Combining Discriminative Appearance and Segmentation Cues for Articulated Human Pose Estimation

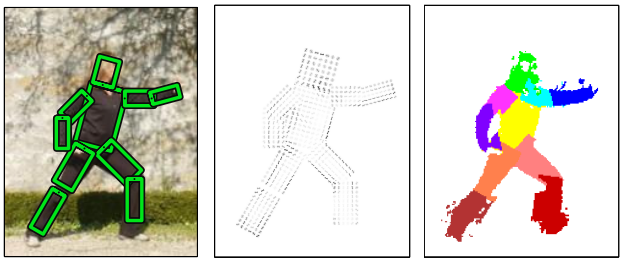
Sam Johnson and Mark Everingham
School of Computing
University of Leeds
{mat4saj|m.everingham}@leeds.ac.uk

Abstract

We address the problem of articulated 2-D human pose estimation in unconstrained natural images. In previous work the Pictorial Structure Model approach has proven particularly successful, and is appealing because of its moderate computational cost. However, the accuracy of resulting pose estimates has been limited by the use of simple representations of limb appearance. We propose strong discriminatively trained limb detectors combining gradient and color segmentation cues. Our main contribution is a novel method for capturing coherent appearance properties of a limb using efficient color segmentation applied to every limb hypothesis during inference. The approach gives state-of-the-art results improving significantly on the "iterative image parsing" method, and shows significant promise for combination with other models of pose and appearance.

1. Introduction

Articulated human pose estimation is an important and challenging goal for computer vision, with numerous applications including pedestrian detection, markerless motion capture, and movie indexing. Robust and accurate pose estimation is a prerequisite for detailed understanding of human activity in video. Numerous approaches using both 2-D and 3-D body models have been proposed. In this work we address the task of 2-D pose estimation, adopting the view that 3-D estimation is best achieved by "lifting" [9] 2-D pose estimates obtained across multiple frames, where implicit ambiguities in the projection can be resolved. The Pictorial Structure Model (PSM) proposed by Felzenswalb & Hutenlocher [7] has proven an effective framework for pose estimation, with appeal in terms of simplicity and supporting computationally efficient maximum *a posteriori* (MAP) estimation and sampling, and we build on this approach in our work. The PSM comprises two components: (i) a statistical prior model of allowable configuration of limb "parts";



(a) Estimated pose (b) Gradient cues (c) Segmentation cues

Figure 1. Overview of the method. (a) The 2-D pictorial structure model is used to estimate articulated pose. Estimation of the pose is driven by two discriminatively learnt limb appearance terms: (b) gradient cues, capturing limb edges and shading; (c) local segmentation cues, capturing limb shape and coherent color properties.

(ii) appearance terms measuring the likelihood of an individual limb being present at a given image location. Our work focuses on the second component, proposing strong discriminatively trained limb detectors capturing both gradient and color segmentation cues without compromising global optimality of the MAP pose estimates or sacrificing computational efficiency.

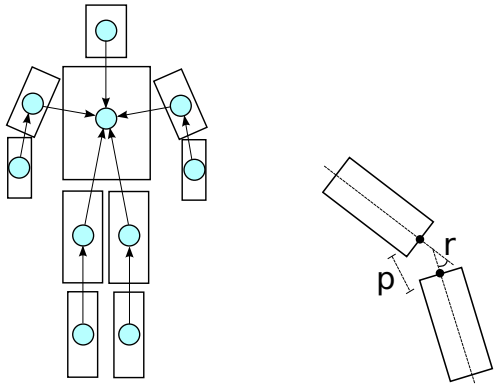
The original PSM formulation required off-line learning of appearance models for each limb, consisting of simple mean color and a hand-built "bar detector", preventing its application to general images, and limiting robustness. Using the model as a sampler, rather than for MAP inference, it was shown [7] that whole-body silhouette appearance could be used to improve on the MAP estimates, however requiring a known background to support foreground/background segmentation. Ramanan [13] proposed to learn person-specific color models for each limb in a tracking context, by detecting a canonical lateral walking pose ("scissor" shape) using a constrained configuration prior and relying on generic bar detectors to initially detect the limbs. This approach was extended to single images by taking an iterative approach: first the marginal pose

