Bayesian Body Localization Using Mixture of Nonlinear Shape Models

Jiayong Zhang[1], Robert Collins[2] and Yanxi Liu[1]

[1]The Robotics Institute
Carnegie Mellon University, USA
{zhangjy,yanxi}@cs.cmu.edu

[2]Dept. of Computer Science and Engineering
The Pennsylvania State University, USA
rcollins@cse.psu.edu

Abstract

We present a 2D model-based approach to localizing human body in images viewed from arbitrary and unknown angles. The central component is a statistical shape representation of the nonrigid and articulated body contours, where a nonlinear deformation is decomposed based on the concept of parts. Several image cues are combined to relate the body configuration to the observed image, with self-occlusion explicitly treated. To accommodate large viewpoint changes, a mixture of view-dependent models is employed. Inference is done by direct sampling of the posterior mixture, using Sequential Monte Carlo (SMC) simulation enhanced with annealing and kernel move. The fitting method is independent of the number of mixture components, and does not require the preselection of a "correct" viewpoint. The models were trained on a large number of interactively labeled gait images. Preliminary tests demonstrated the feasibility of the proposed approach.

1. Introduction

We consider the problem of localizing the nonrigid and articulated shape of human body. Given an image, the task is to detect all human figures and find their limb shapes and positions. This problem has a history of over twenty years [7] and turns out to be very difficult. First, body shapes vary dramatically across subjects, poses and viewpoints. Second, body appearances are hard to model due to the wide variety of color/texture of clothing and skin. These difficulties are compounded by ambiguities caused by self-occlusion, foreshortening, and similarities of body limbs.

In this work, we assume that: 1) the image contains the whole body of a single human target, 2) there is no external occlusion, and 3) the target is approximately parallel to the imaging plane, but can be viewed from an arbitrary, unknown angle. An example of such a scenario is to fit a random shot of a person walking in a circle.

We take a 2D model-based approach to this problem. The body shape is represented by a set of landmarks along the boundary curves. The deformation of the model is constrained by the joint probabilistic distribution of landmark positions. To simultaneously accommodate anthropometric deformation, articulated motion and viewpoint effects, this distribution is inevitably complex and highly nonlinear. To proceed, we apply a hierarchical decomposition. First, the body shape is modeled by a mixture of view-dependent models. Each component model works for a small range of view angles. Second, landmarks are grouped into *parts* and *joints*, thus the nonlinear deformation of the component model can be factored into shape variations of the parts and articulated motions of the joints. Finally, the deformation of each part/joint is modeled by either one or a mixture of simple distributions conditioned on the deformation of other parts. This conditioning is designed to impose anthropometric constraints on the relative lengths of the limbs.

We formulate the matching of this mixture model to the observed image in a Bayesian framework. The likelihood is computed from several cues, including edge gradient, silhouette, skin color and region similarity. Due to the high degree of freedom, optimizing the posterior is intrinsically difficult. Therefore, we impose a sequential structure on the model. This sequential arrangement enables us to expand the configuration space and collect image information incrementally using Sequential Monte Carlo sampling. It also enables a parallel search through all the view-dependent models, where resources are dynamically allocated according to the scores of their partial fits. Besides, we employ two well-known techniques, *i.e.* annealing and MCMC move [4], to enhance the SMC inference performance.

The proposed approach has several features. First, we study the body shape at the category level, *i.e.* across subjects sampled at random from a population. To this end, the model prior is learned from a large number of real gait images that have been interactively labeled. Experiments proved that the system is able to handle subjects with large shape differences. Second, our model is designed to capture the detailed body boundary and encode both shape and pose, which is different from most existing body models. We argue that this can help localize body parts due to a better decoupling of geometric deformation from appearance variation. Third, our use of multiple deformable models does not computationally depend on the model number, nor does it require preselection of a "correct" viewpoint. Therefore it is potentially easy to increase the number of mixture

components in order to increase the modeling accuracy.

1.1. Previous Work

Most work on body fitting focuses on temporal tracking through video sequences (*e.g.,* [2, 3, 14]). In this case, search is constrained by the strong prior propagated from the past and/or future through temporal dynamics. Instead, we focus on spatial analysis of body structure without a dynamic model, and rely purely on kinematic constraints.

Existing spatial methods can be broadly grouped into two categories: learning-based and model-based. Learning-based approaches [1, 6, 13] aim at recovering body pose without extracting body parts. They are appealing because proven statistical learning techniques can be easily applied. They also can be made fast (in test mode) and suited to real-time applications. However, most existing implementations use features extracted solely from silhouette images, and do not recover anthropometric information.

Model-based methods can be divided into two types: bottom-up and top-down. The bottom-up approach [8, 12] assumes simple part models that are loosely connected. It highlights a simple and flexible structure, and thus often targets high-level tasks such as human detection. However, this approach usually depends on a robust part detector which is difficult to build in practice.

The top-down approach [10] directly explores a high-dimensional configuration space. Our method falls into this category. In general, this approach is time consuming, and may be easily trapped in local minima. However, different effects can be delineated and studied individually. When the motion is complex, multiple parametric models can be used [9]. As an extreme case, each training example may be treated as a separate model (or exemplar) [11, 15].

