A Generative Model for Simultaneous Estimation of Human Body Shape and Pixel-Level Segmentation

Ingmar Rauschert and Robert T. Collins

Pennsylvania State University,
University Park, 16802 PA, USA

Abstract. This paper addresses pixel-level segmentation of a human body from a single image. The problem is formulated as a multi-region segmentation where the human body is constrained to be a collection of geometrically linked regions and the background is split into a small number of distinct zones. We solve this problem in a Bayesian framework for jointly estimating articulated body pose and the pixel-level segmentation of each body part. Using an image likelihood function that simultaneously generates and evaluates the image segmentation corresponding to a given pose, we robustly explore the posterior body shape distribution using a data-driven, coarse-to-fine Metropolis Hastings sampling scheme that includes a strongly data-driven proposal term.

1 Introduction

We consider the problem of human body segmentation from still images, which is extremely challenging due to the large variation in body appearance, caused by changing viewpoints, clothing and limb articulation.

Popular methods attempt to extract the whole body using graph-cuts [1, 2], individual part features using super-pixels [3–5], or shape descriptors [6–10]. A hierarchy of global-to-local constraints can be used for parsing low-level features [11]. Body parts can also be detected directly using discriminatively learned body part detectors, employing for example HoG features [12–14], coupled with an inference algorithm that works on a graph-based representation of the human body structure [14–18].

A popular approach to *general* object segmentation is based on Markov Random Fields (MRF), which defines a regular grid on the image and uses graph-cuts or belief-propagation to minimize an energy function defined over local pixel neighborhoods. A natural extension is to use knowledge of the expected object's shape. For example, side views of pedestrians yield a relatively compact set of silhouette shapes that can be used to build a shape prior in the form of level-sets [10], or be clustered into equivalence classes [6–8]. The contour person model [19] is another example of a detailed 2D human shape model, which is derived from a realistic 3D person model.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part V, LNCS 7576, pp. 704–717, 2012.

In [20–22], a concatenation of basic part shapes is used to approximate the human articulated body, which are projected and evaluated against color and shape features in the image. Typically a local search over initial parameter estimates is used to find the best shape configuration. In a step towards more realistic appearance models [9], a detailed parameterization of the human body is used together with shape context descriptors and a mixture of regressors to learn a mapping from image data directly to model parameters. The estimated parameters are then used to initialize a local, generative inference approach to fit the detailed shape model to the image.

Approaching human body segmentation as a two class object segmentation problem (body and background), PoseCut [1] and subsequent work by Kohli and Torr [2] demonstrate the power of integrating human shape priors directly into the segmentation process. Their body shape prior is computed from a dilated version of a stick-figure model, derived from a simple tree-structured parameterization of the human body. A gradient-based optimization scheme searches over the pose parameters in a local neighborhood of the true shape. However, this approach is limited in that the local search procedure requires an initial pose estimate close to the true body pose. Their method also relies on a separately computed foreground mask as an additional cue (obtained from video data), which makes their overall segmentation problem easier than our single image scenario.

Our work removes these limitations by using a multi-label segmentation algorithm to extract the human body and its individual body parts from a single image. This segmentation algorithm is embedded in a global, rather than a local search over pose parameters, with each pose hypothesis providing the set of energy potentials needed to initialize the segmentation process. Hypothesized body shapes and their extracted image segmentations are then effectively evaluated by a simple image likelihood function based on color and contour features. In order to make the global search over model parameters feasible, it is guided by a data-driven proposal function and is further split into a number of coarse-to-fine levels, with coarse levels modeling only a small number of large body parts on low resolution images, while the finest level uses the full body model and image resolution available.

2 Generative Model

To formulate the problem of human body segmentation we employ a generative model approach. The basic idea is to render an expected scene, based on a scene model M into the image and to assess how well this rendition matches the observed image I. Within a Bayesian inference framework we approximate the posterior probability distribution $p(M|I)$ of model parameters M as the product of image likelihood $p(I|M)$ and model prior $p(M)$:

$$p(M|I) \propto p(I|M)p(M) \tag{1}$$

The scene model $M = \{X, Y\}$ consists of a mixture of view-point dependent body models X, and a background model Y.

