# Representation and Matching of Articulated Shapes

Jiayong Zhang, Robert Collins and Yanxi Liu

The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213
{zhangjy,rcollins,yanxi}@cs.cmu.edu

## Abstract

*We consider the problem of localizing the articulated and deformable shape of a walking person in a single view. We represent the non-rigid 2D body contour by a Bayesian graphical model whose nodes correspond to point positions along the contour. The deformability of the model is constrained by learned priors corresponding to two basic mechanisms: local non-rigid deformation, and rotation motion of the joints. Four types of image cues are combined to relate the model configuration to the observed image, including edge gradient map, foreground/background mask, skin color mask, and appearance consistency constraints. The constructed Bayes network is sparse and chain-like, enabling efficient spatial inference through Sequential Monte Carlo sampling methods. We evaluate the performance of the model on images taken in cluttered, outdoor scenes. The utility of each image cue is also empirically explored.*

## 1. Introduction

We consider a model-based approach to simultaneously finding the body boundary shape and locating body parts of a walking human target in a single view (Fig. 1). Such a segmentation can provide discriminative cues for human identification from gait, or can be used to initialize a kinematic body tracker for activity analysis. However, even in a fixed viewpoint scenario, accurate human body extraction is a non-trivial task, due to large variation in observed body shapes caused by articulated motion, anthropometric body variation, and clothing.

Most work on articulated human body fitting focuses on tracking 3D kinematic body models through video sequences. These approaches are often brittle because the likelihood surface relating a high degree of freedom 3D articulated body model to 2D body shape in an image is fraught with local minima [16]. Given the complexity of the likelihood, Monte Carlo sampling techniques for representing the posterior distribution demonstrate the most promising results [3, 8, 15, 17]. Even then, robust fitting is typically achieved only by imposing additional information, such as the use of multiple simultaneous views [3, 8], or strong constraints on the temporal dynamics [15].

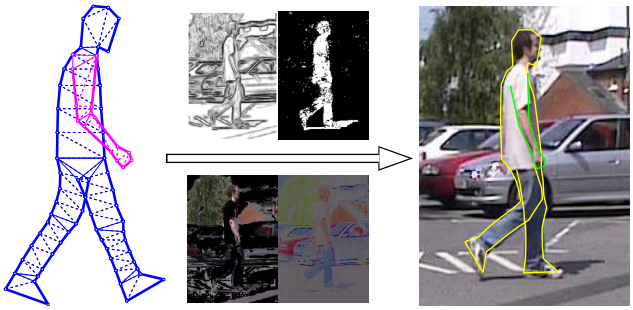


Figure 1: Overview of our approach. An articulated non-rigid 2D body contour model (left) and local image cues (middle) are combined via Bayesian graphical modeling. The model is fit using sequential Monte Carlo to a sample image (right) taken in a cluttered, outdoor scene.

One alternative is to fit a 2D articulated body model to the image instead, in the hope that the likelihood surface will be better behaved [1, 2, 6]. Since each link in the 2D model typically describes the projected image appearance of a corresponding rigid link in the 3D kinematic model, these approaches are, by necessity, viewpoint-specific. Nonetheless, the degrees of freedom left in the projected model are still high enough that gradient descent tracking [1, 6] needs a good initial pose estimate and small inter-frame motion. Methods that recognize that the solution space is multi-modal [2], particularly in the presence of background clutter, again seem the most promising.

Another alternative is to detect and assemble individual body parts in a bottom-up manner [11, 13, 18]. These approaches employ "weak" models, where each body part is represented by a single rectangle or feature point, and target high-level tasks such as human detection.

Conventionally, body parts are approximated by regular shapes such as lines, 2D ribbons or blobs, or 3D cylinders. On the other hand, there is a rich body of research on matching arbitrary deformable shapes. Recently a new polygon representation was proposed using the constrained Delaunay triangulation [5]. It has an attractive property that the globally optimal match of a model to the image can be found via Dynamic Programming (DP), since the dual graph of a triangulated polygon is a tree. However, this method is

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)
1063-6919/04 $20.00 © 2004 **IEEE**

IEEE COMPUTER SOCIETY

restricted to simple polygons and a predefined deformation energy on a collection of cliques of size 3. More importantly, DP can only output a single estimate of the match, which as we have alluded to is typically multi-modal.

In this paper, we propose a new body shape representation based on Bayesian graphical modeling. This representation has several features: 1) it models both the local non-rigid deformation and the global articulated motion; 2) the matching does not use any dynamic constraints on body pose over time; and 3) the model has a chain-like structure allowing spatial inference to be performed efficiently via Sequential Monte Carlo (SMC). It is important to note that we are using SMC to perform inference over a spatial chain for shape fitting, rather than over a temporal chain of poses, as is typical in the body tracking work described above. A similar idea has been used in [9, 12].