Our proposed part-based model is conceptually similar to pictorial structures [5]. The main differences are: 1) our part parameterization is highly flexible to capture natural body deformation, 2) our joint constraints are tight to preserve boundary continuity, and 3) our model handles self-occlusion and constraints on relative lengths of limbs, and is not a simple tree structure for inference purpose.

2. Mixture Shape Model

We take a 2D approach to localizing body shape viewed from an arbitrary and unknown angle. The basic idea is to build a finite number of 2D models, each of which works for a small range of viewpoints. Then we apply these models to the given image and combine their outputs. A Bayesian formulation is as follows. Let χ be a viewpoint index, Ω be a 2D configuration of a body projected to viewpoint χ, and $\mathcal{I}$ be an input image. $p(\chi)$ encodes the prior probability that the image is obtained from a particular viewpoint. $p(\Omega|\chi)$ encodes our prior knowledge of possible shape deformation at viewpoint χ. $p(\mathcal{I}|\Omega,\chi)$ measures the likelihood of seeing a particular image given some body configuration at viewpoint χ. Using Bayes' rule the posterior can be written as,

$$p(\Omega|\mathcal{I}) \sim \sum\nolimits_{\chi} p(\mathcal{I}|\Omega,\chi)\, p(\Omega|\chi)\, p(\chi). \quad (1)$$

This indicates that the posterior is a mixture distribution with χ as the component index. Each component $p(\mathcal{I},\Omega|\chi)$ corresponds to a different view-dependent model.

Currently we use eight component models from angles uniformly distributed in [0,2π]. They are further simplified to five basic models (as depicted in Fig. 1), noting the fact that left facing models can be constructed by flipping their right counterparts. The remaining of this section specifies the component prior $p(\Omega|\chi)$ and likelihood $p(\mathcal{I}|\Omega,\chi)$. Note that all these models are parameterized in the same way, and the viewpoint index χ will be dropped for simplicity.

2.1. Shape Prior

We represent the body shape by a set of piecewise linear boundary curves or, equivalently, by a set of K landmarks $\mathbf{v}_{1:K} = \{\mathbf{v}_k\}_{k=1}^{K}$. The 2D coordinates of these landmarks, $\{(x_k, y_k)\}$, specify the configuration $\Omega \in \Re^{2K}$ of our body model. We further divide $\mathbf{v}_{1:K}$ into M sequentially ordered *parts*, $\mathcal{W} = \{W_i\}_{i=1}^{M}$, where $W_i = \{\mathbf{v}_{i,k}\}_{k=1}^{K_i}$ consists of K_i sequentially ordered vertices (Fig. 1). W_i is virtually attached to a particular parent part, say $W_j (j < i)$, through two edges, say $\mathbf{e}_i^j$ and $\mathbf{e}_j^i$. $\mathbf{e}_i^j$ is specified by the first two vertices of W_i, and $\mathbf{e}_j^i$ is specified by some pair of vertices from W_j. $(\mathbf{e}_i^j, \mathbf{e}_j^i)$ constitute a flexible *joint* that connects W_i and W_j. The M parts are connected into a "tree" structure by a total of $(M-1)$ joints $\mathcal{J} = \{(i,j)\}$. This tree structure can be traversed sequentially by visiting $\{\mathbf{v}_{1,1}\cdots\mathbf{v}_{1,K_1}\}\{\mathbf{v}_{2,1}\cdots\mathbf{v}_{2,K_2}\}\cdots\{\mathbf{v}_{M,1}\cdots\mathbf{v}_{M,K_M}\}$.

Given the fixed landmark ordering, the prior can be decomposed into a series of marginal and conditional distributions. We start from the simplest case. Assuming the following Markov properties,

$$p(W_i|\mathbf{e}_i^j, W_{1:i-1}) = p(W_i|\mathbf{e}_i^j), \quad (2)$$

$$p(\mathbf{e}_i^j|W_{1:i-1}) = p(\mathbf{e}_i^j|\mathbf{e}_j^i), \quad (3)$$

the shape prior can be decomposed as,

$$p(\mathbf{v}_{1:K}) = p(W_1) \prod\nolimits_{(i,j)\in\mathcal{J}} p(\mathbf{e}_i^j|\mathbf{e}_j^i) p(W_i|\mathbf{e}_i^j). \quad (4)$$

This suggests two types of deformation mechanisms. The first mechanism, encoded by $p(\mathbf{e}_i^j|\mathbf{e}_j^i)$, specifies the joint motion. We parameterize this motion by a similarity transform that maps $\mathbf{e}_j^i$ to $\mathbf{e}_i^j$ with the probability given by,

$$p(\mathbf{e}_i^j|\mathbf{e}_j^i) = p(x_i, y_i, \rho_i, \theta_i) = p(x_i, y_i) p(\rho_i) p(\theta_i), \quad (5)$$

where (x_i, y_i) is translational offset, ρ_i is scale and θ_i is rotation angle.

