Body Localization in Still Images Using Hierarchical Models and Hybrid Search

Jiayong Zhang[1], Jiebo Luo[2], Robert Collins[3], and Yanxi Liu[1]

[1]The Robotics Institute
Carnegie Mellon Univ., PA

[2]Kodak Research Lab.
Eastman Kodak Co., NY

[3]Dept. of Computer Science and Eng.
Pennsylvania State Univ., PA

Abstract

We present a 3-level hierarchical model for localizing human bodies in still images from arbitrary viewpoints. We first fit a simple tree-structured model defined on a small landmark set along the body contours by Dynamic Programming (DP). The output is a series of proposal maps that encode the probabilities of partial body configurations. Next, we fit a mixture of view-dependent models by Sequential Monte Carlo (SMC), which handles self-occlusion, anthropometric constraints, and large viewpoint changes. DP and SMC are designed to search in opposite directions such that the DP proposals are utilized effectively to initialize and guide the SMC inference. This hybrid strategy of combining deterministic and stochastic search ensures both the robustness and efficiency of DP, and the accuracy of SMC. Finally, we fit an expanded mixture model with increased landmark density through local optimization. The model hierarchy is trained on a large number of gait images. Extensive tests on cluttered images with varying poses including walking, dancing and various types of sports activities demonstrate the feasibility of the proposed approach.

1. Introduction

An articulated object can be loosely defined as a structure composed of *links* and *joints*. The human body is a typical example of a non-rigid, articulated object. Body localization and 3D pose recovery from a single image has a 20-year history in computer vision, yet remains one of the fundamental unsolved problems. This problem has attracted increasing attention from researchers. This interest is motivated by a wide spectrum of potential applications, such as surveillance, video editing and annotation, human computer interfaces, and entertainment.

In this paper, we focus on localizing the 2D shapes and positions of the body parts. We seek a good summary of both body pose and shape in a given image, while avoiding the ill-posed problem of 3D recovery. We assume that: 1) the torso of the target is approximately parallel to the imaging plane, and 2) there is no serious external occlusion. However, we do not impose any constraint on the body pose or the viewpoint. No background subtraction (*e.g.,* from video) or depth information (*e.g.,* from stereo) is required. A typical example is shown in Fig. 1. Note that body joints can be localized from the open ends of adjacent parts.

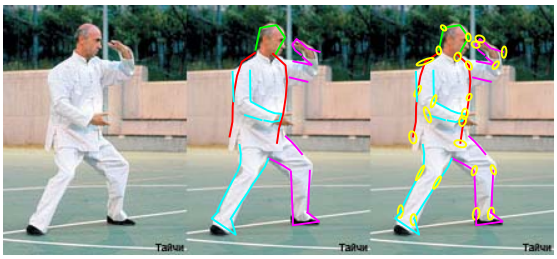

Figure 1. Given a single image (left), we want to get a boundary estimate (middle) and its uncertainty (right; shown as error ellipses of selected landmarks) for each body part.

We use a Bayesian model-based approach to take advantage of the strong priors on body deformation, and to combine multiple image cues in a robust fashion. We start from the dense body model introduced in [13]. The body shape is parameterized by the positions of landmarks densely sampled along the body contours (Fig. 2c). Given the weak assumptions of our problem setting, a direct use of such a detailed model is problematic. We adopt a coarse-to-fine strategy by introducing a 3-level hierarchical model decomposition (Fig. 2). A compact set of landmarks are identified (Fig. 2b) that characterize well the nonrigid and articulated body deformation with reduced complexity. The remaining landmarks are considered only after the inference on these key landmarks is completed.

To locate key landmarks, we employ the mixture model approach introduced in [13], which handles self-occlusion, anthropometric constraints, and large viewpoint changes. This mixture model possesses the potential for high-accuracy localization, but has a complex form for which most inference algorithms do not apply. Thus, we introduced in [13] a sequential structure so that inference could