# 2. Deformable Model

We adopt a Bayesian approach to deformable template matching, as conveyed by the formula

$$p(\Omega|\mathcal{I}) \sim p(\Omega)\, p(\mathcal{I}|\Omega), \tag{1}$$

where $\Omega$ denotes the parameter set of the model, and $\mathcal{I}$ denotes the input image. The shape prior $p(\Omega)$ encodes our knowledge of possible shape deformations, while the imaging likelihood $p(\mathcal{I}|\Omega)$ measures how compatible a given model configuration is with respect to observed image features. The desired model-to-image matching can then be found by exploring the posterior distribution $p(\Omega|\mathcal{I})$.

## 2.1. Shape Prior

We represent the body shape by a set of non-rigid 2D contours, as depicted in Fig. 1. These contours are assumed to be piecewise linear, and thus are completely described by a set of $T$ landmarks $\mathbf{v}_{1:T} = \{\mathbf{v}_t\}_{t=1}^T$. The 2D coordinates of these landmarks, $\{(x_t, y_t)\}_{t=1}^T$, constitute the parameter set $\Omega \in \mathcal{R}^{2T}$ of our body model. To encode the shape prior knowledge, we need to model the joint density distribution of $\Omega$, *i.e.* $p(x_1, y_1, \ldots, x_T, y_T)$, or equivalently $p(\mathbf{v}_{1:T})$.

A common practice in statistical shape analysis is to first remove the rigid part of the deformation (translation, rotation and scaling) and then model the shape residual using some low dimensional linear model. However, direct application of such a global analysis to the shape of a human body is difficult, because the articulated motions of human body parts are so large and independent that the shape residuals no longer reside in a low dimensional linear subspace.

In this paper, we apply graphical modeling to the shape representation, *i.e.* factoring the joint distribution of all landmarks into a series of marginal and conditional distributions. More precisely, the shape is represented by a Bayes net with $T$ nodes corresponding to the $T$ landmark points. Each node can take any continuous vector value $(x, y)$. When the links between nodes are sparse, $p(\mathbf{v}_{1:T})$ can be factored into products of many local terms, each of which only depends on a few neighboring nodes. To this end, the contour model is triangulated as in Fig. 2. The landmark positions and triangulations are designed such that: 1) the landmarks can be ordered in a fixed way; 2) the shape can be constructed sequentially by growing one landmark at a time; and 3) each landmark $\mathbf{v}_t$ is connected to an unique parent edge $\mathbf{e}_t^P = \{\mathbf{e}_t^P(0:1)\} \subset \mathbf{v}_{1:t-1}$, and a parent triangle $\mathbf{g}_t^P$. The form of this representation is essential to the efficient sampling algorithm described below.

Given the fixed landmark ordering, the joint landmark distribution can be expanded as

$$p(\mathbf{v}_{1:T}) = \prod_t p(\mathbf{v}_t|\mathbf{v}_{1:t-1}). \tag{2}$$

To further specify the complete conditional $p(\mathbf{v}_t|\mathbf{v}_{1:t-1})$, we introduce two types of deformation mechanisms.

The first type is designed to model rotation motion of the joints. We select nine joint triangles, with the index set denoted as $\mathcal{J}$, corresponding to neck, shoulder, elbow, hips, knees and ankles. They divide the body shape into ten parts. For each $t \in \mathcal{J}$, $\mathbf{v}_t$ is predicted by perturbing $\mathbf{e}_t^P(1)$ with $(\rho_t, \theta_t)$ in the local polar coordinates determined by $\vec{\mathbf{e}}_t^P$.

$$\mathbf{v}_t = \rho_t \cdot R(\theta_t) \cdot (\mathbf{e}_t^P(1) - \mathbf{e}_t^P(0)) + \mathbf{e}_t^P(0)\,. \tag{3}$$

Although it seems safe to assume that the local lengths $\rho_t$ are independent, we cannot ignore the long range dependencies among joint angles $\Theta = \{\theta_t, t \in \mathcal{J}\}$. Therefore another Bayes network is manually designed to model $p(\Theta)$. Fig. 3 shows its topology.

The second type of deformation mechanism is designed to model local non-rigid deformation. For those triangles within the body parts, we assume the Markov property

$$p(\mathbf{v}_t|\mathbf{v}_{1:t-1}) = p(\mathbf{v}_t|\mathbf{g}_t^P), \tag{4}$$

which implies that the position of the $t$-th landmark $\mathbf{v}_t$ can be predicted from it's parent triangle $\mathbf{g}_t^P$. Our prediction method uses an affine transformation in the local landmark coordinate system:

$$\mathbf{v}_t = (A_t \cdot \bar{\mathbf{v}}_t + b_t) \circ \mathbf{n}_t\,, \tag{5}$$

where $\bar{\mathbf{v}}_t$ is the reference position of the $t$-th landmark. Note that the conditioning variables $\mathbf{g}_t^P$ are implicitly encoded in $A_t$ and $b_t$. To predict the position of $\mathbf{v}_t$, the reference landmark $\bar{\mathbf{v}}_t$ is carried through a linear transformation $A_t$ followed by a shift $b_t$, and then perturbed by noise $\mathbf{n}_t$. $(A_t, b_t)$ is determined by either 1) the affine transformation from the triangle $\bar{\mathbf{g}}_t^P$ in the reference model to the triangle $\mathbf{g}_t^P$ fit previously to the data, or 2) the similarity transform from the reference edge $\bar{\mathbf{e}}_t^P$ to the fitted edge $\mathbf{e}_t^P$. The latter is used for the first triangle of each body part, whose parent is a joint triangle. The noise term $\mathbf{n}_t = (n_t^x, n_t^y)$ is applied in the local Cartesian coordinates determined by $\vec{\mathbf{e}}_t^P$.