The second mechanism, encoded by $p(W_i|\mathbf{e}_i^j)$, models the local part deformation. We parameterize W_i by its Procrustes residuals $\mathbf{r}_{i,:} = \{\mathbf{r}_{i,k}\}_{k=1}^{K_i}$ and $\mathbf{e}_i^j$, where $\mathbf{r}_{i,:}$ is modeled as multivariate normal. To predict W_i, the mean shape

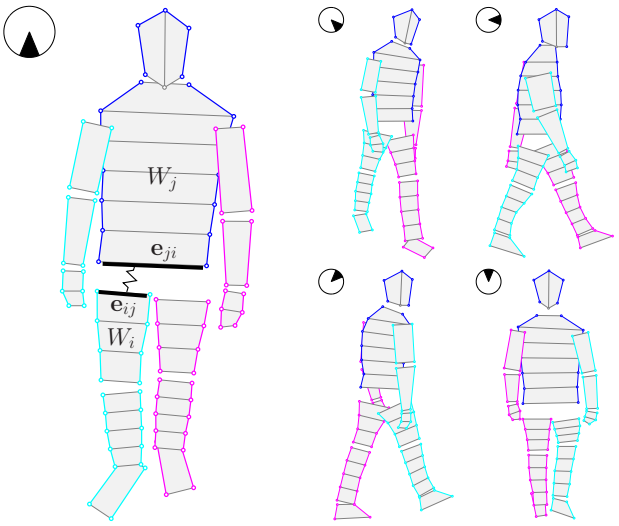

Figure 1: Topology of five basic component models. Landmarks are grouped into a collection of parts with depth order. A fixed landmark ordering is specified such that the shape can be traversed by growing one strip at a time.

of the i-th part is shifted by $\mathbf{r}_{i,:}$, followed by a similarity transform that maps the first two vertices of the shifted mean shape to $\mathbf{e}_i^j$. Assuming that the shape of W_i is independent of its location, rotation and scale, the local deformation probability simplifies to,

$$p(W_i|\mathbf{e}_i^j) = p(\mathbf{r}_{i,:}) = \prod\nolimits_k p(\mathbf{r}_{i,k}|\mathbf{r}_{i,1:k-1}). \quad (6)$$

Plugging Eqs. (5) and (6) into Eq. (4) we get,

$$p(\mathbf{v}_{1:K}) = \prod\nolimits_i p(x_i, y_i)p(\rho_i)p(\theta_i) \prod\nolimits_k p(\mathbf{r}_{i,k}|\mathbf{r}_{i,1:k-1}).$$

Now we examine those assumptions we made in deriving the above decomposition. Although the human body possesses a disaggregated structure, there exist strong dependencies among the body parts. For example, contours of two adjacent parts are mostly continuous at their connection, and anthropometric constraints exist on the relative lengths of the limbs. The continuity constraint can be imposed in our model by proper choice of the origins of joint transforms and labeling of training data. However, the limb length constraint obviously invalidates our independence assumptions in Eqs. (2), (3) and (6). By parameterizing W_i with Procrustes residuals, its length l_i becomes a nonlinear function of both the shape $\mathbf{r}_{i,:}$ and the "scale" $\|\mathbf{e}_i\|$. As a result, imposing constraints on the limb length will induce a correlation between $\mathbf{r}_{i,:}$ and $\|\mathbf{e}_i\|$.

Based on this consideration, we modify Eqs. (2), (3) and (6) as $p(W_i|\mathbf{e}_i^j, W_{1:i-1}) = p(W_i|\mathbf{e}_i^j, l_1)$, $p(\mathbf{e}_i^j|W_{1:i-1}) = p(\mathbf{e}_i^j|\mathbf{e}_j^i, l_1)$ and $p(\mathbf{r}_{i,:}|\mathbf{e}_i^j) = p(\mathbf{r}_{i,:}|\gamma_i^j)$, where l_1 is the length of W_1, and $\gamma_i^j = \|\mathbf{e}_i^j\|/l_1$. The final form of prior is,

$$p(\mathbf{v}_{1:K}) \propto \prod\nolimits_{(i,j)\in\mathcal{J}} p(x_i, y_i|\gamma_j^i)p(\rho_i|\gamma_j^i)p(\theta_i) \prod\nolimits_k p(\mathbf{r}_{i,k}|\mathbf{r}_{i,1:k-1}, \gamma_i^j). \quad (7)$$

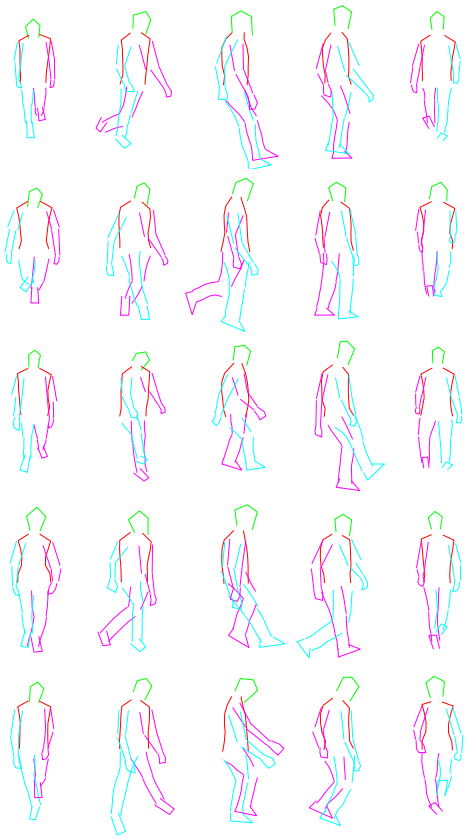

Figure 2: Selected random samples from the learned shape prior. Each row contains five samples corresponding to five component models. Each shape is normalized by aligning the torso with the associated mean shape.