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY

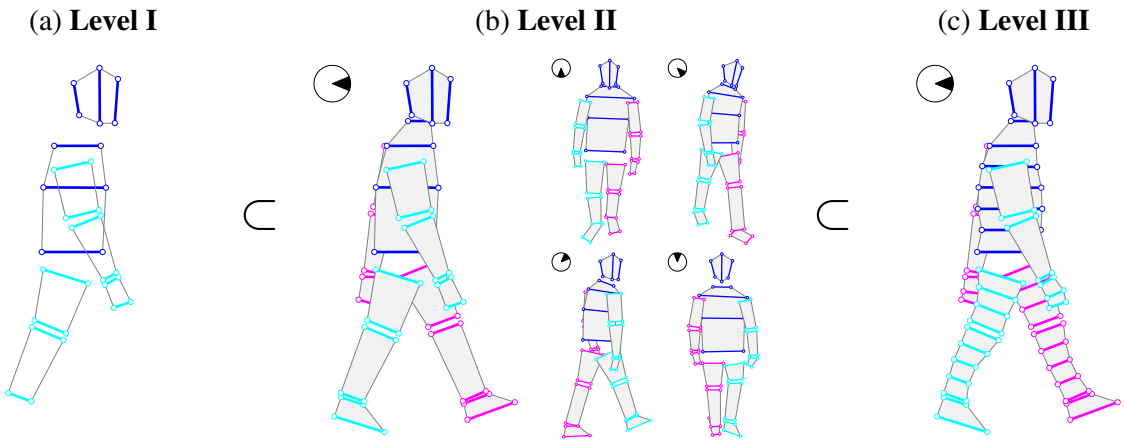

Figure 2. A three-level hierarchy of body models: (a) View-independent tree-structured model. (b) Mixture model, with eight view-dependent components from angles uniformly distributed in $[0, 2\pi]$ (only five basic ones are shown). (c) Boundary model, a mixture model similar to (b) but with increased landmark density. The model at level m is defined on $\mathcal{E}_m$, a set of K_m line segments (drawn in bold and color) that divide the body shape into quadrilaterals. The three levels are designed with a nested hierarchical structure ($\mathcal{E}_1 \subset \mathcal{E}_2 \subset \mathcal{E}_3$) in order to facilitate a coarse-to-fine search.

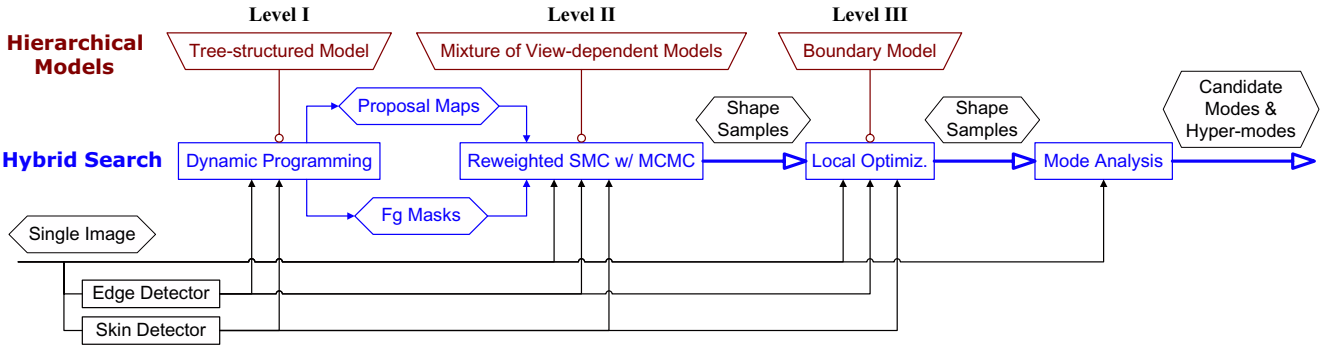

Figure 3. Algorithm flowchart. Two cues are extracted from a single image input: edge gradient, and skin/hair color. A hybrid strategy combining deterministic (DP) and stochastic (SMC) search is conducted over the model hierarchy in three steps: 1) A tree-structured model is fit by DP (from feet to head). The output is a series of proposal maps, together with foreground masks for different body parts. 2) A mixture model is fit by SMC (from head to feet). Proposal maps from DP are combined into an improved proposal function for the SMC search, while foreground masks from DP are used in computing the importance weights. 3) A detailed boundary model is fit by local optimization, initialized by the SMC output. This 3-step hybrid search is followed by a mode analysis module to get a number of candidate modes, which can be further clustered into a few compact hyper-modes.

be done by a simple top-down stochastic search (SMC).

SMC is essentially a probabilistic version of beam search, and lacks a "look-ahead" mechanism. Given a limited number of particles, SMC is prone to diverge if no strong constraints (*e.g.,* from background subtraction) are available during the early stage of the search. This difficulty is compounded by Monte-Carlo variance. One feasible solution is to use bottom-up proposals such as body part detection [4, 10]. Many frontal head detectors exist, and limb detectors have been built based on, *e.g.,* parallel line grouping [3, 8], image segmentation [6, 11], and learning techniques [5, 9]. However, building a robust part detector is difficult due to the simple structure and limited image support for one part alone, especially in situations of self-occlusion and low resolution.

Instead, we propose to generate bottom-up proposals by finding *partial bodies* (*e.g.,* the whole leg, or all parts except the head) in arbitrary poses. These proposals are used to initialize and guide the top-down SMC inference. Specifically, a tree-structured model is defined on a further reduced

set of landmarks (Fig. 2a). This model is fit by deterministic search using Dynamic Programming (DP). DP alone has been used for body localization in the past [1, 7]. In our work, DP and SMC complement each other by searching in opposite directions in the configuration space. At each step of SMC search, there is a "conjugate" proposal map from the DP output that encodes the probabilities of partial body configurations that have not been visited by SMC. These maps effectively initialize and guide the SMC inference, similar to the use of heuristic functions in A* search. This combination of deterministic and stochastic search ensures both the robustness and efficiency of DP, and the accuracy of SMC.