Using the deformation mechanisms described above, a complete sample shape can be sequentially constructed starting from a given position, scale and orientation of the


IEEE COMPUTER SOCIETY

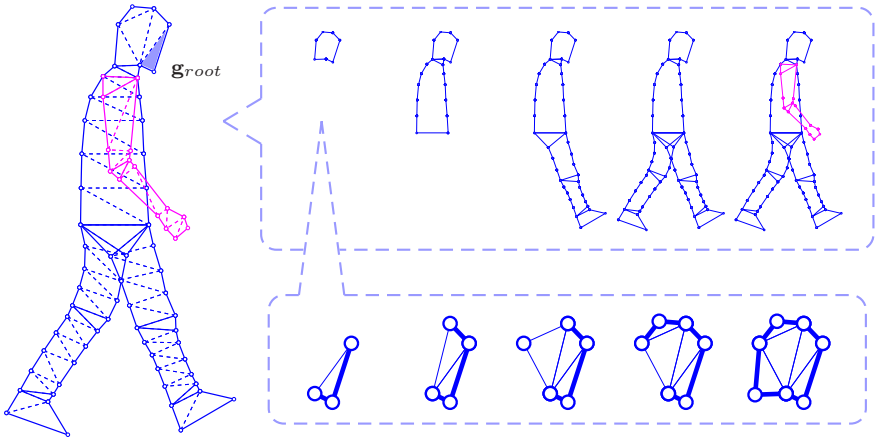



Figure 2: Shape triangulation specifying the elimination order of vertices. Given root triangle $\mathbf{g}_{root}$, the shape is constructed sequentially by growing one triangle (vertex) at a time. Note that this is not the connectivity graph of the shape prior.

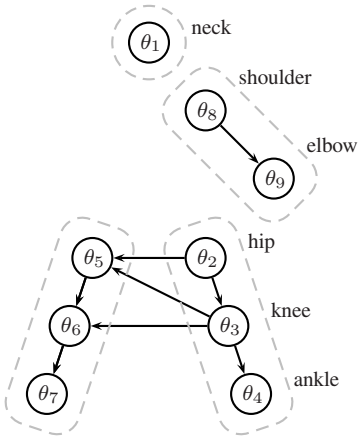



Figure 3: The graph structure specifying the dependencies among nine joint angles $\Theta$.

root triangle $\mathbf{g}_{root} = \mathbf{v}_{0:2}$, which is defined on the face in our shape model (Fig. 2). At each step, a vertex sample $\tilde{\mathbf{v}}_t$ is generated according to either (3) or (5), depending on whether the current triangle is a joint or body triangle.

To summarize, the shape prior can be formulated as

$$p(\mathbf{v}_{1:T}) = p(\Theta) \prod_{t \notin \mathcal{J}} p(\mathbf{n}_t) \prod_{t \in \mathcal{J}} p(\rho_t), \qquad (6)$$

Note that the proposed model is translation invariant because $p(\mathbf{v}_{1:T})$ involves no absolute landmark positions. By expressing $\mathbf{n}_t$ in the local coordinate system of $\vec{\mathbf{e}}_t^P$, the model is also made rotation and scale invariant.

We estimate the densities $p(\mathbf{n}_t)$, $p(\rho_t)$ and $p(\Theta)$ from a set of training images. The details are described in Section 4.2. Fig. 4 shows several samples randomly drawn from the learned shape prior.

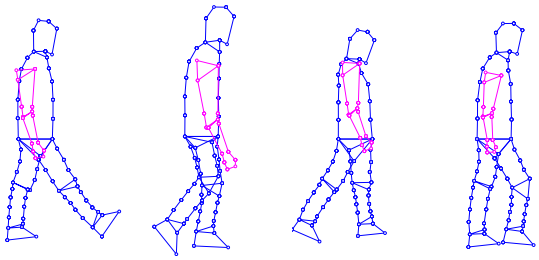


Figure 4: Stochastic samples from the learned shape prior $p(\Omega)$. Each shape is normalized by aligning the bottom edge of the torso with line segment $(0,0)(1,0)$.

## 2.2. Imaging Likelihood

Similar to $p(\Omega)$, we factor the imaging likelihood $p(\mathcal{I}|\Omega)$ into a series of potential functions. First, the shape model is covered by a set of clusters $\mathcal{C}$. Each cluster contains a small number of related nodes, from which local features can be computed. These local features are assumed to be independent, and the product of their probabilities defines the potential function $\phi(\mathbf{v}_C)$ associated with that cluster. Finally, the likelihood model is expressed as

$$p(\mathcal{I}|\Omega) = \prod_{C \in \mathcal{C}} \phi(\mathbf{v}_C). \qquad (7)$$

Four different types of image cues are involved in computing the local features that induce cluster potentials.