2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops
978-1-4244-4441-0/09/$25.00 ©2009 IEEE

is estimated using generic edge-based limb detectors, then color models for each limb are estimated and the pose re-estimated using both edge and color terms. The method is shown [11] to improve accuracy over edge-based limb detectors alone, but relies on reasonable pose estimates in the first iteration, and is computationally expensive. Ferrari *et al*. [8] have applied this scheme to upper-body pose estimation, incorporating a sliding-window upper-body detector to generate initial hypotheses and produce an approximate foreground/background segmentation, reducing the search space and removing some "clutter" edges in the background. They show improved computational efficiency and accuracy in the constrained domain of upper-body pose estimation, but rely on accurate initial upper-body detections, making the approach inapplicable to general images.

Several authors have investigated incorporation of stronger models of limb appearance in the PSM approach, aiming to overcome the limitations of simple hand-built bar models. Ronfard *et al*. [15] trained limb detectors using support vector machine (SVM) classifiers and an image descriptor based on pixel-wise Gaussian derivatives. The effectiveness of the method was modest, most likely limited by the lack of invariance in the image descriptor. Ramanan [11] used discriminatively-learned chamfer templates [6, 17] matched to Canny edges. Buehler *et al*. [3] address upper-body pose estimation in a PSM sampling framework [7], and incorporate Histogram of Oriented Gradient (HOG) [4] templates for the forearms. However, the approach uses simple template matching to score limb hypotheses, and requires *person-specific* color models and HOG templates to be defined, preventing application to general images. Tran & Forsyth [18] have also exploited HOG descriptors of limb appearance, addressing the task of pedestrian detection via partial (lower-body) pose estimation.

In our work we propose two extensions to the PSM approach, providing stronger appearance cues for limb detection. First, we show that the HOG descriptor [4], previously applied to pedestrian detection [4, 18] can be used to effectively capture appearance of the limbs in terms of edge and shading properties. Second, we show how to incorporate color segmentation cues in the limb detectors in a computationally efficient manner. Some previous work [10] has used bottom-up "super-pixel" segmentation to provide cues for limb locations, but gives only modest results, and requires expense combinatorial search over segments to hypothesize limb candidates. Our method incorporates segmentation cues without bottom-up segmentation, and with high computational efficiency. The approach is inspired by the work of Ramanan [12] which uses segmentation to validate object detections output by a sliding-window detector. The essence of the method is: (i) detect candidate window; (ii) segment the window; (iii) classify the segmentation as "object-like" or "non-object-like". By performing window-local segmentation the approach can capture local coherence of object color and object shape, not captured by gradient-based descriptors such as HOG, while not requiring a category-level color model. In Ramanan's work [12] verification of segmentation is applied only to a subset of detections because the segmentation method is computationally expensive. In this work we show how segmentation can be estimated and verified for *every* window (hypothesized limb location) at modest computational expense. This allows segmentation cues to be incorporated in the PSM without losing the guarantee of global optimality of the MAP solution. We show that combining gradient and segmentation cues leads to significantly improved pose estimates. The approach is general and could be incorporated in other approaches to both human pose estimation and part-based object detection.



(a) Graphical model (b) Edge parameters

Figure 2. Pictorial structure model. (a) The human body is modeled as a graph with 10 nodes corresponding to limbs. Edges represent pairwise joint probability distributions on limb position and orientation. The tree-structured graph enables efficient inference; (b) Compatibility between connected parts is represented by a distribution over relative position and orientation, learnt from annotated training data.

Outline. Section 2 briefly reviews the PSM approach. In section 3.1 the HOG-based appearance terms are described. Section 3.2 describes our proposed method for incorporating color segmentation cues. Section 4 presents results on the challenging "Iterative Image Parsing" (IIP) dataset [11], and Section 5 offers conclusions and proposes directions for future work.

2. Pictorial Structure Model

We begin by reviewing the pictorial structure model [7] since our method extends this work.

Model. As shown in Figure 2(a) The human body is represented by a graphical model with nodes corresponding to

the major "parts" or limbs (torso, head, upper/lower arm, etc.) A 2-D pose is parametrized by $L = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n\}$ where the subscript indexes the parts. The state of each part has parameters $\mathbf{l}_i = \langle x_i, y_i, \theta_i \rangle$, specifying its 2-D image coordinates and orientation.