We estimate densities in Eq. (7) from labeled gait images. Fig. 2 shows some random samples drawn from this learned shape prior. Note that we assume independent joint motion without activity specific constraints, thus the model is able to generate poses of activities other than walking.

2.2. Image Likelihood

Let $\Lambda = \{(i, j)\}$ be the image lattice associated with the image $\mathcal{I}$, and let $\mathcal{I}_R$ denote the image patch defined on a region $R \subset \Lambda$. As depicted in Fig. 1, the sequential structure of our model insures that the shape can be traversed in $T = K/2$ steps by growing one quadrangle strip at a time. We denote the quadrangle at step t as $\mathbf{v}_{Q_t}$, where $Q_t = 2t-3 : 2t$, and the associated region as R_t. These quadrangles partition the image into two areas: the body foreground, $R_{FG} = \cup_t R_t$, and the background, $R_{BG} = \cap_t \overline{R_t}$.

Similar to the prior, we seek a marginal and conditional decomposition of the likelihood. We start from the simplest case. Suppose: 1) there is no overlap between foreground regions, 2) $\mathcal{I}_{R_t}$ is an independent realization from a probabilistic foreground model $p(\mathcal{I}_{R_t}|FG)$, and 3) $\mathcal{I}_{R_{BG}}$ is an independent realization from a background model

$p(\mathcal{I}_{R_{BG}}|BG)$. With some assumption on the background model, the likelihood can be simplified as,

$$p(\mathcal{I}|\Omega) \propto \prod_t \frac{p(\mathcal{I}_{R_t}|FG)}{p(\mathcal{I}_{R_t}|BG)} = \prod_t \phi(\mathbf{v}_{Q_t}) \qquad (8)$$

This means $p(\mathcal{I}|\Omega)$ can be factored into the products of many local terms, each of which is a likelihood ratio defined on a local image region. Since $\mathcal{I}$ is constant, the t-th likelihood term $p(\mathcal{I}_{R_t}|FG)/p(\mathcal{I}_{R_t}|BG)$ only depends on the position of the t-th quadrangle $\mathbf{v}_{Q_t}$, and thus is simply denoted as $\phi(\mathbf{v}_{Q_t})$.

Now consider more complex cases. Through the discussion, we will incrementally modify the decomposition given by Eq. (8). First, visual patterns from different parts may not be coherent, and thus should be explained by different models. Accordingly we replace the homogeneous likelihood term $\phi(\mathbf{v}_{Q_t})$ with $\phi(\mathbf{v}_{Q_t}|\ell_{Q_t})$, where ℓ_{Q_t} is the observation model index for R_t.

Second, foreground regions come from the same object so they are very likely correlated. This can be modeled by merging multiple regions to be explained as a whole, or by using conditional terms like $p(I_{R_t}|I_{R_{t-1}})$. In this case, it is more convenient to cover the shape by a set of clusters $\mathcal{C}$. Each cluster $C \in \mathcal{C}$ contains a small number of related vertices, on which a likelihood ratio $\phi(\mathbf{v}_C)$ is defined. We impose a sequential structure on $\mathcal{C}$ by letting $\mathcal{C}_t$ be those clusters that are completely covered only at step t.

Third, due to self-occlusion, foreground regions do overlap. The effect can be modeled by introducing correction terms in the sequential process of shape construction. Suppose at step t we visit a new cluster C which covers the region R_C. By inspecting $\mathbf{v}_C$ and $\mathbf{v}_{\mathcal{C}_{1:t-1}}$, we may detect that R_C overlaps with a cluster region, say $R_{C'}$, that has been visited. In this case, we compute a correction term as follows and multiply it to the likelihood function,

$$\psi(\mathbf{v}_C, \mathbf{v}_{\mathcal{C}_{1:t-1}}) = \frac{\phi(\mathbf{v}_C, \mathbf{v}_{C'})}{\phi(\mathbf{v}_C)\phi(\mathbf{v}_{C'})}. \qquad (9)$$

In fact, $p(\mathcal{I}_R|\cdot)$ does not have to be a precise generative model. An approximate measure, such as a subjective energy term, may be good enough in practice. Another choice is to extract features $\mathcal{F}_R$ from the image patch $\mathcal{I}_R$, and replace the likelihood (ratio) to observe $\mathcal{I}_R$ by the likelihood (ratio) to observe $\mathcal{F}_R$. The definition of ϕ can then be modified as $\phi(\mathbf{v}_C) = p(\mathcal{F}_{R_C}|FG)/p(\mathcal{F}_{R_C}|BG)$. We extract features from different types of image cues. For each cue z, we define a cluster structure $\mathcal{C}^z$, and a set of likelihood terms $\phi^z(\mathbf{v}_C)$. Assuming these cues are independent, the joint likelihood can be computed as their product.

Taking all of the above into consideration, the likelihood model is expressed as,

$$p(\mathcal{I}|\Omega) \propto \prod_t \prod_z \prod_{C \in \mathcal{C}_t^z} \phi^z(\mathbf{v}_C|\ell_C)\, \psi^z(\mathbf{v}_C, \mathbf{v}_{\mathcal{C}^z_{1:t-1}}). \quad (10)$$

Our implementation of the likelihood model involves four types of image cues.

