Guiding Model Search Using Segmentation

Greg Mori*
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada V5A 1S6
mori@cs.sfu.ca

Abstract

In this paper we show how a segmentation as preprocessing paradigm can be used to improve the efficiency and accuracy of model search in an image. We operationalize this idea using an over-segmentation of an image into superpixels. The problem domain we explore is human body pose estimation from still images. The superpixels prove useful in two ways. First, we restrict the joint positions in our human body model to lie at centers of superpixels, which reduces the size of the model search space. In addition, accurate support masks for computing features on half-limbs of the body model are obtained by using agglomerations of superpixels as half-limb segments. We present results on a challenging dataset of people in sports news images.

1. Introduction

In this paper we show how a *segmentation as preprocessing* paradigm can be used to improve the efficiency and accuracy of model search in an image. We use the superpixels of Ren and Malik [10] to operationalize this idea, and test it in the problem domain of human body pose estimation from still images.

Consider the image in Figure 1(a). Given the task of localizing the joint positions of the human figure in this image, a naive search based on a particular body model would require examining every pixel as a putative left wrist location, left elbow location, and so forth. If the body model has a complicated structure (the model we use has $O(N^8)$ complexity, with N pixels in the image), the search procedure is computationally prohibitive. Instead, we use segmentation as a pre-processing step to limit the size of the state space which we must search over for each joint. Figure 1(b) shows an example over-segmentation into superpixels. In our approach we examine every superpixel center, rather than every pixel, as a putative joint position. The images we consider in our experiments are large, $N = 150 - 500K$ pixels, so the reduction to $N_{sp} \approx 1000$ superpixels provides a clear computational improvement.

(a) (b)

Figure 1: (a) Input image of 160K pixels. (b) An over-segmentation of 933 superpixels. We approximate the position of each joint of the human figure as a superpixel center, and each half-limb as being composed of superpixels.

In addition to reducing the state space of search, the superpixels can provide accuracy improvements by defining support masks on which to compute features. For the problem we are interested in, human body pose estimation, model-based approaches typically define a particular shape (such as rectangle in 2D) of half-limb on which to compute image features. Instead we use the image boundaries given by the superpixels to define support masks for half-limbs. These more accurate support masks that adhere closely to image boundaries result in features that include less background clutter.

The structure of this paper is as follows. We start by reviewing previous work in Section 2. Section 3 describes our human body model based on the superpixel representation. Section 4 describes our inference procedure. Results are presented in Section 5, and we conclude in Section 6.

2. Related Work

Some of the earliest research related to the problem of human body pose estimation is the pedestrian tracking work

*This work is supported by grants from NSERC (RGPIN-312230) and the SFU President's Research Fund.

of Hogg [3]. A vast quantity of work continued in this vein, using high degree-of-freedom 3D models of people, rendering them in the image plane, and comparing them with image data. Gavrila [2] provides a survey of this work. These approaches typically require a hand-initialized first frame, and the large number of parameters in their models lead to difficult tracking problems in high dimensional spaces.

The complexities in 3D model-based tracking have led researchers to pose the problem as one of matching to stored 2D exemplars. Toyama and Blake [19] used exemplars for tracking people as 2D edge maps. Mori and Malik [6], and Sullivan and Carlsson [17] directly address the problem of pose estimation. They stored sets of 2D exemplars upon which joint locations have been marked. Joint locations are transferred to novel images using shape matching. Shakhnarovich et al. [12] address variation in pose and appearance in exemplar matching through brute force, using a variation of locality sensitive hashing for speed to match upper body configurations of standing, front facing people in background subtracted video sequences.

