of deformable shapes. *Graphical Models and Image Processing*, 60(3):173 – 195, 1998. 3

[17] X. Lan and D. Huttenlocher. Beyond trees: Common factor models for 2D human pose recovery. *ICCV*, pp. 470–477, 2005. 1, 2

[18] Z. Lin and L. Davis. Shape-based human detection and segmentation via hierarchical part-template matching. *PAMI*, 32(4):604–618, 2010. 3

[19] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 26(5):530–549, 2004. 6

[20] E.-J. Ong and S. Gong. A dynamic human model using hybrid 2D-3D representations in hierarchical PCA space. *BMVC*, pp. 33–42, 1999. 2

[21] D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. *CVPR*, vol. 2, pp. 467–474, 2003. 1, 2

[22] C. Rother, V. Kolmogorov, and A. Blake. GrabCut: Interactive foreground extraction using iterated graph cuts. *ACM ToG*, 23:309–314, 2004. 2, 3, 6

[23] L. Sigal and M. J. Black. Predicting 3D people from 2D pictures. *Articulated Motion and Deformable Objects (AMDO)*, LNCS 4069, pp. 185–195, 2006. 1, 2

[24] M. Vaillant, M. Miller, L. Younes, and A. Trouvé. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23:161–169, 2004. 4