Edge Gradient. The edge potential ϕ^e is defined on the external boundary sides of each quadrangle. Given a line segment $\mathbf{e}$, we compute the average gradient strength perpendicular to $\mathbf{e}$ over all color channels. This strength is then quantized and indexed into a precomputed likelihood ratio table. If $\mathbf{e}$ is occluded, we simply set $\phi^e = 1$.

Silhouette. The silhouette potential ϕ^f is computed from a binary foreground mask $\mathcal{B}$ 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. In our experiments, we use a static camera and compute the mask using background subtraction. Assume that each mask pixel 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. The probability to observe foreground mask $\mathcal{B}$ is

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

where N_{10} and N_{11} are numbers of pixels inside the model that are labeled background and foreground respectively, γ is a constant independent of Ω. Let $\widetilde{R}_t$ be the area within R_t which is not covered by visited quadrangles, *i.e.* $\widetilde{R}_t = R_t \cap (\cap_{i<t}\overline{R_i})$. Noting that $N_{1\cdot}$ can be decomposed as $N_{1\cdot} = \sum_t N_{1\cdot}(\widetilde{R}_t)$, we have,

$$\phi^f(\mathbf{v}_{Q_t}) \propto \exp\{\alpha_f N_{10}(\widetilde{R}_t) + \beta_f N_{11}(\widetilde{R}_t)\}, \qquad (12)$$

where α_f and β_f are coefficients depending on p_{10} and p_{00}.

Skin Color. The skin potential ϕ^s is only defined on the head and arm. We use a simple skin detector based on color histogram. As the detector outputs a binary mask, a potential function similar to ϕ^f is used. Note that we only count skin and non-skin pixels in observable regions.

Region Similarity. The region similarity potential ϕ^r is defined by comparing appearances of image patches. Given two adjacent quadrangles, we compute their 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 d_{ij} is indexed to retrieve the associated likelihood ratio. This reflects the observation that appearances of the same part are likely to be coherent. We also compare each part with its surrounding area in a similar way. This reflects the observation that appearances of body and background are likely to be different.

3. Inference

There are two common strategies in using multiple deformable models. One is to fit each model completely then select the one that fits the best. This approach requires high computational cost when the model is complex. The other is to identify the "correct" model by a preprocessing step. However, sometimes it might not be possible to completely remove the uncertainty without fitting the model.

Our formulation of the fitting problem leads to the exploration of a posterior mixture. Combining the equations for shape prior (7) and imaging likelihood (10) with the Bayes

equation (1), the posterior mixture can be written as,

$$p(\chi, \Omega|\mathcal{I}) \propto p(\chi) \prod\nolimits_t \Gamma_t \cdot \Phi_t, \tag{13}$$

where,

$$\Gamma_t = \begin{cases} p(x_i, y_i|\gamma_j^i, \chi)p(\rho_i|\gamma_j^i, \chi)p(\theta_i|\chi) & Q_t \text{ is joint} \\ p(\mathbf{r}_{i,k-1:k}|\mathbf{r}_{i,1:k-2}, \gamma_i^j, \chi) & \text{otherwise} \end{cases}$$

$$\Phi_t = \prod\nolimits_z \prod\nolimits_{C \in \mathcal{C}_t^z} \phi^z(\mathbf{v}_C|\ell_C, \chi)\, \psi^z(\mathbf{v}_C, \mathbf{v}_{\mathcal{C}_{1:t-1}^z}|\chi).$$

Eq. (13) shows that the prior and likelihood terms of each component are factored into a series of simple terms with the same sequential structure. This enables us to directly sample the posterior mixture using Sequential Monte Carlo methods, which is equivalent to searching parallelly through all component shape models.

We traverse the shape model in $T = K/2$ steps. At step t, we grow two landmarks, expanding the configuration space by four dimensions. The proposal function π_t is the partial shape prior on $\mathbf{v}_{1:2t}$, which has an iterative form $\pi_t = \pi_{t-1}\Gamma_t$. The (unnormalized) importance weights is $w_t \propto w_{t-1}\Phi_t$. We may grow more landmarks, or even a whole body part, at a time. The choice depends on the balance between how much uncertainty we can remove by collecting new image information, versus how much uncertainty we will introduce by expanding the configuration space. A complete answer to this question is beyond the scope of this paper. We initialize the shape from the face region, which is the most visually informative part of a human body. The output of SMC inference procedure $\left\{\chi^{(i)}, \mathbf{v}_{0:K}^{(i)}\right\}_{i=1}^{N}$ is the sample representation of the posterior mixture. Note that viewpoint parameter χ can be marginalized out from the output if we are only interested in localizing the positions of body contours.

For complex models like ours, the basic particle filters may not work well. Here we employ two well-known improvement techniques.

3.1. Annealing

The basic idea of annealing is to gradually increase the peakness of likelihood term in order to avoid being trapped in local maxima during the early stage of the search. At each step, we compute a correction term from all visited clusters based on the change in their observation model, and multiply this correction term to the importance weight. For silhouette potential ϕ^e, we adjust the parameters $\{p_{00}, p_{01}\}$ in Eq. (11). The reason is that, when we fit a partial shape, the foreground area which the partial shape did not cover should be considered as background. As a result, a pixel in this background is more probable to be labeled as 1. This implies that using the same background model during the search procedure is inherently inappropriate. For other image cues, we use the formula $\ln \phi(t) = \xi_t \cdot \ln \phi$, where ξ_t increases linearly from $1/T$ to 1.