**Edge Gradient Map**

The edge potential $\phi_e(\mathbf{e})$ is defined on clusters of size two, namely the vertices of a boundary line segment. We use a color edge detector called the compass operator [14]. At each pixel, this operator outputs a vector $\mathbf{u}$ ($\|\mathbf{u}\| \in [0,1]$) which encodes the strength and orientation of the edge feature at that point. Fig. 5b shows an example strength image. Given a line segment $\mathbf{e}$, we compute the boundary energy

$$E(\mathbf{e}) = \int_s \mathbf{u}(s) \cdot \mathbf{e} / \|\mathbf{e}\|^2 \, ds, \qquad (8)$$

and then model $\phi_e(\mathbf{e})$ with a truncated Gaussian

$$\phi_e(\mathbf{e}) \propto \exp\{-[1 - E(\mathbf{e})]^2 / \sigma_e^2\}, \quad E(\mathbf{e}) \in [0,1]. \qquad (9)$$

**Foreground/Background Mask**

The foreground potential $\phi_f(\mathbf{g})$ is defined on clusters of size three. It is computed from a binary foreground mask that labels pixels as 1 if they are likely to be on the person, and 0 if they are more likely to come from the background. This mask could be computed from a prior model of the color distribution of the person's clothing, via histogram backprojection [19]. However, in our experiments, we use a static camera and compute the mask using background subtraction. A standard background model of the mean and covariance of each pixel is used, and a binary mask $\mathcal{B}$ is generated by thresholding the Mahalanobis distance. Further, we assume that each pixel in $\mathcal{B}$ is drawn independently from the Bernoulli distribution $\{p_{10}, p_{11}\}$ if the pixel is in the foreground, or $\{p_{00}, p_{01}\}$ if it is in the background ($p_{\cdot 0}$+$p_{\cdot 1}$=1, $p_{00}$>$p_{01}$, $p_{10}$<$p_{11}$). Given a model configuration $\Omega$, the probability of observing foreground mask $\mathcal{B}$ is derived as

$$p(\mathcal{B}|\Omega) = \gamma \left(p_{10}/p_{00}\right)^{N_{10}} \left(p_{11}/p_{01}\right)^{N_{11}}, \qquad (10)$$

(a) Raw Image (b) Edge Mag. (c) Skin Segment. (d) FG/BG Mask

Figure 5: A sample image (a) and three low-level visual cues (b)–(d) combined in the imaging model.

where $N_{10}$ is the number of pixels inside the model that are labeled background, $N_{11}$ is the number of pixels inside the model that are labeled foreground, and $\gamma$ is a constant independent of $\Omega$. Noting that $N_{1\cdot}$ can be decomposed as $N_{1\cdot} = \sum_t N_{1\cdot}(\mathbf{g}_t)$, we have

$$\phi_f(\mathbf{g}) \propto \exp\{\alpha_f N_{10}(\mathbf{g}) + \beta_f N_{11}(\mathbf{g})\}, \qquad (11)$$

where $\alpha_f$ and $\beta_f$ are coefficients depending on $p_{10}$ and $p_{00}$.

**Skin Color**

The skin potential $\phi_s(\mathbf{g})$ helps to locate the head and arm. We use a simple color-based skin detector based on Gaussian mixture models. The detector is learned from a training set of hand-labeled skin pixels. Because the face area is often very small in gait images shot from a side view, we extend the training set with hair pixels such that the resulting detector can detect both skin and hair. Note that the skin color mask can be very noisy and contain large false positive areas (Fig. 5c). However, this is not a problem when complemented by other image cues. As the detector outputs a binary mask, we use a potential function similar to $\phi_f$:

$$\phi_s(\mathbf{g}) \propto \exp\{\alpha_s N_{10}(\mathbf{g}) + \beta_s N_{11}(\mathbf{g})\}. \qquad (12)$$

**Appearance Consistency**

The appearance consistency potential $\phi_c(\mathbf{g}_i, \mathbf{g}_j)$ is designed to reflect the observations that: 1) appearances of adjacent triangles are likely to be similar; 2) appearances of symmetrically corresponding leg triangles are likely to be similar; and 3) appearance of foot and leg triangles are likely to be different. Given two triangles $\mathbf{g}_i$ and $\mathbf{g}_j$, we first compute the normalized color histograms $h_i$ and $h_j$. Their distance is then defined using Bhattacharya coefficient $d_{ij} = \sqrt{1-\rho_{ij}}$, where $\rho_{ij} = \sum_k \sqrt{h_i(k)h_j(k)}$. Finally we model $d_{ij}$ with a truncated Gaussian

$$\phi_c(\mathbf{g}_i, \mathbf{g}_j) \propto \exp\{-d_{ij}^2/\sigma_c^2\}, \quad d_{ij} \in [0,1]. \qquad (13)$$