The result of hierarchical model fitting is a set of shape samples. Each sample is associated with a particular viewpoint. One way to summarize the output is "winner-take-all", *i.e.,* to first identify the viewpoint with most samples and then apply mode analysis only on the samples associated with that viewpoint. There are several arguments against this choice. First, in some cases body parts are in such a pose that they should be best explained by different viewpoints. Second, the samples labeled with a sub-optimal viewpoint may fit poorly on one leg but fit well on the other. Discarding them would be a waste of resources. Third, there is inherent ambiguity among different viewpoints (*e.g.,* front and back facing targets have very similar boundary shapes due to human body bilateral symmetry). Based on these considerations, we compute the final output by *viewpoint combination* rather than viewpoint selection.

We train the model hierarchy on hand-labeled gait images from the CMU Mobo Database [2]. A large number of virtual examples are also generated to increase the deformability of the shape prior. We obtain promising test results on over 100 cluttered still images with varying poses including walking, dancing and various sports activities.

2. Hierarchical Models

Our system employs three models with increasing levels of complexity: the tree-structured model, the mixture of view-dependent models, and the boundary model (Fig. 2). In this section, we discuss the definitions of prior and likelihood terms for each model. Section 3 presents the inference process by a hybrid strategy combining deterministic and stochastic search in a coarse-to-fine manner (Fig. 3).

We represent the body shape by a set of piecewise linear boundary curves, or equivalently, by L landmarks $\mathbf{v}_{1:L} = \{\mathbf{v}_l\}_{l=1}^{L}$. Landmarks on opposite boundary curves are paired into $K = L/2$ line segments $\mathbf{e}_{1:K} = \{\mathbf{e}_k\}_{k=1}^{K}$. We divide $\mathbf{e}_{1:K}$ into P sequentially ordered *parts*, $\mathcal{W} = \{W_p\}_{p=1}^{P}$, where $W_p = \{\mathbf{e}_{p,k}\}_{k=1}^{K_p}$ consists of K_p sequentially ordered line segments. Each part is virtually attached to a particular parent part through two edges, which constitute a flexible *joint*. The P parts are connected into a "tree" structure by a total of $(P-1)$ joints $\mathcal{J} = \{(p,q)\}$. The shape can be traversed sequentially by visiting $\{\mathbf{e}_{1,1}\cdots\mathbf{e}_{1,K_1}\}\{\mathbf{e}_{2,1}\cdots\mathbf{e}_{2,K_2}\}\cdots\{\mathbf{e}_{P,1}\cdots\mathbf{e}_{P,K_P}\}$.

We further divide line segments into three nested subsets, on which a 3-level hierarchical model is defined (Fig. 2). The k-th segment of the m-th level is denoted as $\mathbf{e}_k^m$. The superscript will be dropped for simplicity when it can be easily determined from the context.

2.1. Tree-structured Model for DP

The tree-structured model is defined on a small set of line segments that capture the basic body structure (Fig. 2a). These segments are grouped into 7 body parts {head, torso, thigh, calf, upper arm, lower arm, hand}. Note that the topology is simplified to only one leg and one arm. Left and right legs/arms are mapped to the same line segments. The model is made view-independent by pooling together training data from all viewpoints.

We design the prior and likelihood functions in such a way that it is possible to obtain globally optimal solutions by deterministic search. These solutions are then used to initialize and guide the inference of more complex models.

Given two adjacent segments $\mathbf{e}_{k-1}$ and $\mathbf{e}_k$, the deformation between them is parameterized by a similarity transform that maps $\mathbf{e}_{k-1}$ to $\mathbf{e}_k$ in the local coordinates of $\mathbf{e}_{k-1}$. The prior of the model is given by,

$$\begin{aligned} H(\mathbf{e}_{1:K}) &= \prod\nolimits_k H(\mathbf{e}_{k-1:k}) \\ &= \prod\nolimits_k p(x_k)p(y_k)p(\rho_k)p(\theta_k) \end{aligned} \quad (1)$$

where (x_k, y_k) is translational offset, ρ_k is relative scale and θ_k is rotation angle. Note that parts and joints are parameterized in the same way without any constraint on the form of deformation. $p(x)$, $p(y)$, $p(\rho)$ and $p(\theta)$ are modeled as histograms learned from multi-view training data.

The likelihood of the model is given by,

$$\begin{aligned} G(\mathbf{e}_{1:K}) &= \prod\nolimits_k G(\mathbf{e}_{k-1:k}) \\ &= \prod\nolimits_k \begin{cases} 1 & \text{joint} \\ \phi^s(\mathbf{e}_{k-1:k})\phi^e(\mathbf{e}_{k-1:k}) & \text{part} \end{cases} \end{aligned} \quad (2)$$

The skin potential ϕ^s is defined on head and hand segments, and is computed as the product of skin/hair probabilities at fixed positions in the quadrilateral $\mathbf{e}_{k-1:k}$. Note that the skin detector is very lenient due to the simultaneous detection of both skin and hair. The edge potential ϕ^e is computed as the average boundary probability along the two segments connecting $\mathbf{e}_{k-1}$ and $\mathbf{e}_k$.