Another family of approaches use a 2D model to find or track people. The approach we describe in this paper falls into this category. Felzenswalb and Huttenlocher [1] score rectangles using either a fixed clothing model or silhouettes from background subtraction of video sequences and then quickly find an optimal configuration using the distance transform to perform dynamic programming on the canonical tree model. Morris and Rehg [8] use a 2D Scaled Prismatic Model to track people and avoid the singularities associated with some 3d models. A subset of these 2D approaches apply a simple low-level detector to produce a set of candidate parts, and then a top-down procedure makes inferences about the parts and finds the best assembly. Song et al. [15] detect corner features in video sequences and model their joint statistics using tree-structured models. Ioffe and Forsyth [4] use a simple rectangle detector to find candidates and assemble them by sampling based on kinematic constraints. Ramanan and Forsyth [9] describe a self-starting tracker that builds an appearance model for people given salient rectangular primitives extracted from video sequences.

One difficulty with the tree-based body models that are often used to reduce the complexity of search is that there is no direct mechanism for preventing the reuse of image pixels. An arm with a good low-level score could be labeled as both the right and left arm. Felzenswalb and Huttenlocher [1] address this by sampling from the tree-based distribution over body poses, which is computed extremely efficiently using the distance transform, and then evaluating these samples using a more complicated model. In this work, we instead use a model that incorporates occlusion reasoning directly and use superpixels to reduce the computational difficulties in model search.

A few recent works are of particular relevance to this paper. The inference algorithm we will use to sample the distribution over human body poses is a Markov Chain Monte Carlo (MCMC) algorithm. Lee and Cohen [5] presented impressive results on pose estimation using *proposal maps*, based on face and skin detection, to guide a MCMC sampler to promising regions of the image. Tu et al. [20] perform object recognition and segmentation simultaneously, combining face and letter detectors with segmentation in a DD-MCMC framework. Sigal et al. [14] and Sudderth et al. [16] track people and hands respectively, using *loose-limbed models*, models consisting in a collection of loosely connected geometric primitives, and use non-parametric belief propagation to perform inference. Sudderth et al. build occlusion reasoning into their hand model. Sigal et al. use *shouters* to focus the attention of the inference procedure.

The idea of using an over-segmentation as support masks on which to compute features has been developed previously by Tao et al. [18] who used colour segmentation as pre-processing for stereo matching. The superpixels we use in this paper are obtained via the Normalized Cuts algorithm [13], and at this scale (1000 superpixels) provide better support masks than colour segmentation, particularly in the presence of texture.

3. Body Model

We use a 2D human body model that consists in 8 half-limbs (upper and lower arms and legs), a torso, and two occlusion variables describing relative depth orderings of the arms, legs and torso. This model is similar in spirit to the commonly used "cardboard person" models (e.g.[1, 9]) in which the torso and each half-limb is represented by a pre-defined 2D primitive (typically a rectangle), and the kinematics of the body are modelled as a collection of 2D angles formed at links between these primitives.

There are two differences between our model and these others. First, the half-limbs are restricted to respect the spatial constraints imposed by the reduction of the original image to a collection of superpixels. The endpoints of the half-limbs (i.e. the positions of the joints: elbows, shoulders, etc.) are restricted to lie in the center of one of the N_{sp} superpixels. This restriction drastically reduces the search space of possible half-limbs ($N = 150K - 300K$ pixels as endpoints to $N_{sp} \approx 1000$) while yielding only a minimal amount of lost precision in spatial location of the half-limbs. Further, each half-limb is formed as an agglomeration of superpixels. In addition to forming more complex shapes than would be possible with any particular 2D primitive, this allows for efficient computation of features for half-limbs by combining features computed on a per superpixel basis.

The second difference is the addition of the two occlusion variables, one representing the depth ordering of

the left and right arms with respect to each other and the torso, and one the depth ordering of the left and right legs. These occlusion variables allow half-limbs to claim exclusive ownership over regions of the image, avoiding the double counting of image evidence that often occurs in models that lack such information (such as tree-structured models). Further, they allow us to predict appearance of partially occluded limbs.

Given the locations of the half-limbs, and the depth ordering in the occlusion variables, we render the upper body and, separately, lower body in a back-to-front ordering to determine which superpixels are claimed by each half-limb.