Prior. In order to constrain the model to limb configurations representing plausible human poses, a prior $P(L)$ is defined over the state of the parts L. The joint distribution over all parts is factorized into a product of pairwise conditional distributions according to the set of directed edges E in the model (Figure 2(a)):

$$p(L) = p(\mathbf{l}_1) \prod_{(\mathbf{l}_i, \mathbf{l}_j) \in E} p(\mathbf{l}_i | \mathbf{l}_j) \tag{1}$$

where $\mathbf{l}_1$ represents some "root" part (the torso). The set of edges E is naturally defined as a kinematic tree. As discussed below, restriction of the graph to a tree structure leads to computationally efficient inference.

The pairwise conditional probabilities $p(\mathbf{l}_i|\mathbf{l}_j)$ model the compatibility of connected parts in terms of their relative position and orientation (Figure 2(b)). The distributions are modeled as Gaussian in a transformed space [7, 14]:

$$p(\mathbf{l}_i|\mathbf{l}_j) = N(T_{ji}(\mathbf{l}_i)|T_{ij}(\mathbf{l}_j), \Sigma^{ji}) \tag{2}$$

where Σ^{ji} is the covariance between the parts. The function $T(\cdot)$ transforms the configuration of each part to a shared coordinate system:

$$T_{ji}(\mathbf{l}_i) = \begin{pmatrix} x_i + d_x^{ji}\cos\theta_i - d_y^{ji}\sin\theta_i \\ y_i + d_x^{ji}\sin\theta_i + d_y^{ji}\cos\theta_i \\ \theta_i + \tilde{\theta}_{ji} \end{pmatrix} \tag{3}$$

where $d^{ji} = (d_x^{ji}, d_y^{ji})^T$ denotes the mean relative position between part i and its parent part j, and $\tilde{\theta}_{ji}$ is the angle between them.

This model corresponds to a "loose limbed" approach [16] in that the end points of the limbs are not constrained to be equal, but are allowed to displace according to the pairwise covariance. This improves robustness over a strict kinematic model (where *e.g.* upper- and lower-arm always meet precisely at the elbow), since the model is at best an approximation of the human skeletal structure.

Appearance terms. In order to estimate the pose (configuration of parts) for a given image I, the model is augmented with a term measuring the likelihood of the image given an estimated pose L:

$$p(L|I) \propto p(I|L)p(L) \tag{4}$$

The appearance term $p(I|L)$ is simplified by assuming independence between all parts, giving

$$p(I|L) = \prod_{i}^{n} p(I_i|\mathbf{l}_i) \tag{5}$$

where $p(I_i|\mathbf{l}_i)$ measures the image evidence for part i at location $\mathbf{l}_i$. As described in Sec. 3.3 we define these terms using discriminatively-trained limb detectors.

Inference. By restricting the graphical model to a tree structure, inference is computationally efficient [7]. In our work we estimate pose as the MAP pose $\arg\max_L p(L|I)$. The MAP pose is found using a dynamic programming algorithm and exploiting the generalized distance transform [7] with complexity $O(np)$ where n is the number of possible limb locations (position and orientation) and p is the number of parts (10 in our model). The PSM also supports efficient computation of marginals for each limb [14] and sampling [7, 3], and our proposed appearance terms can also be applied to these approaches.