Given the partial body configuration $\mathbf{e}_{1:k}$, the marginal posterior $Q(\mathbf{e}_k)$ has a recursive form,

$$\begin{aligned} Q(\mathbf{e}_k) &= \sum\nolimits_{\mathbf{e}_{1:k-1}} P(\mathbf{e}_{1:k})G(\mathbf{e}_{1:k}) \\ &= \sum\nolimits_{\mathbf{e}_{k-1}} H(\mathbf{e}_{k-1:k})G(\mathbf{e}_{k-1:k})Q(\mathbf{e}_{k-1}). \end{aligned} \quad (3)$$

IEEE COMPUTER SOCIETY

2.2. Mixture of View-dependent Models

The mixture model is defined on a compact set of line segments that characterize well the articulated body deformation (Fig. 2b). The mixture model consists of eight part-based component models. Each component works for a small range of view angles. Details of this mixture and part-based decomposition can be found in [13], and the main results are summarized below for completeness.

We design the prior and likelihood functions of the component model in such a way that self-occlusion and anthropometric constraints can be handled, while the model can still be fit via stochastic sequential search.

We define two deformation mechanisms: 1) shape variation of the parts, which is parameterized by Procrustes residuals $\mathbf{r}_{p,:} = \{\mathbf{r}_{p,k}\}_{k=1}^{K_p}$ where $\mathbf{r}_{p,:}$ is modeled as a multivariate normal, and 2) articulated movement of the joints, which is parameterized similar to the tree-structured model. Accordingly, the prior is decomposed as,

$$p(\mathbf{e}_{1:K}) = \prod_p p(x_p, y_p)p(\rho_p)p(\theta_p) \prod_i p(\mathbf{r}_{p,i}|\mathbf{r}_{p,1:i-1}).$$

To impose anthropometric constraints on the relative lengths of the limbs, we introduce conditioning variables $\gamma_p = \|\mathbf{e}_{p,1}\|/l_1$ for parts and $\gamma_q = \|\mathbf{e}_{q,K_q}\|/l_1$ for joints, where l_1 is the length of the face line segment, and $\mathbf{e}_{p,1}$ and $\mathbf{e}_{q,K_q}$ are line segments through which two parts are virtually attached. The final form of prior is,

$$p(\mathbf{e}_{1:K}) \propto \prod_{(p,q)\in\mathcal{J}} p(x_p, y_p|\gamma_q)p(\rho_p|\gamma_q)p(\theta_p) \prod_i p(\mathbf{r}_{p,i}|\mathbf{r}_{p,1:i-1}, \gamma_p). \quad (4)$$

We define potential functions on a set of clusters $\mathcal{C}$ that cover the body shape. Each cluster $C \in \mathcal{C}$ contains a small number of related line segments. Four types of potentials are defined based on edge, skin color, foreground appearance, and region similarity. The foreground mask is generated from the intermediate output of the tree-structured model and will be described in Sect. 3.1. The likelihood is computed as the product of all types of potentials,

$$p(\mathcal{I}|\mathbf{e}_{1:K}) \propto \prod_z \prod_{C\in\mathcal{C}^z} \phi^z(\mathbf{e}_C), \quad (5)$$

where z is the potential type index.

To handle self-occlusion caused by viewpoint change, a depth ordering of parts is assigned to each view-dependent model. The ordering is considered in the computation of potentials, resulting in clusters that contain many line segments across different parts (*e.g.,* two overlapping legs).

The final posterior has a recursive form,

$$p(\mathbf{e}_{1:K}|\mathcal{I}) \propto \prod_k \Gamma_k \cdot \Phi_k, \quad (6)$$

where,

$$\Gamma_k = \begin{cases} p(x_p, y_p|\gamma_q)p(\rho_p|\gamma_q)p(\theta_p) & \text{joint} \\ p(\mathbf{r}_{p,i}|\mathbf{r}_{p,1:i-1}, \gamma_p) & \text{part} \end{cases}$$

$$\Phi_k = \prod_z \prod_{C\in\mathcal{C}_k^z} \phi^z(\mathbf{e}_C)$$

and $\mathcal{C}_k$ is the set of clusters newly "activated" at step k.

2.3. Boundary Model

The boundary model has the same component topology as the mixture model of the previous level, except that landmarks are more densely sampled along the body contours (Fig. 2c). The new landmarks are introduced to model the detailed boundary deformation of each body part. We assume that this local deformation is conditionally independent. Thus the shape prior can be written as,

$$p(\mathbf{e}^3) = p(\mathbf{e}^2)\, p(\mathbf{e}^{3\setminus 2}|\mathbf{e}^2) = p(\mathbf{e}^2) \prod_p p(\mathbf{r}_p^{3\setminus 2}|\mathbf{r}_p^2), \quad (7)$$

where $\mathbf{e}^{3\setminus 2}$ denotes those line segments belonging to the boundary model but not to the mixture model. Note that $p(\mathbf{e}^2)$ is exactly the mixture model prior defined by Eq. (4). In the current implementation, we simply model $p(\mathbf{r}_p^{3\setminus 2}|\mathbf{r}_p^2)$ as a multivariate normal.