3.2. MCMC Move

In standard SMC procedure, all new samples of a vertex, say $\mathbf{v}_j$, are generated at step j. The number of distinct values of these samples, say S_j, is finite. As every resampling after step j results in a decrease in S_j, it will gradually diminish and eventually we lose the accuracy of the distribution of $\mathbf{v}_j$. This phenomenon is sometimes referred to as sample attrition, or particle degeneracy. To alleviate this problem, we move each particle once after every resampling procedure, using Metropolis update. Specifically, given a particle $\{\chi^{(i)}, \mathbf{v}_{0:2t}^{(i)}\}$ at step t, a new particle $\{\chi^{(i)}, \tilde{\mathbf{v}}_{0:2t}^{(i)}\}$ is generated from a Gaussian proposal density $N(\mathbf{v}_{0:2t}^{(i)}, \eta_t\Sigma_t)$, where Σ_t is the covariance matrix estimated from the current particle set, and $\eta_t < 1$. $\tilde{\mathbf{v}}_{0:2t}^{(i)}$ is accepted with the probability $\min(1, p(\chi^{(i)}, \tilde{\mathbf{v}}_{0:2t}^{(i)}|\mathcal{I})/p(\chi^{(i)}, \mathbf{v}_{0:2t}^{(i)}|\mathcal{I}))$. Currently we have not implemented jump transition between different viewpoint models.

4. Preliminary Experiment

4.1. Data Collection and Training

The first challenge in building the proposed body shape model is to obtain realistic, multi-view training data. Here we use the CMU MoBo dataset which contains 25 subjects walking on a treadmill. The subjects perform four different activities: slow walk, fast walk, incline walk and walking with a ball. All subjects are captured using six synchronized cameras distributed evenly around the treadmill. We use 150 slow-walk sequences for the training purpose. Labeling these data is inevitably laborious and difficult. We completed this task by combining interactive tracking and the presented localization method. Given a walking sequence, we first hand-labeled a number of key frames and used them to initialize an appearance-based body tracker. We then edited the tracking errors by hand and, if necessary, added more key frames. This procedure was repeated until all frames in the sequence were correctly labeled. We only labeled arms and legs on one side using interactive tracking, as their counterparts suffered from severe occlusion thus were very difficult to track. Instead, we fit the missing limbs using the presented shape model, which was learned from the partially labeled data. For each sequence, we labeled around 50 frames which covers more than a complete gait cycle. Fig. 3 shows some training examples overlayed with the labeled body contours.

An interesting fact is that, since the cameras are synchronized, the labeling at six discrete views can be interpolated to generate virtual contours at an arbitrary angle. This potentially enables us to construct densely populated body shape models. However, there are two $\pi/2$ angular gaps in MoBo camera setup which are too large to get realistic interpolation. This problem can be fixed using the periodic and symmetric property of human walking, and Fig. 4 shows an

Figure 3: Example training images of two subjects, overlayed with body contours obtained by interactive tracking (cyan) and fitting the presented shape model (magenta). Synchronized frames from all six views were shown for each subjects.

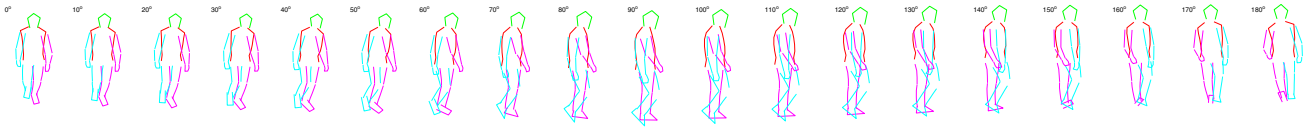

Figure 4: Virtual contours generated by interpolating labeled synchronous data (right in Fig. 3), at 19 view angles uniformly distributed between 0 and π.

example of virtual body contours generated by linear interpolation. Note that we did not use these virtual contours for training in this paper.

4.2. Test Results

We applied our mixture shape model to both indoor and outdoor cluttered scenes. First, we tested the model on CMU MoBo incline-walk and fast-walk sequences. For each sequence we randomly selected one frame, resulting in a test set of 300 images. These images were obtained from view angles similar to the training data, but the target performed different activities. Fig. 5 shows some example results. Plotted are the output of a simple mode selection procedure, which was used to deal with the possible swapping between left and right limbs in the inference output. First, the sampled body shapes generated by each component model are split into two clusters based on hand and foot positions respectively. Then we select the cluster with the highest fitting score, and output its mean shape and associated component model index (plotted as a compass in the top left corner of each image). As can be observed in Fig. 5, both the viewpoint and body boundary estimates are quite accurate. Considering the fact that the variations of body shape among these 25 subjects are quite large, the results do demonstrate the superior modeling ability of our shape model. Note that the target poses in some images are quite different from the slow-walk training data but are correctly handled. This is because that our model assumes independent joint motion without activity specific constraints.