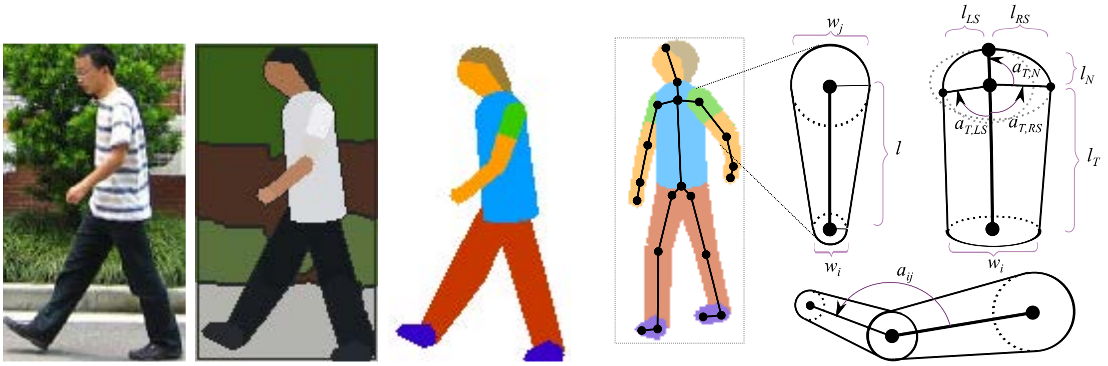

Fig. 1. a) We use a generative body model with viewpoint and clothing attributes to segment body parts from images. b) The human body is modeled using a tree-structured graph with 20 individual parts. Clothing of upper and lower body are modeled independently, each using an indicator variable **v** with three discrete states: *long*, *medium* and *short* sleeved. The face has eight different appearances, based on viewpoint.

2.1 Body and Background Model

A kinematic tree model is used to capture the articulated structure of the human body. In order to provide a tight constraint on valid configurations, we employ a mixture model of eight classes, representing poses of pedestrians as seen from eight different viewpoints (e.g. front, front-left, left, left-back, etc). The body model consists of a root and 19 parts: torso, neck, head, left/right shoulders and hips, left/right upper/lower arms and legs and left/right hands and feet. The complete set of parameters describing a particular body model incarnation $X_j \in \mathcal{X}$ is defined by a set of three state vectors $\{\mathbf{x},\ \mathbf{m},\ \mathbf{v}\}$ and a view-point indicator variable k.

We define the configuration vector $\mathbf{x}$ as a concatenation of individual configuration vectors θ_i for each of the n body parts, so that $\mathbf{x} = (\theta_1, ..., \theta_n)$. There are three parameters for each body part: its absolute length l, width w and relative rotation a with respect to its parent part. The root node of the model, which does not represent a real body part, is parameterized slightly differently in that its three different values correspond to absolute body location (x, y) and orientation a in the image. A depth layering indicator $\mathbf{m} \in \mathrm{I}^n$ is an integer vector of length $n = 20$, where the ith element is set to the layer number in which the ith part resides. Each layer contains exactly one body part. Finally, we define a visualization indicator $\mathbf{v} \in \mathrm{I}^n$ as an integer vector of length $n = 20$, where the value of v_i defines how the part should be rendered into the image. In particular, a value of $v_i = 0$ indicates that the ith part is not visible at all, $v_i \in [1, 3]$ indicates one of three possible sleeve/pants lengths, and $v_i \in [4, 12]$ indicates one of eight viewpoint renditions of the head. The neck, shoulder and hip parts of our model exist purely to add flexibility to the model, and therefore are set to be not visible at all times.

The shape of each body part is modeled as a trapezoid with a half circle at each end. One exception is the torso. We found that a simple trapezoid does not

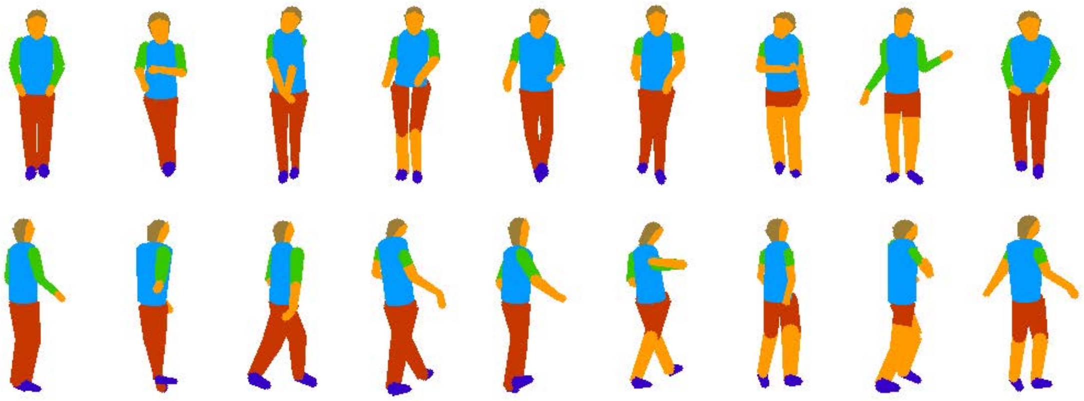

Fig. 2. We use a mixture model of viewpoint-dependent body models with eight components (e.g. front, left-front, left-side, etc.). Shown here are the mean pose (left most column) and random pose samples (other columns) of *front* (top) and *right-back*, (bottom) models.

adequately model the actual upper body shape. To improve its expressiveness, we combine torso, neck and left/right shoulders into a single shape by replacing their half circles with three ellipses, one at the bottom and two connecting the left/right shoulders to the neck (see Figure 1). This adds more flexibility to model the actual shape of the human upper body.

Because a generative model seeks to explain the entire image, we also need a model $Y = (y_1, ..., y_4, \beta)$ for the background. In this work we use four rectangular background segments of width y, that span the size of the image either horizontaly or verticaly, as indicated by β (see Sec. 4).