3. Inference Using Hybrid Search

We adopt a hybrid strategy that combines deterministic and stochastic search in a coarse-to-fine manner in the configuration space of 92 landmarks. A flowchart of the complete algorithm is shown in Fig. 3.

3.1. Dynamic Programming

We first fit the tree-structured model to the input image. The posterior in Eq. (3) has a recursive form and can be computed by DP. The basic evaluation of $Q(\mathbf{e}_k)$ is a weighted sum over quantization bins of $\mathbf{e}_{k-1}$, which can be considered as a convolution. The complexity of DP is $O(N^2)$, where N is the number of bins. Given that the typical value of N is $32 \times 32 \times 16 \times 32 \sim 10^5$, the cost of a naive DP implementation is unacceptable. Some fast algorithms exist [1], but are not applicable here because: 1) H and G are modeled by non-Gaussian histograms, and 2) the convolution kernel $(H \cdot G)$ is not homogeneous. We use two properties to derive an efficient implementation:

1. Both H and G have limited spatial support, so that most bins can be pruned during the convolution;
2. H is decomposable, so that we only need to do four 1D convolutions (over x, y, ρ and θ respectively).

Note that the second property only works for joints because G is not decomposable. As a result, part likelihood evaluation becomes the bottleneck for the speed of our DP implementation. This is why we only use simple potential functions in the tree-structured model.

The output of DP, $\{Q(\mathbf{e}_k)\}_{k=1}^K$, constitutes a series of proposal maps that encode the probabilities of partial body configurations, from which we learn an appearance model for each of {head, torso, thigh, calf} respectively.

IEEE COMPUTER SOCIETY

Figure 4. Example foreground classification based on DP outputs.

Fig. 4 shows an example. First, a number of shape samples are computed by sampling backwards from $Q(\mathbf{e}_K)$ to $Q(\mathbf{e}_1)$ (left bottom). Next, weighting masks are constructed, where pixel value is the number of sample shapes covering that pixel (left bottom). Finally, four discriminative quadratic classifiers are learned using weighted training pixels, and applied to the original image to obtain binary foreground masks (right top). Note that these masks can be very noisy and contain large false positive areas. To alleviate this problem, we perform a validity check on each scanline. A scanline is valid if the ratio between the number of its foreground and background pixels is less than a threshold (set to 4 as default in our experiments). Only valid lines are fed to the following modules as a strong cue. A similar method has been used in [7] to construct appearance models for human tracking.

3.2. Reweighted SMC with MCMC

We next fit the mixture model to the input image. Eq. (6) shows that the posterior of the view-dependent model has a recursive form, thus can be fit via Sequential Monte Carlo. A naive implementation of SMC uses shape prior Γ as the proposal, and the likelihood potential Φ as the importance function. However, without strong constraints such as background subtraction, the search is prone to diverge. Here we employ the deterministic search of DP to reduce the effect of Monte Carlo variance, and to provide a "look-ahead" mechanism. DP and SMC are designed to search in opposite directions in the configuration space, such that at step k of SMC, there is a conjugate proposal map $Q(\mathbf{e}_k)$ which encodes the probabilities of partial body configurations $\mathbf{e}_{k+1:K}$ that SMC has not yet visited. $Q(\mathbf{e}_k)$ plays a similar role in SMC as heuristic lookahead functions in A* search.

We initialize the SMC procedure by sampling $\mathbf{e}_1$ from $Q(\mathbf{e}_1) + U(\mathbf{e}_1)$, where U is a regularization term as a uniform distribution. At step k, we use reweighted importance sampling [12] to draw samples from a distribution closer to the true posterior. The proposal is modified as,

$$\Gamma'_k = \Gamma_k \cdot Q(\mathbf{e}_k), \tag{8}$$

while the importance function is the same as regular SMC. Another reweighting procedure is applied after resampling, which multiplies the weights of all particles by $1/Q(\mathbf{e}_k)$, to keep the objective function unchanged.

To sample from Γ'_k, we reformulate it as,

$$\Gamma'_k = \frac{\Gamma_k Q(\mathbf{e}_k)}{\int \Gamma_k Q(\mathbf{e}_k) d\mathbf{e}_k} \int \Gamma_k Q(\mathbf{e}_k) d\mathbf{e}_k. \tag{9}$$

The first term has an irregular distribution that cannot be directly sampled. We employ MCMC to solve this problem, with Γ_k as the transition kernel. The second integration term $\int \Gamma_k Q(\mathbf{e}_k) d\mathbf{e}_k$ is difficult to compute, so we approximate it as $Q(\tilde{\mathbf{e}}_k)$, where $\tilde{\mathbf{e}}_k$ is the MCMC output. This term is used as a weight associated with $\tilde{\mathbf{e}}_k$.

Besides the guidance of DP, we also use an annealing procedure that gradually increases the peakiness of the likelihood term in order to avoid being trapped in local maxima during the early stage of the search.