We also applied our model to a widely tested outdoor video sequence (from Michael Black) of a person walking in a circle. Sample results are shown in Fig. 6. The video contains a total of 174 frames with the size of 320×240 pixels. Note that we did not use the sequential nature of the data to impose dynamic constraints on the body pose over time. Each frame is fit independently. This test set is challenging in several ways. First, it contains continuous change of viewpoint, while the gap between our shape models is 45^o. Second, the circle radius is quite small. The head, torso, legs and feet of the target are almost never in the same direction. Third, the elevation angle of this test data is different from our training data by 10^o–15^0 for side views, and 25^o for front and back views. The fitting algorithm shows reasonable performance on estimating the shape boundaries of body parts. However, we observed large noise in the viewpoint estimate. One obvious reason is the difference between training and testing conditions. Another reason is that the elevation angle of the test data is close to zero, in which case the inherent ambiguity between symmetric viewpoints becomes more evident.

5. Summary

We have presented a novel statistical representation of non-rigid and articulated shapes using mixture and part-based decomposition. We also proposed an effective yet computationally feasible algorithm to fit multiple view-dependent models. Preliminary experiments demonstrated the ability of our model to localize human body in images viewed from arbitrary, unknown angles. Future works include the incorporation of new image cues for robust viewpoint estimation, and the extension to poses of more general activities.

References

[1] A. Agarwal and B. Triggs. 3D human pose from silhouettes by relevance vector regression. In *Proc. CVPR*, vol. 2, pp. 882–888, 2004.

[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 and D. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comp. Vision*, 61(1):55–79, 2005.

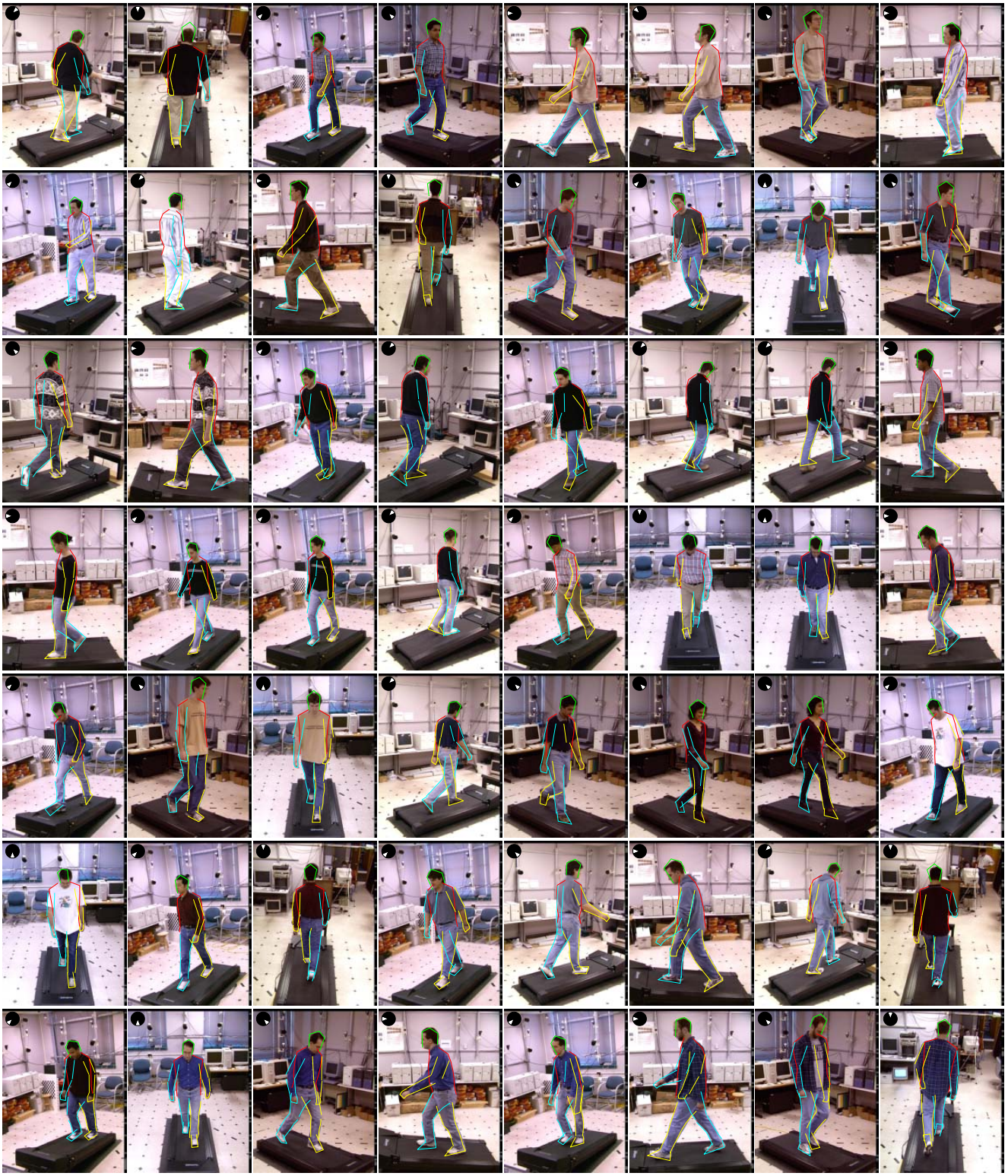

Figure 5: Sample test results on frames randomly drawn from incline-walk and fast-walk sequences of CMU MoBo dataset.