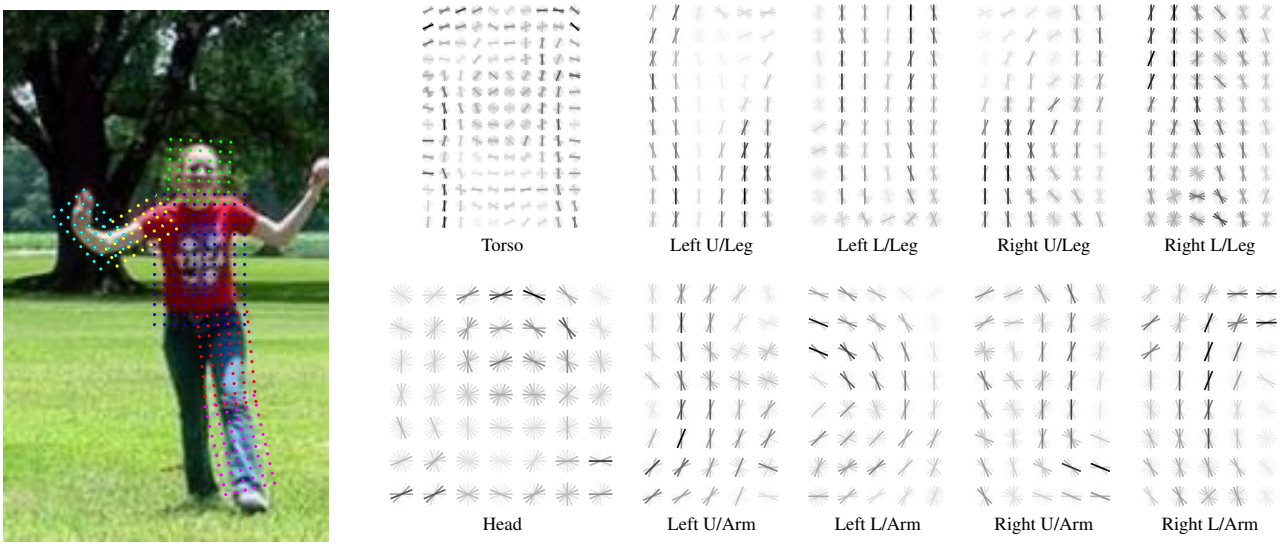
3. Limb detection

This section describes our proposed methods for limb detection, which are used as the appearance terms in Eqn. 5. As noted in the introduction we model two aspects of a limb's appearance: (i) edge and shading properties based on image gradients (Section 3.1); (ii) shape derived from color segmentation cues (Section 3.2). For each appearance descriptor a discriminatively-trained classifier (Section 3.3) is used to estimate the likelihood that a limb is present at a particular location.

3.1. Gradient descriptor

To capture general appearance of a limb we use a form of the Histogram of Oriented Gradients (HOG) descriptor [4]. Figure 3(b) visualizes the HOG descriptors extracting for each limb. HOG represents an image region as a joint histogram over quantized gradient orientation and spatial position. This descriptor has recently been applied successfully to fore-arm [3] and leg [18] localization. The intuition of the HOG descriptor is to capture local appearance such as edges and shading while incorporating a controlled level of invariance to local deformation, *e.g.* variation in limb shape, by quantization and spatial pooling of the image gradients.

Given a hypothesized limb position and orientation, a HOG descriptor is computed about the center of the limb oriented along its medial axis. The image gradient $\langle \frac{d}{dx}I, \frac{d}{dy}I \rangle$ is estimated by convolution (we use simple $[-1, 0, 1]$ filters [4]) and the magnitude and orientation are computed. The orientation is quantized into a fixed number of discrete "bins" and the gradients are then pooled into a set of spatial "cells" by recording a histogram of quantized orientation over the image region of the spatial bin, with each orientation weighted by the corresponding gradient magnitude. As in previous work [4] we used unsigned gradient orientation (dark/light vs. light/dark transitions are considered equal) since we do not know *a priori* whether limbs will appear light or dark against the background.



(a) HOG locations

(b) HOG descriptors

Figure 3. HOG descriptors for limb detection. (a) Colored blocks indicate the location and size of the HOG descriptors in ground truth position for the image shown (corresponding descriptors are extracted for the other arm and leg, omitted for clarity). Each dot marks the center of a spatial bin; (b) The extracted descriptors represent the image patch by histograms of oriented gradient magnitude over a grid of spatial cells. Bars represent the magnitude in each orientation bin, with darker bars indicating stronger gradient.

Linear interpolation is applied between neighboring orientation bins [4] to avoid quantization artifacts, so each gradient vector contributes to two orientation bins. The spatial cells are arranged on a rectangular grid (Figure 3) and Gaussian weighting is used when pooling orientations into each cell [4] to avoid quantization artifacts. We do not apply the "block" scheme proposed by Dalal & Triggs [4] (which renormalizes cells over larger spatial neighborhoods), as this is relevant only where there is significant local variation in contrast.

As shown in Figure 3 the HOG descriptors extracted for each limb capture the edges present at the boundary of the limb, characteristic shading across the limb, and internal features such as the hairline. In the case of strongly textured clothing such as the T-shirt worn here, the descriptor also represents this texture, which is undesirable for generic limb detection. As described in Section 3.3, we learn a classifier from training images which can suppress such unstable internal features.