2.2 Model Priors

We learn a full prior model configuration $p(\mathbf{x})$ for each viewpoint model in order to exploit the inherent correlations between body parts. For example, we can model correlations between angles of the left and right arms and those of the legs, which will be quite distinct for a frontal view model as compared to a left walking model. Each prior distribution is modeled as a two component Gaussian Mixture Model with parameters learned from 30 viewpoint related pose training samples using k-means and EM-like clustering.

We make the following independence assumptions: 1) Upper and lower body configurations are independent from each other, which is reasonable because the upper body can rotate 90 degrees with respect to the lower body, and 2) *angle*, *length* and *width* parameters are modeled independently, because the true length of a body part is obscured by the effect of foreshortening (when projected from 3D world to 2D image). With these simplifications, we have three separate parameter distributions for parts in both the upper and the lower body, for a total of six distributions $(X_{ub,a}, X_{ub,l}, X_{ub,w}, X_{lb,a}, X_{lb,l}, X_{lb,w})$. The prior over the viewpoint indicator variable k is assumed to be uniform. Examples of learned prior body model configurations are shown in Figure 2.

2.3 Image Likelihood

We employ an image likelihood function that looks for image edges where the model predicts body contours, and that evaluates the inner part regions for homogeneity and, if applicable, for skin color. Therefore, the likelihood $p(I|\mathrm{M})$ of observing the image I given a proposed model M is computed by rendering the body and background model into the image, assigning part labels i to each pixel $y \in I$ via a label assignment function $l_{\mathrm{M}}(y)$, and evaluating the combined color and edge likelihood of all image pixels y:

$$p(I|\mathrm{M}) = -\sum_{y} \log p(I(y)|H_i, l(y) = i) + \log p(E(y)|C(y)) \tag{2}$$

The first term expresses the likelihood of observing pixel color $I(y)$ given that pixel y falls into a segment of part i with expected color distribution H_i. This likelihood is then given by

$$p(I(y)|H_i, l(y) = i) = \frac{H_i(I(y))}{||H_i||}. \tag{3}$$

Thus, the color likelihood essentially measures the homogeneity of a segment's color distribution, since we expect the clothing of individual body parts to be fairly uniform in color. In the case when pixel y falls into a body part not covered by clothing, a predefined skin color model H_S is used instead[1].

The second term expresses the likelihood of observing an image edge $E(y)$ given an expected contour edge $C(y)$ of the rendered body shape at pixel y,

$$p(E(y)|C(y)) = \gamma(C(y), E(y)) = C(y) \cdot E(y) \tag{4}$$

which is formulated as an edge correlation operator γ that takes the dot-product of two edge vectors. A segment's color distribution, contour edge and part assignment label $l(y)$ for a given pixel y are obtained by rendering the body model into the image and performing a localized, pixel-level segmentation (see next Section 2.4).

2.4 Multi-label Body Part Segmentation

Above formulation of the image likelihood requires knowledge of pixel assignments to background or body part segments, in order to compute a part's color distribution. However, this information is not available unless it either is modeled explicitly as part of the body model, or is extracted from the image directly. Although detailed human body shape models with relatively low dimensional parameterizations exist [20], they cannot account for the large shape variation caused by clothing. We argue that it is easier and more feasible to extract the exact body

[1] To model skin color, a normal distribution over blue and red channels in the YCbCr color space is used with $\mu = (102, 153)$ and $\sigma = (14, 12)$.

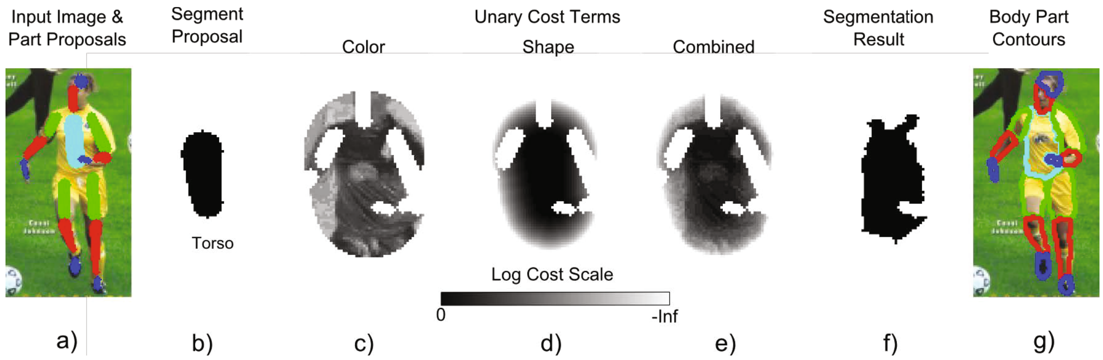

Fig. 3. Unary energy potentials for segmenting the torso are shown: a) Proposed body pose, b) the color distribution of the torso is calculated from the inner region R of the torso, c) the unary color potential, d) the unary shape potential, e) the total unary potential, f) the resulting segmentation for torso and g) for all body parts.

shape, and hence the label assignment, directly from the image. While a joint optimization of higher level (pose) and lower level (pixel) energy functions has been proposed in [22], we argue that it is unnecessary to include pixel-level constraints directly into the model parameterization. In our approach, after hypothesizing a body pose, we build a predictive model of the type of clothing and its color. Then, we employ a separate, locally constrained image segmentation to extract the body's exact shape. This is much simpler because it completely decouples the inference of low-level image features from high-level pose parameters.