# 3. Inference by Sequential Monte Carlo

Combining the equations for shape prior (6) and imaging likelihood (7) with the Bayes equation (1), the posterior distribution can be written as

$$p(\mathbf{v}_{1:T}|\mathcal{I}) \propto p(\Theta) \prod_{t \notin \mathcal{J}} p(\mathbf{n}_t) \prod_{t \in \mathcal{J}} p(\rho_t) \prod_{C \in \mathcal{C}} \phi(\mathbf{v}_C) \qquad (14)$$

Several tools are available to explore this posterior distribution, such as Dynamic Programming, Belief Propagation, and Markov Chain Monte Carlo. In this paper, we adopt the method of Sequential Monte Carlo (SMC), also known as particle filters [4], which has the special property of drawing simultaneously a population of independent samples from the given distribution. This is possible because the proposed shape model possesses a simple chain-like structure. It is important to distinguish our use of particle filters for body model fitting in a single view, from the usual one in tracking body pose across time. Here, our chain is spatial, representing the sequential decomposition of contour landmark points, instead of a temporal chain of poses across time.

To apply the recursive SMC smoother, we need to pick a proposal function $\pi(\mathbf{v}_{1:t})$ with corresponding importance weights $w(\mathbf{v}_{1:t})$. As described in section 2.1, our shape model can be constructed sequentially by adding one vertex at a time, thus making it a natural choice for the proposal function. Let $\Theta_t$ be the subset of joint angles that are visited as of time $t$, *i.e.* $\Theta_t = \{\theta_k | k \le t, k \in \mathcal{J}\}$. Let $\mathcal{C}_t$ be those clusters that are completely covered only at time $t$, *i.e.* $\mathcal{C}_t = \{C | t \in C, C \subseteq [1:t], C \in \mathcal{C}\}$. The proposal function $\pi_t$ is the partial shape prior on $\mathbf{v}_{1:t}$, which has an iterative form

$$\begin{aligned} \pi_t &= \pi_{t-1} \cdot p(\mathbf{v}_t|\mathbf{v}_{1:t-1}) \\ &= \pi_{t-1} \cdot \begin{cases} p(\mathbf{v}_t|\mathbf{e}_t^P, \Theta_{t-1}) & \text{if } t \in \mathcal{J}, \\ p(\mathbf{v}_t|\mathbf{g}_t^P) & \text{otherwise} \end{cases} \end{aligned} \qquad (15)$$

with the (unnormalized) importance weights

$$w_t \propto w_{t-1} \cdot \prod_{C \in \mathcal{C}_t} \phi(\mathbf{v}_C) \qquad (16)$$

Another key element in SMC is resampling in order to deal with a high number of dimensions. We use stratified resampling proposed in [7], which is optimal in terms of variance in the class of unbiased resampling schemes.

The inference procedure is summarized as follows.

SMC INFERENCE PROCEDURE

1. INITIALIZATION.
   - For $i$ = 1 to $N$, sample $\mathbf{v}_{0:2}^{(i)} \sim p_0(\mathbf{v}_{0:2}|\mathcal{I})$ and set $t = 3$.
2. IMPORTANCE SAMPLING.
   - For $i$ = 1 to $N$, if $t \notin \mathcal{J}$, sample $\tilde{\mathbf{v}}_t^{(i)} \sim p(\mathbf{v}_t|\mathbf{g}_t^{P(i)})$, otherwise sample $\tilde{\mathbf{v}}_t^{(i)} \sim p(\mathbf{v}_t|\mathbf{e}_t^{P(i)}, \Theta_{t-1}^{(i)})$. Set $\tilde{\mathbf{v}}_{0:t}^{(i)} = \left(\mathbf{v}_{0:t-1}^{(i)}, \tilde{\mathbf{v}}_t^{(i)}\right)$.
   - For $i$ = 1 to $N$, evaluate the importance weights
     $$\tilde{w}_t^{(i)} = \prod_{C \in \mathcal{C}_t} \phi(\tilde{\mathbf{v}}_C^{(i)})$$
     Normalize the importance weights.
3. STRATIFIED RESAMPLING.
   - Resample $N$ particles $\left\{\mathbf{v}_{0:t}^{(i)}\right\}_{i=1}^N$ from the set $\left\{\tilde{\mathbf{v}}_{0:t}^{(i)}\right\}_{i=1}^N$ according to the importance weights.
   - Set $t \leftarrow t+1$ and go to step 2.

The procedure is initialized by uniformly sampling the root triangle over a range of position, rotation and scale.

# 4. Experiments

## 4.1. Image Dataset

We evaluate our deformable model using the Southampton HumanID gait database[1], which was originally collected for research in automatic gait recognition. The database contains video sequences of walking individuals. Only sequences filmed from the side view are used in our experiments. Our training data consists of 112 sequences of 28 subjects filmed inside the lab, under controlled lighting with a green chroma-key backdrop (Fig. 6a). Our testing data consists of 10 sequences of 10 subjects shot outdoors with cluttered background and natural lighting (Fig. 6b).

Although the raw data are video sequences, it is important to note that we did not use the sequential (video) nature of the data to impose dynamic constraints on the body pose over time. For the purpose of body contour fitting, each frame is treated independently.