We search in parallel through all the view-dependent models. The number of samples N_i associated with a particular viewpoint χ_i is proportional to its posterior probability, reflecting a mechanism of dynamic resource allocation. In practice, however, we often observed large fluctuation of N_i during the search, which negatively affects the quality of the estimate. In addition, there are a fair number of reasons to maintain multiple models instead of a single "correct" one (Sect. 1). Therefore, we divide the eight component models into three groups of ambiguous viewpoints: {front/back-facing}, {left-facing}, {right-facing}. A regularized resource allocation scheme is employed such that,

1. Resources of the viewpoints in the same group are always kept the same.

2. Resource reallocation for the three groups is applied only at selected steps, when enough discrimination information has been accumulated.

The robust allocation is achieved by maintaining a buffer of discrete distributions, which, multiplied by the number of samples (N_i), keep track of the posterior of the viewpoint.

3.3. Local Optimization

Given the inference output of the mixture model, the boundary model can be fit by local optimization techniques. In the current implementation, a one-step importance sampling is employed. We plan to use a more accurate curve fitting method in the future work.

3.4. Mode Analysis

The output of the 3-step hybrid search described above is a set of shape samples with prior and likelihood scores.

Each sample is associated with a particular viewpoint (component model). Instead of choosing a single "correct" viewpoint, we employ a break-and-reassemble strategy to generate the final output. First, sample shapes of the whole body are broken into three body groups {torso/head, legs/feet, arms/hands}. Each body group contains those parts that are strongly correlated. Next, we apply cluster analysis to get a number of modes for each group (typically two due to flipping ambiguity of left/right limbs). The clustering is done to the samples of each viewpoint respectively, and also to the samples pooled from each ambiguous viewpoint combination. Accordingly, we get on average around 10–30 candidate modes for each body group. The exact number varies depending on the number of component models that survive the search. Finally, we extract ridge and blob features for each candidate, and sort these candidates by a linear combination of likelihood and ridge/blob scores. Note that this output of body group candidates is different from that of commonly used part detectors. Experiments demonstrate that our group candidates well satisfy the geometric constraints between different groups. Thus, given an ideal ranking function, the final output can be generated by combining the top-scoring candidate from each group. In contrast, part detectors produce unorganized output that must be assembled with the help of some geometric model.

Because many candidates are very similar, we can also cluster them into visually compact hyper-modes. This will be discussed in the Experiments section.

4. Experiments

We trained the 3-level model hierarchy using hand-labeled gait images from CMU Mobo Database. The images were captured by synchronized cameras around a treadmill. Virtual contours (at 5^o intervals) were generated by interpolating the labeling of different views. Details of this data collection process can be found in [13].

The tree-structured model was trained by pooling together samples from all view angles. Each view-dependent model was trained with (both real and virtual) samples within a 90^o view range. This range was chosen deliberately large in order to increase the deformation ability of the shape prior. In addition, models trained on gait images were relaxed by expanding the allowed range of joint angles and part deformation in order to handle arbitrary poses.

We tested three types of images: 1) 90 walking images (50 from USH Outdoor and 40 from CMU Mobo); 2) 150 break dancing images from a TV advertisement captured from a moving camera on a rainy night. The original frame is 428×240, from which the human target is roughly cropped out by hand; 3) 100 images collected from the web, with poses vary from walking to various sports activities. Note that, although some test data are originally video sequences, we did not use background subtraction or any motion cues.

The output of our system contains three sets of candidates, categorized as torso, legs and arms. The average number of candidates per category is 10 for torso, 15 for legs, and 30 for arms. We are conservative in candidate pruning to insure a high true positive rate on the diverse and challenging test data. As many candidates are very similar, we clustered them into N_h visually compact hyper-modes for better interpretation. To determine an appropriate N_h, we used a hierarchical clustering method without specifying the number of clusters. As a result, the average number of hyper-modes reaches 2.9 for torso, 5.5 for legs and 10.6 for arms. By visualizing the results at different N_h values, we observed that the compactness of hyper-modes is still satisfactory even when clustering the output into 2 for torso, 4 for legs and 5 for arms. Fig. 5 shows the complete candidate sets on 5 example images that are organized in this way. As can be observed, these candidates well satisfy the geometric constraints, and a final body assembly can be constructed by simply picking a (top-scoring) candidate from each body group.

We consider fitting to have failed if the correct torso or leg position is missing from the output. On the 340 test images, our success rate is around 40%. Fig. 6 shows a sample of successful cases, where a "preferred" mode is manually selected from each candidate set. These assembled results have two implications: 1) They demonstrate the ability of our system to generate compact candidate sets that contain good candidates; 2) They are accurate enough to provide a good starting point for ground truth labeling. Fig. 7 shows a few failures, most of which result from the presence of clutter or unusual poses.

5. Summary and Discussion

We have presented a 2D model-based approach for human body localization in still images. A hybrid search is conducted, combining stochastic and deterministic strategies in a coarse-to-fine manner, and facilitated by a hierarchical model decomposition. Improvements in both speed and accuracy have been achieved compared to using only top-down SMC. The time to fit one image is around 5 minutes on a 2GHz PC.