Therefore, given a proposed body pose, we use a very efficient implementation of a multi-label segmentation algorithm by Alahari et al. [23] to assign part labels to each pixel. It is based on Graph-Cuts to find a minimal cut through energy potentials defined on an undirected graph where nodes correspond to pixel labels and edges specify 8-neighbor connectivity among pixels.

For each body part and background area, a proposed segment of half the part's size is used to form an energy function with unary and pair-wise potentials. The unary potential represents the likelihood of a pixel belonging to its proposed segment and has two terms, a color likelihood and a shape prior. The color likelihood is defined as the probability of being generated from the color distribution that is computed over the proposed segment. The shape prior is a weighted distance transform around the proposed segment and truncated at some distance from the proposed segment boundary. In our implementation, the segment color distributions are computed as one dimensional histograms over color indices, and the maximum shape width is set at two times the proposed segment width (see Figure 3 for an illustrative example).

3 Model Inference

Adopting a Bayesian framework, we search for model parameters M that maximize the posterior probability distribution $p(M|I)$ formed by the product of image likelihood $p(I|M)$ (Sec.2.3) and pose prior $p(M)$ (Sec.2.2). The maximum a posterior estimate (MAP) then provides a body pose (and background configuration) that can be used to partition the image into individual body part and background segments.

Due to the high-dimensional search space of our model and the multi modality of the target distribution, a straightforward greedy method for finding the MAP estimate is prone to fail by getting stuck in local minima. Furthermore, because we consider up to eight body models $M_1, ..M_8$ simultaneously, we require a method for selecting the best suitable *a posteriori* model. We therefore explore the posterior probability distribution using the Metropolis-Hastings (MH) algorithm, a variant of Markov Chain Monte Carlo sampling, which has been used in a similar setting, albeit using a single 3D model, in previous work [21]. The MH-algorithm is a sampling-based method with guided proposal moves that explore the pose space effectively by focusing on high likelihood areas, while also allowing moves to less likely regions and jumps over valleys of low likelihood [24].

3.1 Proposal Function

We use two extensions to the standard MH-algorithm: data-driven moves and dynamic mixing. Data-driven moves provide a way to focus the search based on features extracted from the image data. We scan the image with an edge-based part detector that responds well to parallel edges. The resulting response maps (one for each considered part dimension and orientation) are then used directly as the proposal maps from which body part configurations are drawn. This works well for arm and leg parts. Of course more sophisticated part detectors could be used as well, for example those suited to detect head, shoulder or hand. The dynamic mixing extension incorporates proposal moves that promote the exchange of plausible part and clothing configurations between different body models. By focusing proposals on such regions, we have a good chance of finding body parts that visually “stick out” but are not yet found.

The sampler also uses a set of more standard global and local random moves. Local perturbation of individual part parameters, such as *width*, *length* and/or part *orientation* allow refining the pose of body parts to local image features. Cross-over moves switch the subtrees of two symmetric parts such as legs or arms, and depth proposals move a body part up or down in the depth layering. Global moves allow large jumps in the parameter space to help escape local minima, but need to be crafted carefully to avoid unnecessary rejection of the move. Using the model prior, we sample a part and all its children from either their respective marginal distributions or jointly from their conditional distribution (see Section 2.1). Alternatively, we sample from the posterior distribution of the previous hierarchy level (see Section 3.2) and sample new parts at the current level according to a data-driven proposal.

3.2 Hierarchical Search

Within the tree-like structure of the human body, parts closer to the root (e.g. torso) are large and those closer to the leaves (e.g. hands) are small, which allows for a coarse-to-fine approach where both the number of parts n and the image resolution can be significantly reduced.

Our implementation of this idea uses six hierarchy levels. Starting with only the torso, more and more body parts are added to the model until all body parts are represented at the finest level. At each level, the image is re-sampled to a resolution at which the minimum width of the current model's smallest part corresponds to six pixels in the image; the resolution at which a body part can still be reliably detected. By only searching in a progressively constrained search space, from a single body part to the entire body structure, we are effectively searching a lower dimensional space at any given level of coarseness without making premature hard decisions such as committing to a single MAP estimate at each level. Down-sampling the image during search also saves significant computation time during evaluation of the image likelihood. An added bonus is that potentially distracting clutter elements in the image get smoothed out or disappear entirely at low levels of image resolution.

4 Experiments

For the evaluation of our generative model approach we use the Penn-Fudan dataset, which includes 169 images of walking pedestrians. We compare the ability of our approach to segment the body into its constituent parts with the parsing of super-pixels method by Bo and Fowlkes [5]. We also compare their performance on estimating the viewpoint with that of our approach.