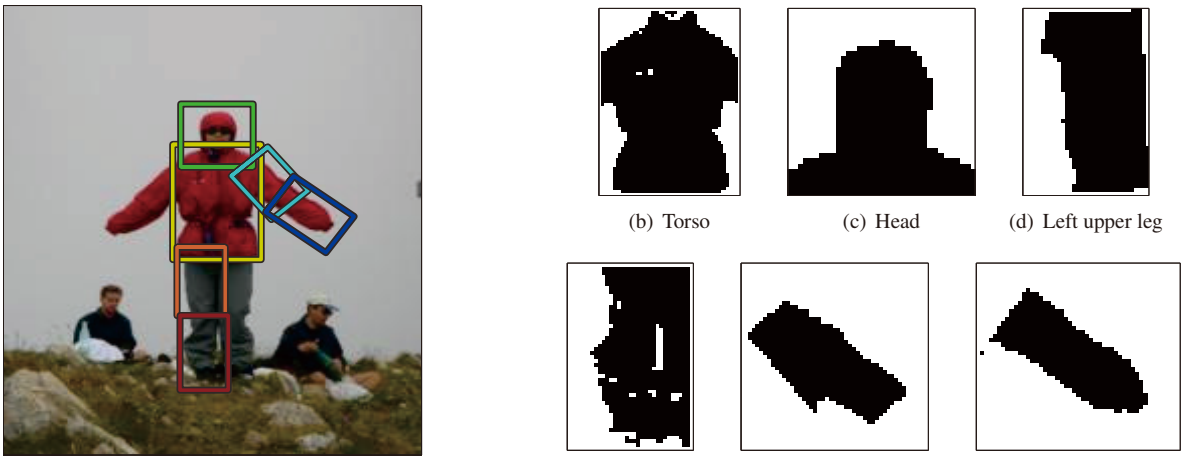
Efficient implementation. To apply the HOG descriptor scheme within the PSM approach, HOG descriptors must be computed exhaustively for all possible limb locations. While seemingly onerous, in practice this involves only moderate computational expense: (i) gradients are computed using small (3×1) 1-D kernels; (ii) conversion of gradient vectors to orientation is performed using a lookup table, avoiding costly $\arctan$ computations; (iii) Spatial binning is accomplished by separable convolution with 1-D Gaussian filters – four 1-D convolutions are required per orientation bin to implement linear interpolation. Given this pre-computation, extracting the descriptor at a given location requires only mn memory lookups, where m is the number of orientation bins and n is the number of spatial cells. As described in Section 3.3 we use a linear classifier to compute the confidence that a limb is present at a given location, requiring just $2mn$ arithmetic operations.

3.2. Segmentation cue

As reported in Section 4 the HOG descriptor proves an effective means of capturing limb appearance, but it does not exploit color cues or capture the coherent properties of a limb *e.g.* uniform color. We therefore propose to augment the PSM with an appearance term based on local segmentation of the image about each hypothesized limb location.

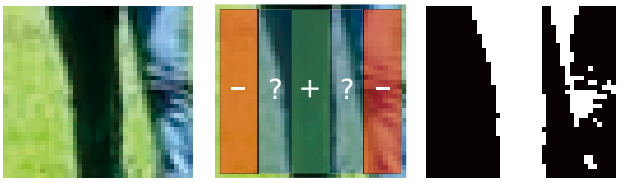
Figure 5 illustrates the proposed scheme for extraction of segmentation cues. As noted, the method is inspired by the work of Ramanan [12] on verifying object detection hypotheses by segmentation. The essence of the method is to perform a local segmentation of the image region around a hypothesized limb location, and "score" the resulting segmentation as limb-like or non-limb-like.

Given a hypothesized limb location, two image regions R_f and R_b relative to the limb are predicted to be respectively foreground (limb) and background (green and red in Figure 5(b)). The appearance of the foreground and background regions is modeled, and pixels are assigned to foreground or background (Figure 5(c)) by the nearest model. The resulting segmentation is "scored" by a classi-



(a) Segmentation descriptor regions. (b) Torso (c) Head (d) Left upper leg (e) Left lower leg (f) Right upper arm (g) Right lower arm

Figure 4. Segmentation descriptors for limb detection. (a) Colored regions indicate the location and size of the segmentation descriptors in ground truth regions for the image shown (corresponding descriptors are extracted for the other arm and leg, omitted for clarity). (b)-(g) The descriptors extracted from the colored regions in (a) using the segmentation method described in Section 3.2.