More precisely, a model state X is defined as follows:

$$X = (X_{lua}, X_{lla}, X_{rua}, X_{rla}, X_{lul}, X_{lll}, X_{rul}, X_{rll}, h_u, h_l)$$

where X_{lua} represents the left upper arm, X_{rll} the right lower leg, and so forth. Each X_i takes a value which is an index into the set of all possible half-limbs, $X_i \in \{1, 2, \ldots, S\}$. The number of possible half-limbs $S \approx N_{sp}^2$, where N_{sp} is the number of superpixels. The details of constructing the set of all possible half-limbs are presented below (Section 3.1); essentially half-limbs of a few widths are placed between all pairs of superpixels within some bounded distance of each other. Note that the torso is defined implicitly in this representation, based on the shoulder and hip locations of the upper limbs.

The variables h_u and h_l represent the upper body (arms and torso) and lower body (legs) occlusion states respectively. h_u can take values representing one of four depth orderings: left-arm/right-arm/torso, left-arm/torso/right-arm, right-arm/left-arm/torso, right-arm/torso/left-arm (both arms may not be behind torso, and an upper arm cannot occlude its adjacent lower arm). h_l can take two values, representing left-leg/right-leg and right-leg/left-leg depth orderings.

We will denote by U the upper body variables, $U = \{X_{lua}, X_{lla}, X_{rua}, X_{rla}, h_u\}$, and L the lower body variables $L = \{X_{lul}, X_{lll}, X_{rul}, X_{rll}, h_l\}$.

A particular model configuration is deemed plausible if:

1. The half-limbs form a kinematically valid human body.

2. Each individual half-limb chosen looks like a half-limb itself.

3. There is symmetry in appearance of corresponding left and right half-limbs.

4. Adjacent half-limbs and corresponding left and right half-limbs have similar widths.

As such, we define a distribution over model states X as a product of four distributions:

$$p(X) = p_k(X) \cdot p_l(X) \cdot p_a(X) \cdot p_w(X) \tag{1}$$

Kinematic constraints forcing the half-limbs that comprise the body to be connected are represented in $p_k(X)$:

$$p_k(X) \propto \psi_k(X_{lua}, X_{lla}) \cdot \psi_k(X_{rua}, X_{rla}) \cdot \psi_k(X_{lul}, X_{lll}) \cdot \psi_k(X_{rul}, X_{rll}) \cdot \psi_{kt}(X) \tag{2}$$

where

$$\psi_k(X_i, X_j) = \begin{cases} 1 & \text{if } X_i, X_j \text{ adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

A pair of half-limbs are defined to be adjacent if they share an elbow/knee superpixel. $\psi_{kt}(X) \in \{0, 1\}$ enforces constraints on the size and shape of the torso induced by the upper limbs of state X.

The other distributions, $p_l(X)$, $p_a(X)$, and $p_w(X)$, are defined in the following subsections.

3.1. Half-limb Model

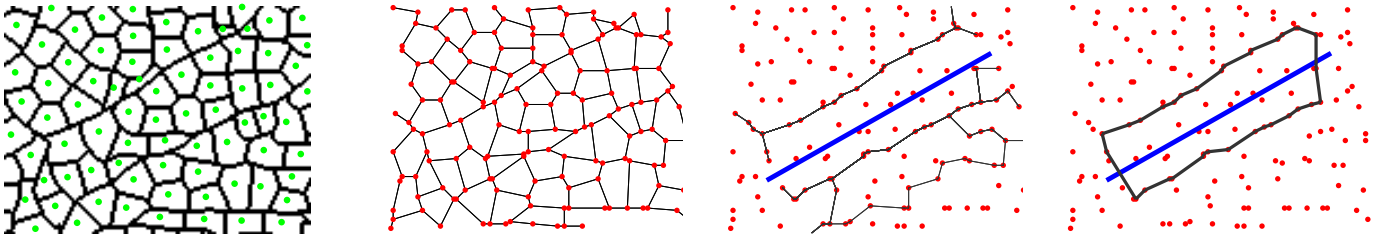
In building our set of S half-limbs, we would like to consider elongated segments, composed of superpixels, of various widths around a bone-line connecting every nearby pair of superpixels. We model a half-limb as a connected region bounded by a pair of polylines and the line segments connecting their endpoints. This modelling assumption is motivated by the available spatial structure of the superpixel segmentation, as described below.