Experimental results demonstrate the ability of our system to generate compact candidate sets that contain good candidates. However, ideally a single "optimal" solution should be found that best matches human perception. Preserving multiple solutions has been a common practice in state-of-the-art pose estimation systems (*e.g.,* [4, 6]). This common practice reflects the difficulty of designing an objective criterion that perfectly matches human perception, particularly when given a generic problem setting and challenging data as the case in our work. We have made a preliminary attempt to design scoring functions to automati-

IEEE COMPUTER SOCIETY

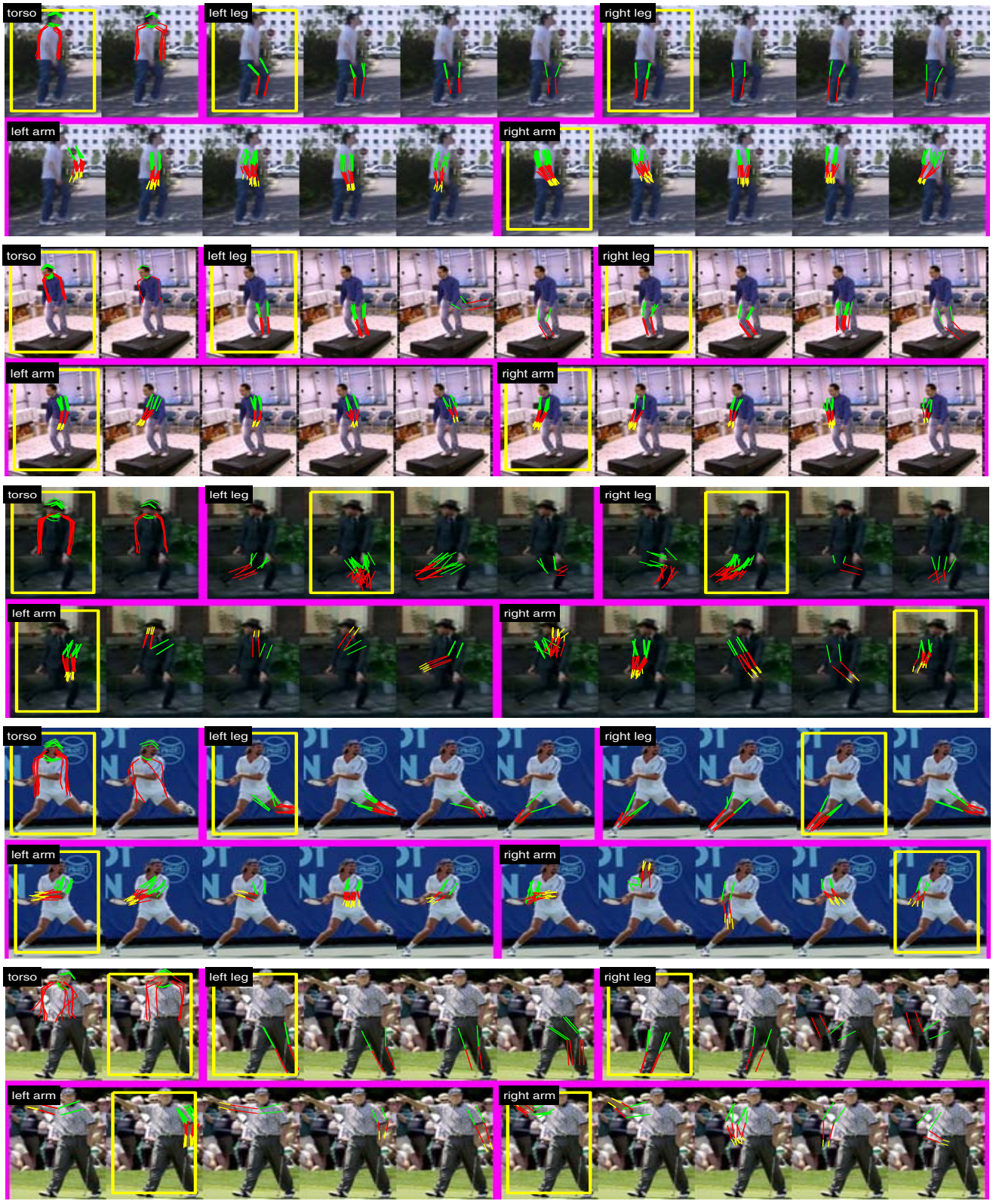

Figure 5. Complete candidate sets on five example images, clustered into {2, 4, 4, 5, 5} hyper-modes for {torso, left leg, right leg, left arm, right arm} respectively. These hyper-modes are sorted by a combination of likelihood and blob/ridge scores. "Preferred" modes are marked by yellow frames. The ideal automated scoring function would have the top-ranked mode (leftmost in each candidate set) coinciding with the preferred mode.