(a) Indoor (training) (b) Outdoor (testing)

Figure 6: Example sequences from Southhampton gait database. Each displayed image is merged from five selected frames (including the starting and ending frames).

## 4.2. Learning Shape Model Parameters

The body shape model was created by the following bootstrapping procedure. First, we built the triangulated body contour and identified its rotation joints by hand-labeling one frame of the indoor data. We then fit this model to all 3,126 indoor training frames using a uniform shape prior, and good fitting was obtained since the indoor green-screen images are very clean (Fig. 7). The fits obtained were then used to learn a more informative empirical prior distribution on body shape parameters. We represent densities $\{p(\mathbf{n}_t)$, $p(\rho_t)$, $p(\Theta)\}$ in the shape prior by discrete probability tables. For each fit in the training set, a set of deformation parameters $\{\mathbf{n}_t, \rho_t, \Theta\}$ was calculated based on the posterior mean estimate, then discretized and pooled to compute the probability tables. Note that each table's dimension is at most three. The final model, including the learned shape prior, was then used for testing in cluttered scenes.

We also trained the skin/hair color model using the indoor images. This leads to a weak classifier, since the lighting conditions of the indoor training images are very different from the outdoor natural illumination of the test scenes. Other parameters of the imaging model were also determined experimentally.

[1] Available online at `http://www.gait.ecs.soton.ac.uk`

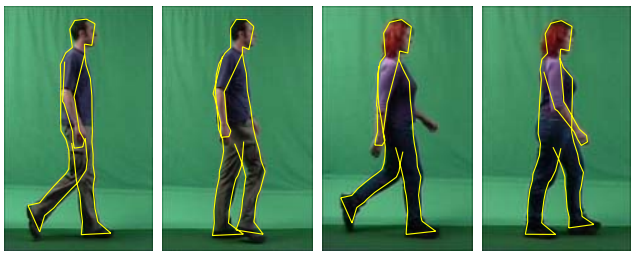


Figure 7: Sample results on fitting the indoor training set, using a uniform shape prior. Plotted are the posterior means.

## 4.3. Test Result and Evaluation

We applied the learned shape model to 963 images taken from the cluttered, outdoor gait sequences. Fig. 8 shows two examples illustrating the incremental SMC inference procedure. For each time $t$, we plot the mean shape up to $\mathbf{v}_t$, with the marginal distribution of $\mathbf{v}_j$ $(j \leq t)$ summarized by its covariance ellipse (*i.e.* error ellipse). In the first image, the two legs are close to each other and a large uncertainty is observed when fitting the front leg. This uncertainty diminishes after both legs are fitted. The second image has a background color similar to that of human skin, thus the head is not reliably detected until the body information has been incorporated.

A post processing procedure was used to deal with cases where both arms are visible. First the sampled arm shapes are divided into two clusters based on hand positions, and the mean shape of each cluster is computed. Then we compare the hand distance between these two mean shapes to the width of the torso. If the ratio is above a threshold of 0.6, then both arms are assumed to be detected.

To quantitatively evaluate the proposed model, we randomly selected 50 images and hand-labeled the ground truth boundaries of body parts. The posterior distribution computed by the SMC algorithm for each image is then summarized by a mean contour, which is compared to the ground truth using two types of metrics. One is symmetric Chamfer distance reflecting the global average error, and the other is symmetric Hausdorff distance reflecting the local worst-case error. Given two point sets $\mathcal{U}$ and $\mathcal{V}$, the Chamfer distance $d_{cham}(\mathcal{U}, \mathcal{V})$ is defined as the mean of the distances between each point in $\mathcal{U}$ and its closest point in $\mathcal{V}$. The symmetric distance is obtained by averaging $d_{cham}(\mathcal{U}, \mathcal{V})$ and $d_{cham}(\mathcal{V}, \mathcal{U})$. The Hausdorff distance is defined similarly except that we replace the mean with the maximum. We evaluate the fitting errors of body and arm separately, since it was expected that the core body shape (head, torso and legs) would be fit more accurately than the arms. Evaluation

Table 1: Evaluation of model fitting by symmetric Chamfer and Hausdorff distances between mean contours and hand-labeled ground truth. The mean and standard deviation (in pixels) over 50 images are given in the form of MEAN ±STD. Each row corresponds to one combination of image cues. If selected, the source is marked with '•'.