The cues which we use when considering half-limbs in isolation, without any global assembly constraints, are (1) amount of edge energy on, and (2) overall shape of the boundary of the half-limb. We desire a representation for these constraints which can be efficiently computed given cues defined based on superpixels. As these cues are defined on the boundaries of segments, we construct a superpixel dual graph on which to compute these cues. The superpixel dual graph, shown in Figure 2(b) is constructed by taking a polygonal approximation to the original superpixel boundaries (Figure 2(a)) and creating a vertex where 3 or more superpixels meet, and an edge between vertices which are endpoints of a side of a superpixel.

Image edge energy and graph edge orientation are associated with each graph edge in this dual graph. Given a half-limb X_i bounded by a pair of polyline paths in this dual graph, we define a "limb-ness" potential for the half-limb based on the average amount of edge energy $\bar{e}_i$, average amount of orientation variation $\bar{o}_i$, and total length of these bounding paths l_i:

$$\psi_l(X_i) = e^{\frac{-1}{2\sigma_e^2}(\bar{e}_i - \mu_e)^2} \cdot e^{\frac{-1}{2\sigma_l^2}(l_i - \mu_l)^2} \cdot e^{\frac{-1}{2\sigma_o^2}(\bar{o}_i - \mu_o)^2} \tag{4}$$

Considering half-limbs bounded by all possible paths through the dual graph would be a daunting and unnecessary task. Instead, between a pair of dual graph vertices we restrict ourselves to the path which is shortest, using straight line distance edge costs. For a particular pair of superpixels



(a) (b) (c) (d)

Figure 2: Finding half-limbs. (a) Superpixel centers and boundaries. (b) Superpixel dual graph, red dot denote vertices, black lines edges. (c) Given two superpixels, shown with thick blue line between their centers, shortest paths between dual graph vertices near each superpixel are considered. (d) A chosen half-limb is shown outlined in gray. A small collection of half-limbs at various widths is kept for each pair of superpixels.

used as the endpoints of the bone-line we only consider dual graph vertices within a small range of allowable widths near these superpixels as start and end points of the polylines forming the half-limb boundaries. These shortest paths are precomputed using Dijkstra's algorithm. We also precompute edge energy, orientation, and length counts along edges in the dual graph, and hence can efficiently evaluate the potential $\psi_l(\cdot)$ for any half-limb. Figures 2(c) and (d) illustrate this process. We keep the best half-limbs, those with the highest half-limb potential, between a pair of superpixels, at a few different widths.

Note that we have not yet taken into account occlusion reasoning. When evaluating the limb-ness potential for a half-limb, we discount edges that are occluded by another half-limb and replace their edge energy and orientation variation counts by outlier costs D_e and D_o. If ω_i is the fraction of half-limb X_i occluded by other half-limbs in the upper or lower body, the limb potential is computed using:

$$\bar{e}_i = (1-\omega_i)\cdot\hat{e}_i + \omega_i D_e \quad (5)$$

$$\bar{o}_i = (1-\omega_i)\cdot\hat{o}_i + \omega_i D_o \quad (6)$$

where $\hat{e}_i$ and $\hat{o}_i$ are the average energy and orientation deviation on the unoccluded portions of the limb respectively.

The distribution $p_l(X)$ is defined as the product of these individual limb potentials, in addition to a similar potential for the torso $\psi_t(X)$:

$$p_l(X) \propto \psi_t(X)\cdot\prod_{i\in Half-limbs}\psi_l(X_i) \quad (7)$$

3.2. Appearance Consistency

We assume that the human figures in our images wear clothing that is symmetric in appearance. For example, the colour of the left upper arm should be the same as that of the right upper arm. The appearance consistency potential for corresponding left and right body parts is defined based on this assumption.