(a) Image region (b) Hypothesis (c) Segmentation

Figure 5. Segmentation cue. Given an image region (a), an initial hypothesized labeling of the image as foreground, background, or unknown is made (b). Color models for foreground and background are estimated and the region is segmented. The resulting segmentation (c) is scored by a limb/non-limb classifier.

fier trained on automatic segmentations extracted from the training images (Section 3.3).

In the work of Ramanan [12], color histograms are used to model foreground/background color distributions, and a graph-cut method [2] to obtain a smooth segmentation. As for the HOG descriptors (Section 3.1), segmentation descriptors need to be extracted for *every* possible limb location. This requires a significantly more efficient method than that proposed by Ramanan. We obtain such efficiency by assuming a simple model for the distribution of colors in foreground and background regions, and making a foreground/background decision for each pixel independently.

For each region R_c where $c \in \{f|b\}$ the distribution of RGB pixel values $p(\mathbf{x}|c)$ is assumed to be Gaussian with mean $\boldsymbol{\mu}_c$ and covariance $\Sigma_c = \sigma^2 \mathtt{I}$. Note that we assume *equal* scalar covariance for both foreground and background regions. A segmentation is then obtained by assigning each pixel i a label λ_i of foreground/background:

$$\lambda_i = \arg \max_{c \in \{f|b\}} p(\mathbf{x}_i|c)P(c) \tag{6}$$

Because of the Gaussian model employed, this function takes a particularly simple form, namely a linear discriminant [5]:

$$\lambda_i = \begin{cases} f & : \quad \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) \geq 0 \\ b & : \quad otherwise \end{cases} \tag{7}$$

where

$$\mathbf{w} = \boldsymbol{\mu}_f - \boldsymbol{\mu}_b \tag{8}$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_f + \boldsymbol{\mu}_b) - \frac{\sigma^2}{||\mathbf{w}||^2} \ln \frac{P(f)}{P(b)} \mathbf{w} \tag{9}$$

We can further assume that the prior probabilities $P(c)$ are equal, so that the second term in Eqn. 9 can be dropped. Using this formulation, to obtain a local segmentation around a particular limb hypothesis we need only the sum and difference of mean RGB values in the hypothesized foreground and background regions (Eqn. 8 & 9).

As shown in Figure 4 the extracted segmentation descriptors capture both the shape of the body part and the coherent color properties. The close proximity of the legs in the image shown leads to both left and right limbs being captured in the lower left leg descriptor. As described in Section 3.3, we learn a classifier from training images which can suppress the incorrectly segmented parts of each descriptor.

Efficient implementation. Estimation of the local foreground and background means can be performed efficiently for *every* limb position using the integral image [19] since the hypothesized regions are axis-aligned rectangles (Figure 5(b)). A total of just 27 addition/subtraction operations are required per limb position to compute these parameters. Segmenting an image region then takes 6 arithmetic operations per pixel in the region.

Method	Torso	Upper Leg		Lower Leg		Upper Arm		Forearm		Head	Total
IIP [11], 1st	39.5	21.4	20.0	23.9	17.5	13.6	11.7	12.1	11.2	21.4	19.2
IIP [11], 2nd	52.1	30.2	31.7	27.8	30.2	17.0	18.0	14.6	12.6	37.5	27.2
PSR [1]	81.4	67.3	59.0	63.9	46.3	47.3	47.8	31.2	32.1	75.6	55.2
HOG	73.2	62.0	55.1	54.2	50.2	51.2	44.4	36.6	28.3	62.4	51.8
HOG+Seg	**77.6**	**64.9**	**58.1**	**57.6**	**52.2**	**55.6**	**50.7**	**42.4**	**36.1**	**68.8**	**56.4**

Table 1. Comparison of body part localization rates (in percentages) for recent approaches and ours on the "Iterative Image Parsing" dataset [11]. This dataset contains 205 test images with 10 parts each.