| $\phi_e\phi_f\phi_s\phi_c$ | Chamfer | | Hausdorff | |
|---|---|---|---|---|
| | Body | Arm | Body | Arm |
| • ○ ○ ○ | 4.00 ±3.52 | 4.41 ±4.08 | 16.7 ±13.7 | 10.7 ±7.20 |
| ○ • ○ ○ | 2.53 ±0.91 | 6.25 ±5.95 | 10.2 ±3.52 | 13.0 ±9.20 |
| ○ • • • | 2.19 ±0.61 | 2.36 ±0.94 | 8.80 ±2.59 | 7.13 ±2.77 |
| • ○ • • | 2.77 ±1.62 | 4.13 ±6.83 | 11.8 ±6.14 | 9.49 ±8.41 |
| • • ○ • | 2.00 ±0.59 | 2.96 ±1.51 | 9.17 ±2.49 | 8.30 ±3.88 |
| • • • ○ | 2.02 ±0.53 | 2.25 ±1.30 | 8.81 ±2.19 | 6.77 ±3.52 |
| • • • • | 1.87 ±0.42 | 2.18 ±0.99 | 8.35 ±2.07 | 6.62 ±2.88 |

results are summarized in the last row of Tab. 1. To interpret these scores, note that average body height is roughly 200 pixels in the dataset. Some example images at all levels of performance are given in Fig. 10, with the last row showing some typical fitting errors. More results are available online at *http://www.cs.cmu.edu/~zhangjy/cvpr04/*.

We also evaluated the utility of each image cue by comparing fitting accuracy both with and without that source of data. The results are shown in the first six rows of Tab. 1. It is observed that removing the foreground mask information decreases the performance most, while appearance consistency affects performance the least. The scatter plot in Fig. 9c suggests that, even with no foreground/background information, we still get reasonable fittings on a considerable portion of images.

It is important to realize that SMC inference procedure produces not simply a single estimate of model configuration, but an entire population of samples from the posterior distribution for the configuration. These samples can be summarized either by the mean or by the maximum a posteriori (mode). We observed that considerable differences between the mean and mode occasionally occur, indicating that the underlying posterior is indeed multimodal (Fig. 11). Hence representing the result of shape matching by a distribution may be preferable if *e.g.* the shape model is biased or the available data is insufficient. Alternatively, methods for more intelligent mode selection could be used [10].

# 5. Summary and Discussion

We have presented a novel approach to localizing the articulated and deformable shape of a walking person in a single view. A learned shape prior and four types of local image cues are combined in a Bayesian framework. The simple chain-like model structure enables efficient spatial inference through sequential Monte Carlo.

The method can be tailored to situations where only a single image or image-pair is available, noting the fact that foreground masks can be generated from many sources other than background subtraction, *e.g.* stereo depth maps or color segmentation. For the results shown here, we used on the order of $10^4$ particles during sampling, and the inference algorithm took around one minute for each image on a 2GHz PC. It may be possible to drastically reduce the number of particles by, *e.g.* incorporating dynamic constraints or using an extended SMC with Markov transition kernel.

The model trained for one view works well for a small range of viewpoints. The method, of course, is applicable to other views, and other articulated objects. To tolerate large viewpoint changes during body fitting, we need to train new models with more flexible joint constraints and explicitly treat self-occlusion. However, a complete solution that incorporates a robust method for model selection remains an interesting open problem.

# Acknowledgments

This work was supported by DARPA/IAO HumanID under ONR contract N00014-00-1-0915 and by NSF/RHA grant IIS-0208965 and IIS-0099597.

# References

[1] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. CVPR*, pp. 8–15, 1998.

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

[3] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Proc. CVPR*, vol. 2, pp. 126–133, 2000.

[4] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[5] P. Felzenszwalb. Representation and detection of deformable shapes. In *Proc. CVPR*, vol. 1, pp. 102–108, 2003.

[6] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. In *Proc. Int. Conf. on Autom. Face and Gesture Recog.*, pp. 38–44, 1996.

[7] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comp. Graph. Stat.*, 5(1):1–25, 1996.

[8] M. Lee and I. Cohen. Human body tracking with auxiliary measurements. In *Proc. Workshop on Anal. and Model. of Faces and Gestures*, pp. 112–119, 2003.

[9] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *Proc. ECCV*, pp. 3–19, 2000.