We measure the appearance similarity between a pair of half-limbs by comparing the colour histograms of the two half-limbs. We precompute a vector quantization of the colours of superpixels so that these histograms can be efficiently computed for any half-limb. Each superpixel in an input image is given the mean colour, represented in LAB colour space, of the pixels inside it. We then run kmeans on the mean superpixel colours. The appearance of a superpixel i is represented by its vector quantization label c_i, keeping the size (number of pixels) s_i of each superpixel.

The occlusion reasoning described above is used to obtain S_i, the set of superpixels comprising half-limb X_i. A colour histogram C_i is then efficiently computed using the precomputed superpixel colour labels and sizes. The j^{th} bin of C_i is:

$$C_i(j) = \sum_{k\in S_i, c_k=j} s_k \quad (8)$$

We compare the colour histograms of the two segments using the earth mover's distance (EMD), with the implementation provided by Rubner et al. [11]. The appearance consistency potential on a pair of limbs is a function of the EMD between the colour histograms of the limbs, along with an outlier cost D_c for the occluded portion of the limbs.

$$\psi_a(C_i, C_j) = e^{\frac{-1}{2\sigma_c^2}((1-\omega_{ij})\cdot EMD(C_i,C_j)+\omega_{ij}D_c)^2} \quad (9)$$

Following the notation above, $\omega_{ij} = \max(\omega_i, \omega_j)$ is the larger of the fractions of the two segments lost under occlusion.

This same $\psi_a(\cdot)$ is applied to upper and lower arms and legs to form the appearance distribution $p_a(X)$:

$$\psi_a^u(U) = \psi_a(C_{lua}, C_{rua})\cdot\psi_a(C_{lla}, C_{rla}) \quad (10)$$

$$\psi_a^l(L) = \psi_a(C_{lul}, C_{rul})\cdot\psi_a(C_{lll}, C_{rll}) \quad (11)$$

$$p_a(X) \propto \psi_a^u(U)\cdot\psi_a^l(L) \quad (12)$$

3.3. Width Consistency

We also assume that the widths of adjacent and left/right pairs of limbs are similar. A potential that measures the similarity in width of the adjacent ends of upper and lower arms and legs (width at elbow or knee), as well as widths of corresponding left/right half-limbs is also included. This potential $\psi_w(\cdot)$, and the distribution $p_w(X)$, take a similar form to those previously described.

4. Inference

Even with the reduction in state space achieved by means of the superpixels, the final inference task using our model is still a difficult one. Exact inference in this model would still require $O(N_{sp}^8)$ time. Even though $N_{sp} << N$, the number of pixels in the image, this is still intractable.

Instead, we employ Gibbs sampling, a Markov Chain Monte Carlo algorithm, to obtain samples from the distribution $p(X)$. An important detail is that the Gibbs sampling procedure operates on joint positions rather than the half-limb labels (X_i), since the kinematic constraints p_k are brittle and assign 0 probability to any disconnected body pose.

We initialize our model to a neutral standing pose in the center of the image. At each step of the algorithm we choose a particular joint J_k or occlusion label (h_u or h_l) at random. We then set the value of the occlusion label or limb(s) X_k adjacent to the joint J_k by sampling from the conditional distribution $p(h_i|X_{\hat{k}})$ or $p(X_k|X_{\hat{k}})$, where $X_{\hat{k}}$ denotes the remaining variables with those adjacent to joint J_k, or the relevant occlusion variable, removed. Computing this conditional distribution in our model is relatively simple, and involves setting the position of J_k to be any of N_{sp} locations, and re-evaluating upper or lower body potentials for the half-limbs that are adjacent at that superpixel. Our MATLAB implementation takes about 1 second per iteration of Gibbs sampling on a 2GHz AMD Opteron 246. Note that the ideas of shouters [14] or proposal maps [5] could be used in conjunction with the superpixel representation for improved performance.