4.1 Segmentation Estimation

We compare segmentation results with those given in [5], using their provided groundtruth data and evaluation script. Segmentation accuracy on individual body parts and combinations of these are computed with an evaluation metric $f(s)$ that correlates the amount of correctly labeled pixels to incorrectly labeled pixels:

$$f(\mathrm{M}_i) = \frac{||\mathrm{S}_i \cap \mathrm{S}_i^{gt}||}{||\mathrm{S}_i \cup \mathrm{S}_i^{gt}||} \tag{5}$$

The numerator counts the number of pixels where groundtruth segment S_i^{gt} and proposed segment S_i overlap. The denominator normalizes the number of pixels by the combined area of both groundtruth and proposed segments.

Results for estimated body shape and its segmentation are shown in Figures 6 and 7. Segmentation accuracies are reported for three levels of detail, denoted as evaluation class one through three. Class 3 evaluates segmentation accuracy of background (BG) versus full body (FG) segmentation. For Class 2 the body is divided into head, upper (UC) and lower body (LC) segments. For Class 1 these segments are further divided into more detailed components. Table 4.1 details the exact performance results for each evaluation class and is compared to those of Bo et al. [5].

Overall, our approach seems to perform slightly better for evaluation classes two and three, for which we achieve 79.6% and 71.7% accuracy compared to 77.2% and 69.5% respectively. On the more detailed third level our approach

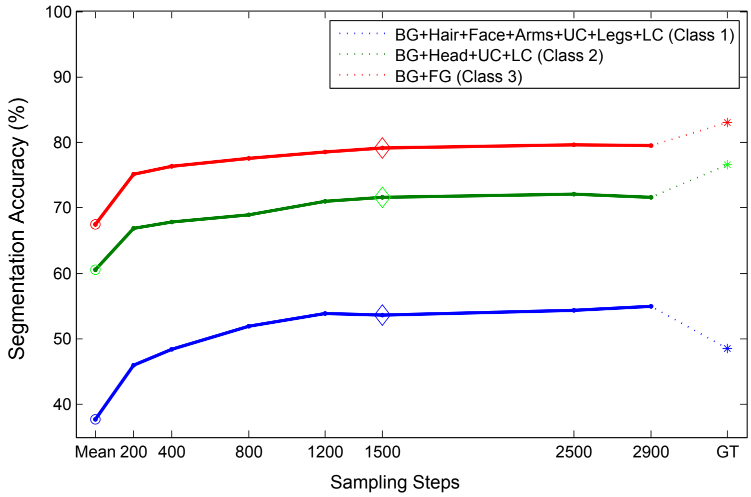

Fig. 4. Performance of proposed algorithm with varying number of sampling steps. (Marker o) Only the average prior model is evaluated. (Marker *) Performance of our annotated groundtruth model.

performs slightly worse with 55.0% compared to 57.0%, which seems to stem primarily from a decreased ability of our algorithm to properly distinguish face and hair regions. In particular, when evaluating segmentation accuracy of the head (face and hair combined) in Class 2, we perform better than [5] with 58.2% and 51.8% respectively. Otherwise there seem to be little differences in the relative performances amongst body parts. We also evaluated our algorithm with respect to the number of employed sampling steps. Figure 4 shows average segmentation accuracies as the number of samples is increased.

We can make the following observations: 1) Performance jumps quickly and levels off very gradually. Using a very low number of samples (e.g. 800) already yields a performance close to the measured maximum. 2) Interestingly, segmentation accuracy obtained from using our own groundtruth models does not yield 100% accuracy. In fact only a few percentage points are gained beyond the best performance (under evaluation class 2 and 3), while we observe even a decrease under evaluation class 1 (compared to results obtained with 800 samples or more). While we found a number of possible reasons that could explain this

Table 1. Segmentation Accuracy

Class 3:

Appr.	FG	BG	Avg
[5]	73.2	81.0	77.2
Ours	76.2	83.0	79.6

Class 2:

Appr.	Head	Up. Body	Lo. Body	BG	Avg
[5]	51.8	73.6	71.6	81.0	69.5
Ours	58.2	72.5	72.9	83.0	71.7

Class 1:

Appr.	Hair	Face	UC	Arms	LC	Legs	BG	Avg
[5]	44.9	60.8	74.8	26.2	71.2	42.0	81.0	57.3
Ours	40.0	42.8	75.2	24.7	73.0	46.6	83.0	55.0

behavior, the by far strongest cause lies in the particular way our model is designed. That is, it is designed around (essentially naked) body parts, linked by a kinematic tree, not by a more loose collection of clothing segments as provided by the groundtruth in [5]. Our groundtruth emphasizes the correct placement of body parts, and by doing so does not yield a good alignment with the visible clothing. The reason why our algorithm achieves better performance than under our groundtruth is that the flexibility of our kinematic tree model allows for deformations that, together with multi-label segmentation can accommodate a variety of different clothing scenarios.