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY

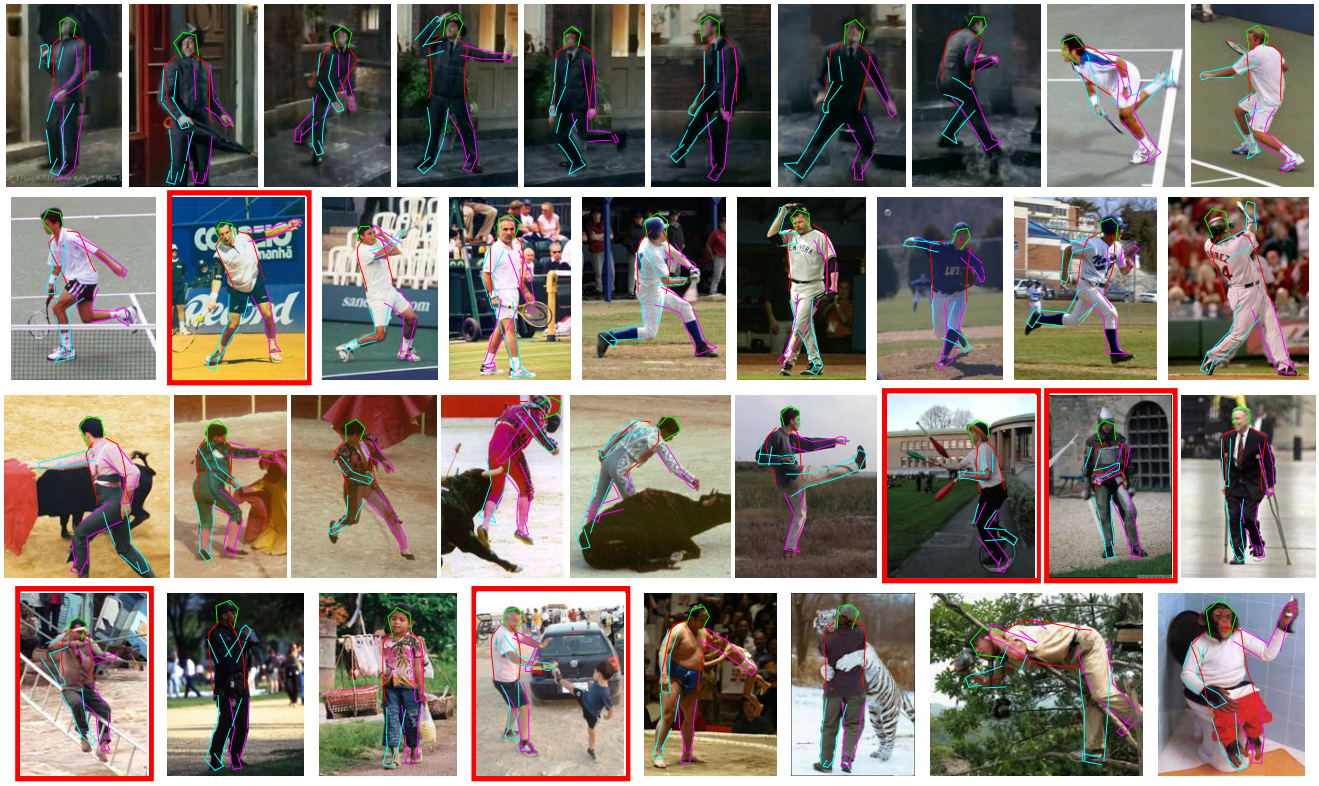

Figure 6. Example results assembled from "preferred" modes manually specified in each candidate set. Images with missing or mislabeled arms are marked with red frames.

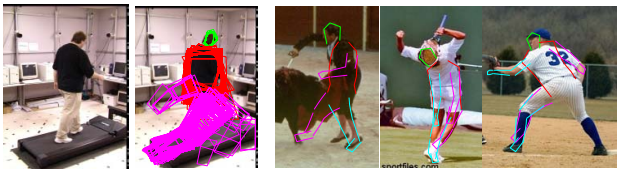

Figure 7. Examples of failure in model fitting (left for DP, and right for SMC).

cally select the "preferred" modes (Fig. 5). However, there is still a gap between our result and the ideal one, where the "preferred" mode would be ranked 1. This gap is especially obvious for the category of arms. Building a fully automatic mode selection scheme in the single image scenario remains an interesting open problem.

References

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

[2] R. Gross and J. Shi. The CMU Motion of Body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, 2001.

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

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

[5] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. ECCV*, volume 1, pages 69–81, 2004.

[6] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *Proc. CVPR*, volume 2, pages 326–333, 2004.

[7] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. CVPR*, pages 271–278, 2005.

[8] X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *Proc. ICCV*, pages 824–831, 2005.

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

[10] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *CVPR*, pages 421–428, 2004.

[11] N. Sprague and J. Luo. Clothed people detection in still images. In *Proc. ICPR*, pages 585–589, 2002.

[12] J. Sullivan, A. Blake, M. Isard, and J. Maccormick. Bayesian object localization in images. *Int. J. Comp. Vision*, 44(2):111–135, 2001.

[13] J. Zhang, R. Collins, and Y. Liu. Bayesian body localization using mixture of nonlinear shape models. In *Proc. ICCV*, volume 1, pages 725–732, 2005.