5. Results

The dataset we use is a collection of sports news photographs of baseball players. This dataset is very challenging, with dramatic variations in pose and clothing, and significant background clutter. Four images from our dataset were used to set the free parameters in our body model, and 53 images were used for testing. For each input image, we compute superpixels[1] and colour, edge energy, orientation, and length cues upon them in a pre-processing step. The Gibbs sampling algorithm is then run 10 times, with 200 sampling iterations per run. Modes from the distribution $p(X)$ are obtained by running kmeans on the set of 2d joint positions of the set of samples.

Figure 3 shows quantitative results, histograms of pixel error in joint positions. The scale of the people in the test images is quite large - the average height is approximately 400 pixels. Figures 3(a-c) show results using the entire test set. There are a few very large errors in these histograms. However, we are able to detect when such difficulties have occurred. Figures 3(d-f) show results using only the top 15 images from our test set, as sorted by unnormalized $p(X)$ values. Further, joint localization errors are broken down into upper and lower body errors. Lower body joints (hips, knees, ankles) are reasonably well localized, while upper body joints (shoulders, elbows, wrists) prove extremely difficult to find. These quantitative results, while by no means accurate, are compare favourably to the state of the art on this extremely difficult problem.

In Figure 4 we show qualitative results on images from our test set. The top row of each set shows input images overlayed with superpixel boundaries, followed by recovered human body poses, and segmentation masks corresponding to the half-limbs. The first five input images (top row) are the top five matches from our test set chosen in a principled fashion (sorted $p(X)$ values), while the remaining examples are of the usual judiciously chosen variety. In order to shed more light on the quantitative results, average joint errors in the top five images are $42.4, 27.7, 57.3, 37.4, 25.4$, significant error measurements for qualitatively reasonable results.

6. Discussion

In this paper we have shown how segmentation can be used as a preprocessing step to improve the efficiency and accuracy of model search in an image. We have demonstrated two advantages of this approach – reducing the state space of model search and defining accurate support masks on which to compute features. Using these ideas, we have shown promising results on the difficult task of human body pose estimation in still images.

The results presented in this paper are comparable in quality to those in our previous work [7]. In our previous method, an initial coarse segmentation followed by a classifier was used to provide candidate half-limbs. An ad-hoc assembly method that required solving a constraint satisfaction problem was then used to assemble these candidate half-limbs. However, this CSP step was brittle, requiring that at least 3 half-limbs were found by the initial segmentation-classification stage, and caused unrecoverable errors.

In contrast, the method presented in this paper performs inference over superpixel locations using a body model. This idea seems generally useful, and we believe it could be applied to other object recognition problems.

[1] Sample MATLAB code for computing the superpixels is available at: `http://www.cs.sfu.ca/~mori/research/superpixels`