[6] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3D structure with a statistical image-based shape model. In *Proc. ICCV*, vol. 2, pp. 641–648, 2003.

[7] D. Hogg. Model-based vision: A program to see a walking

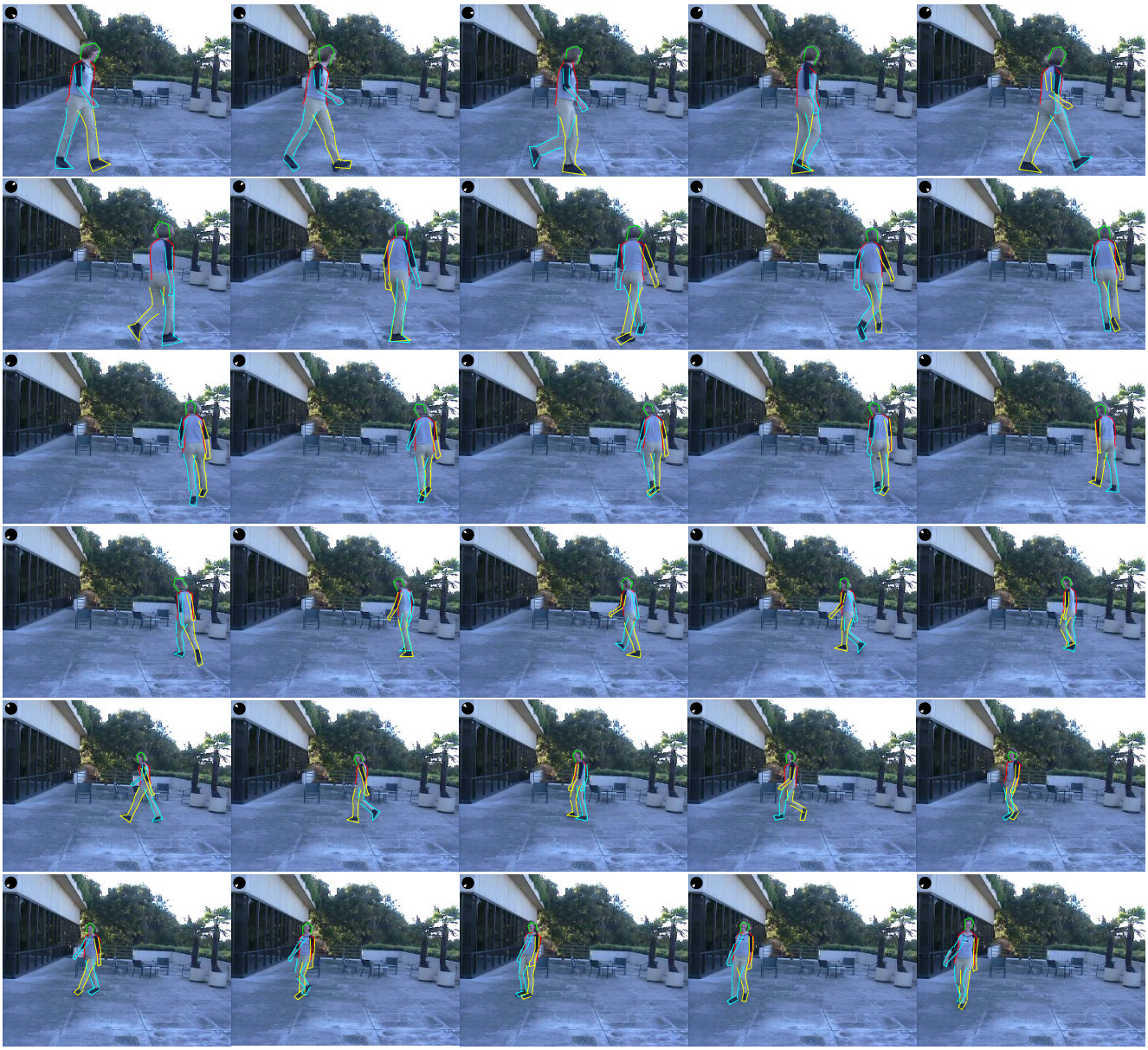

Figure 6: Sample test results on a 174 frame sequence of a person walking in a circle. Each frame is fit independently. Complete results are available at *http://www.cs.cmu.edu/~zhangjy/iccv05/*.

person. *Image and Vision Computing*, 1(1):5–20, 1983.

[8] S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *Int. J. Comp. Vision*, 43(1):45–68, 2001.

[9] X. Lan and D. Huttenlocher. A unified spatio-temporal articulated model for tracking. In *Proc. CVPR*, vol. 1, pp. 722–729, 2004.

[10] M. Lee and I. Cohen. Proposal maps driven MCMC for estimating human body pose in static images. In *Proc. CVPR*, vol. 2, pp. 334–341, 2004.

[11] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *Proc. ECCV*, vol. 3, pp. 666–680, 2002.

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

[13] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. In *Proc. CVPR*, vol. 2, pp. 721–727, 2000.

[14] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *Proc. CVPR*, vol. 1, pp. 421–428, 2004.

[15] K. Toyama and A. Blake. Probabilistic exemplar-based tracking in a metric space. In *Proc. ICCV*, vol. 2, pp. 50–57, 2001.