An often seen example is that of a person who wears a long T-shirt that extends beyond the waist down to the upper legs. In our groundtruth data, the torso and upper legs are annotated truthfully so that parts of the T-Shirt that lay below the waist are assigned to the upper legs. However, our algorithm is able to extend the torso below the waist in order to model the entire T-Shirt as one entity for the torso, while shortening the upper legs. While this results in a suboptimal performance in evaluation class 3, it does not so in evaluation class 1, because here torso and legs are both assigned to the foreground and therefore the labeling distinction between the two body parts no longer exists.

From this finding we expect to achieve improved performance by modifying the model such that it more closely models the shape and configuration of different clothing styles, rather than assuming a practically naked person model.

4.2 Viewpoint Estimation

We now briefly explain and compare performance on viewpoint estimation. We use the mixture model of viewpoint dependent body models to infer a single best body model and viewpoint. However, the distinction between similar viewpoints is not always clear. A person walking to the left and seen from the side could be classified as a *left*-walking person, or, if there is a slight inclination of viewpoint into a *left-front-* or *left-back* walking person.

To handle the ambiguity in proper viewpoint classification, instead of using the viewpoint model with the maximum a posterior probability (MAP), we form a new score $p'(\mathrm{M}_i|I)$ for each viewpoint model M_i averaging the model evidence[2] of related viewpoint models (e.g. *left-front*, *front* and *right-front* models):

$$p'(\mathrm{M}_i|I) = \sum_{\mathrm{M}_j=\{\mathrm{M}_i,\mathcal{N}_i(j)\}} w_i(j) \int p(\mathrm{M}_j|I) \qquad (6)$$

where $\mathcal{N}_i$ is a set of similar viewpoint models, and $w_i(\cdot)$ is a weighting function. We experimented with a flat and a truncated exponential weighting function and found the latter to give slightly better results.

Figure 5 shows the confusion matrix for our viewpoint estimation. As expected, similar viewpoints get mixed up with related viewpoints. Overall our

[2] We obtain the model evidence from the inferred posterior distribution by summing over the (unnormalized) posterior probability of all accepted viewpoint models.

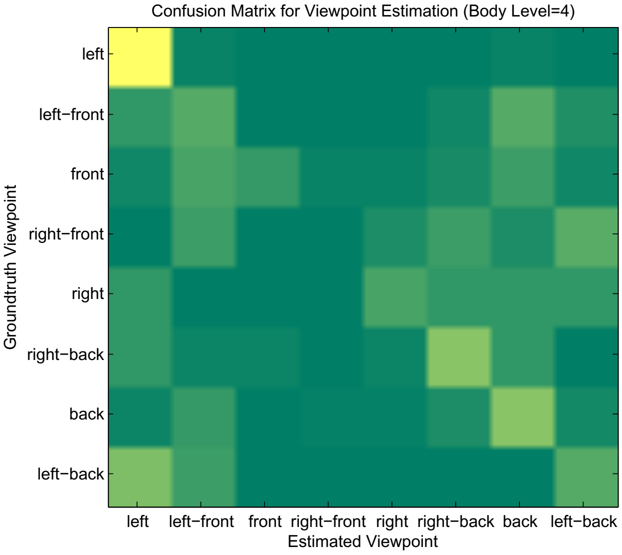

Fig. 5. Confusion matrix for predicting eight viewpoints. We achieve a correct detection rate of 46% across all viewpoints, and an average accuracy of 41% compared to 37% from the approach in [5] (they do not provide a recall rate).