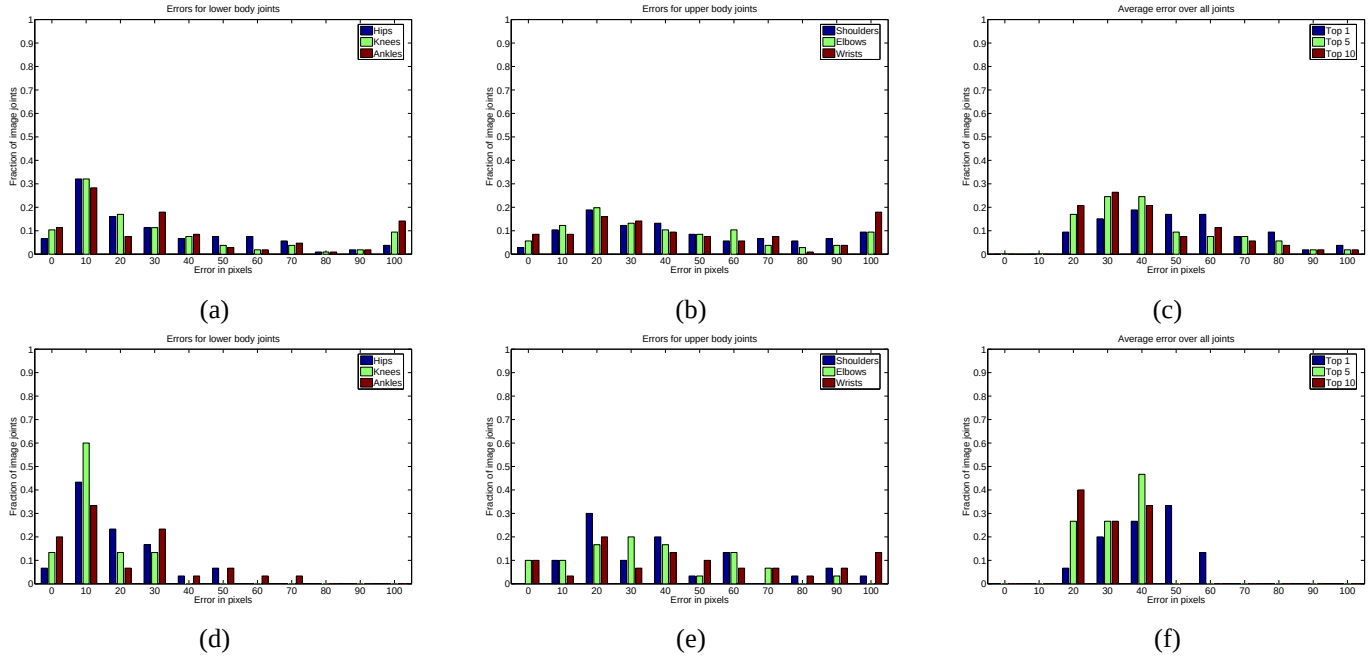


Figure 3: Histograms over pixel error in joint positions. (a-c) Histograms computed using all images in our test set. (d-f) Histograms using best 15 matching images in dataset (highest unnormalized $p(X)$). (a,d)/(b,e) Error in lower/upper body joint positions for overall best configuration out of top 10 modes of $p(X)$. (c,f) Average error over all joints.

References

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

[2] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.

[3] D. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[4] S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *Int. Journal of Computer Vision*, 43(1):45–68, 2001.

[5] M. W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, volume 2, pages 334–341, 2004.

[6] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV LNCS 2352*, volume 3, pages 666–680, 2002.

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

[8] D. Morris and J. Rehg. Singularity analysis for articulated object tracking. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages 289–296, 1998.

[9] D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 2003.

[10] X. Ren and J. Malik. Learning a classification model for segmentation. In *Proc. 9th Int. Conf. Computer Vision*, volume 1, pages 10–17, 2003.

[11] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. Journal of Computer Vision*, 40(2):99–121, 2000.

[12] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *Proc. 9th Int. Conf. Computer Vision*, volume 2, pages 750–757, 2003.

[13] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. PAMI*, 22(8):888–905, 2000.

[14] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 2004.

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

[16] E. Sudderth, M. Mandel, W. Freeman, and A. Willsky. Distributed occlusion reasoning for tracking with nonparametric belief propagation. In *NIPS*, 2004.

[17] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *European Conference on Computer Vision LNCS 2352*, volume 1, pages 629–644, 2002.

[18] H. Tao, H. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *Proc. 8th Int. Conf. Computer Vision*, 2001.

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

[20] Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image parsing: segmentation, detection, and recognition. In *Proc. 9th Int. Conf. Computer Vision*, pages 18–25, 2003.

Figure 4: Two sets of sample results. In each set, top row shows input image with overlayed superpixel boundaries, followed by recovered pose (upper arms in red, lower in blue, upper legs in green, lower in purple), and segmentation associated with each half-limb.