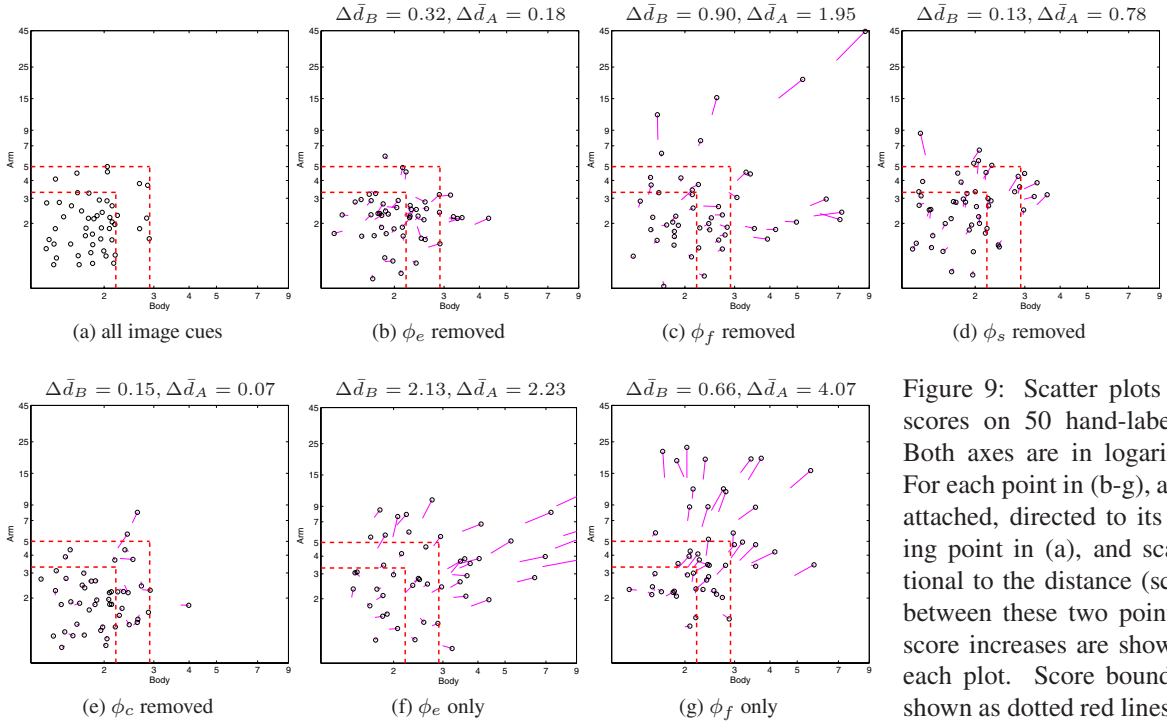
[10] T. Moeslund and E. Granum. Sequential Monte Carlo tracking of body parameters in a sub-space. In *Proc. Workshop on Anal. and Model. of Faces and Gestures*, pp. 84–91, 2003.

[11] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Trans. PAMI*, 23(4):349–361, 2001.



IEEE COMPUTER SOCIETY

Figure 8: Two examples demonstrating the inference process of Sequential Monte Carlo. Plotted are the posterior means up to the $t$-th vertex, with the distribution of each vertex summarized by the shape of its covariance ellipse (error ellipse).




(a) all image cues (b) $\phi_e$ removed (c) $\phi_f$ removed (d) $\phi_s$ removed

(e) $\phi_c$ removed (f) $\phi_e$ only (g) $\phi_f$ only

Figure 9: Scatter plots of Chamfer scores on 50 hand-labeled images. Both axes are in logarithmic scale. For each point in (b-g), a short line is attached, directed to its corresponding point in (a), and scaled proportional to the distance (score change) between these two points. Average score increases are shown on top of each plot. Score bounds in (a) are shown as dotted red lines.


[12] P. Perez, A. Blake, and M. Gangnet. Jetstream: probabilistic contour extraction with particles. In *Proc. ICCV*, pp. 524–531, 2001.

[13] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *Proc. ECCV*, pp. 700–714, 2002.

[14] M. Ruzon and C. Tomasi. Edge, junction, and corner detection using color distributions. *IEEE Trans. PAMI*, 23(11):1281–1295, 2001.

[15] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *Proc. ECCV*, pp. 702–718, 2000.

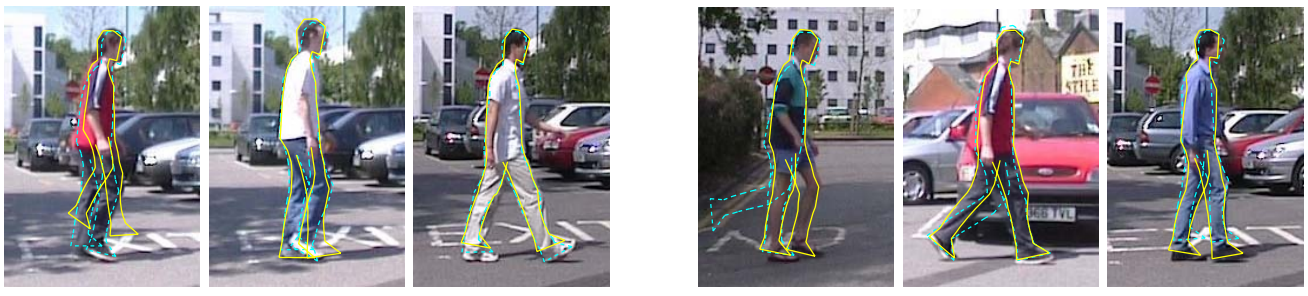
[16] C. Sminchisescu and B. Triggs. Building roadmaps of local minima of visual models. In *Proc. ECCV*, vol. 1, pp. 566–582, 2002.

[17] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. In *Proc. CVPR*, vol. 1, pp. 69–76, 2003.

[18] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Trans. PAMI*, 25(7):814–827, 2003.

[19] M. Swain and D. Ballard. Color indexing. *Int. J. Comp. Vision*, 7(1):11–32, 1991.


Figure 10: Sample results on the outdoor test set. Plotted are the posterior means, with symmetric chamfer distance scores shown in the top corners (body on the left, and arm on the right). A lower score usually indicates a better fit.



(a) (b)

Figure 11: Some cases where discrepancies exist between the mean (solid yellow) and the maximum (dotted cyan) a posteriori of the SMC output. (a) Maximum (mode) is significantly better. (b) Mean is significantly better.



IEEE COMPUTER SOCIETY