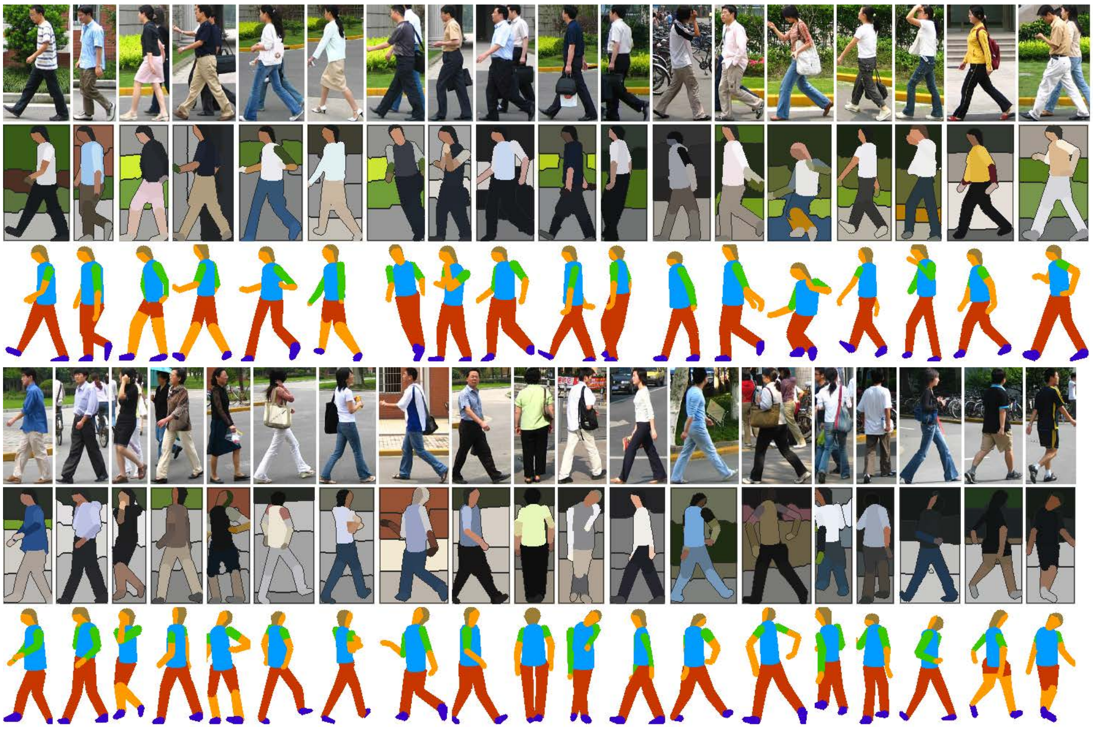

Fig. 6. Segmentation, viewpoint and clothing style results using a mixture of eight viewpoint and six clothing models. Rows show the original image (top), the best scoring body model (bottom), and the image as it is generated by the model (middle).

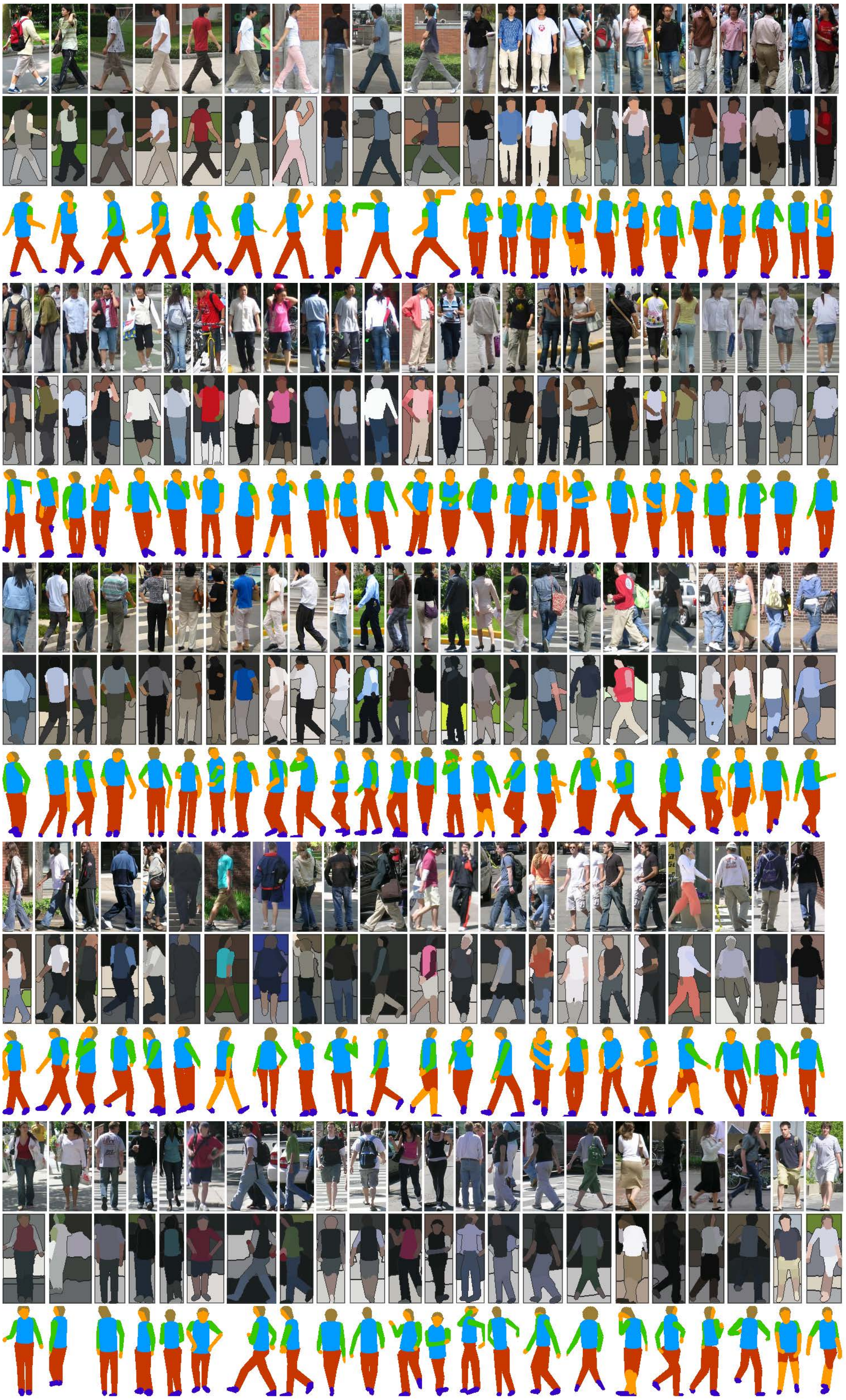

Fig. 7. Segmentation, viewpoint and clothing style results using a mixture of eight viewpoint and six clothing models. Rows show the original image (top), the best scoring body model (bottom), and the image as it is generated by the model (middle).

approach estimates the correct viewpoint in 46% of all cases with an average accuracy of 41% compared to 37% by [5].

5 Conclusion

We have developed a Bayesian method for segmenting human bodies from single images. Body segmentation is formulated as a multi-region image segmentation problem in which the human body is treated as a collection of geometrically linked segments, one for each body part, while the background is split into a small number of distinct regions. A hierarchical and data-driven variant of Metropolis Hastings sampling is used to search the large parameter space of possible body part shapes for a globally optimal body segmentation at a pixel-level. Our approach is a direct extension to earlier work on pose specific image segmentation [2], that performs only a single, full-body image segmentation and that uses a precomputed foreground mask to find only a local solution to the segmentation problem.

Performance evaluations on a large pedestrian dataset have shown that our method performs better than recent work on bottom-up human body parsing [5]. We have also used the inferred body model for viewpoint estimation for which we also achieve improved performance.

The key aspect of our approach lays in the joint estimation of body shape and image segmentation. Together the multi-label image segmentation and the generative body model extract potential body part segments from the image that are simultaneously evaluated against the expected shape and color distribution of a particular, pose specific body model. This is similar to recent work on parsing people that also use pose specific, but discriminatively learned part detectors [14, 15]. Like in our approach, their employed Histogram of Gradients (HoG) features strongly encode the shape of pose-specific body (sub-)parts.

While approaches based on discriminatively trained body part detectors might outperform our method on pose estimation, there is a growing number of work [5, 9, 25, 26] that suggest that the integration of top-down and bottom-up approaches will yield even better results. The method described in this paper provides a flexible framework for such an integration.

References

1. Bray, M., Kohli, P., Torr, P.H.S.: PoseCut: Simultaneous Segmentation and 3D Pose Estimation of Humans Using Dynamic Graph-Cuts. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 642–655. Springer, Heidelberg (2006)
2. Kohli, P., Rihan, J., Bray, M., Torr, P.: Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. International Journal of Computer Vision 79(3), 285–298 (2008)
3. Mori, G.: Guiding model search using segmentation. In: ICCV, pp. 1417–1423 (2005)
4. Srinivasan, P., Shi, J.: Bottom-up recognition and parsing of the human body. In: CVPR 2007 (2007)

5. Bo, Y., Fowlkes, C.C.: Shape-based pedestrian parsing. In: CVPR (2011)
6. Leventon, M.E., Eric, W., Grimson, L., Faugeras, O.: Statistical shape influence in geodesic active contours. In: CVPR, pp. 316–323 (2000)
7. Huang, R., Pavlovic, V., Metaxas, D.N.: A graphical model framework for coupling mrfs and deformable models. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 739–746 (2004)
8. Cremers, D., Osher, S., Soatto, S.: Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. International Journal of Computer Vision 69(3), 335–351 (2006)
9. Sigal, L., Balan, A., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: Neural Information Processing Systems Conference, NIPS (2007)
10. Cremers, D.: Nonlinear dynamical shape priors for level set segmentation. Journal of Scientific Computing 35(2-3), 132–143 (2008)
11. Sapp, B., Toshev, A., Taskar, B.: Cascaded Models for Articulated Pose Estimation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 406–420. Springer, Heidelberg (2010)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of CVPR (2005)
13. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Pose search: retrieving people using their pose. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1–8 (2009)
14. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR (2011)
15. Singh, V.K., Nevatia, R., Huang, C.: Efficient Inference with Multiple Heterogeneous Part Detectors for Human Pose Estimation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 314–327. Springer, Heidelberg (2010)
16. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: Proceedings of CVPR (2009)
17. Ramanan, D.: Learning to parse images of articulated bodies. In: Advance in Neural Information Processing Systems (2006)
18. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient matching of pictorial structures. In: Proc. IEEE Computer Vision and Pattern Recognition Conf., pp. 66–73 (2000)
19. Freifeld, O., Weiss, A., Zuffi, S., Black, M.J.: Contour people: A parameterized model of 2d articulated human shape cvpr (2010)
20. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E.: W. Haussecker, H.: Detailed human shape and pose from images. In: Proceedings of CVPR (2007)
21. Lee, M.W., Cohen, I.: A model-based approach for estimating human 3d poses in static images. IEEE Trans. Pattern Anal. Mach. Intell. 28(6), 905–916 (2006)
22. Wang, H., Koller, D.: Multi-level inference by relaxed dual decomposition for human pose segmentation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, CVPR (2011)
23. Alahari, K., Kohli, P., Torr, P.: Reduce, reuse and recycle: Efficiently solving multi-label mrfs. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
24. Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.I.: An introduction to mcmc for machine learning. Machine Learning 50(1), 5–43 (2003)
25. Borenstein, E., Ullman, S.: Combined top-down/bottom-up segmentation. PAMI 30(12) (2011)
26. Bouchard, G.: Overview of generative and discriminative hybrid models. In: IDIAP 15 Anniversary Workshop, September, 12-13 (2006)