3.3. Classification

For each limb we train separate linear classifiers for the HOG and segmentation cues respectively. For the segmentation cue the feature vector is the binary-valued vectorized segmented image region. Linear classifiers have proven suitable for such features in previous work [4, 12] – the high dimensionality of the feature vectors to some extent obviates the need for kernel methods. The elements of the weight vector serve to emphasize stable features and suppress unstable features such as internal texture (HOG) or pixels which are often shared by limbs (segmentation). The classifiers are trained using logistic regression with L2 regularization of the weight vector, with the regularization parameter determined by cross-validation on the training set. The classifier output can be interpreted as the probability of a limb being present at the hypothesized location, conditioned on the underlying image patch, and is substituted directly into Eqn. 5, assuming independence between the HOG and segmentation descriptors. For the HOG classifier we found it profitable to use bootstrapping [4], collecting false positives in multiple rounds and re-training. For the segmentation classifier, bootstrapping did not improve performance.

4. Experimental results

We evaluated our proposed method on the "Iterative Image Parsing" (IIP) dataset [11]. This dataset is challenging due to the high variability in poses, and the presence of significant background clutter. There are 305 annotated images in total, of which the first 100 are defined for use as training data. From the training set alone we train the limb detectors and estimate parameters of the configuration prior in a maximum likelihood fashion. For evaluation we adopt the criteria proposed by Ferrari *et al*. [8] and adopted in work concurrent to ours by Andriluka *et al*. [1]. For a given pose estimate a part (limb) is considered correctly localized if its predicted endpoints are within 50% of the part length from the corresponding ground truth endpoints.

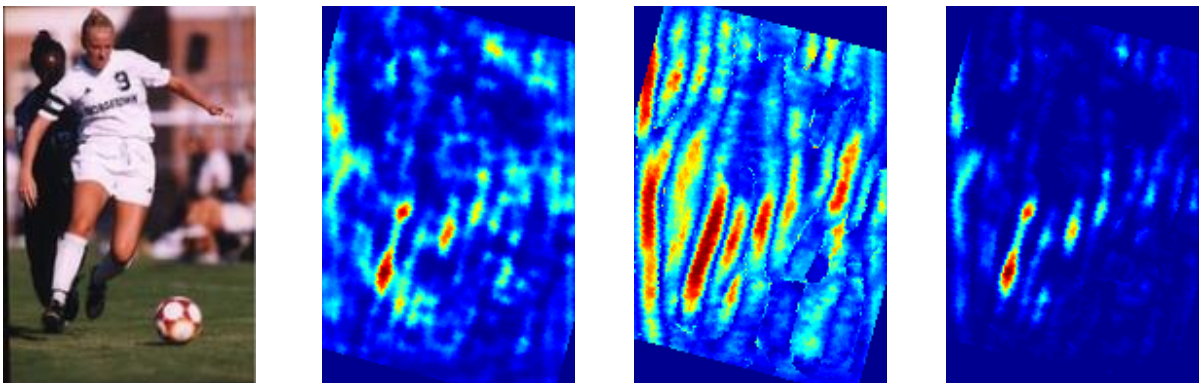
Examples of pose estimates from our method are shown in Figure 7, annotated with the number of correctly localized parts as determined by the defined criteria. As can be seen, qualitatively accurate pose estimates can be obtained over a wide range of pose and imaging conditions.

Table 1 reports quantitative results of the proposed method and comparison to recent methods representing the state-of-the-art on this dataset. "HOG" denotes our method using the HOG descriptor alone, and "HOG+Seg" our method combining both gradient and segmentation cues. "IIP" denotes the Iterative Image Parsing approach of Ramanan [11] – we report results of this method using edge cues alone (IIP, 1st) and using edge features plus the color models learnt from the 1st edge-only iteration (IIP, 2nd). The method denoted "PSR" ("Pictorial Structures Revisited") is work concurrent with ours, presented at CVPR 2009 [1].

Comparison to IIP. Compared to the IIP method [11] our method gives significantly more accurate pose estimates. Using the HOG descriptor alone the percentage of correct localizations is increased for all body parts, both for IIP with and without learnt color models. We conclude that the invariance built into the HOG descriptor is effective compared to the chamfer templates used in the IIP method [11]. Combining HOG and segmentation cues improves the accuracy further, improving significantly on the IIP use of color models (IIP, 2nd). Overall, our method gives 56.4% accurate part localizations compared to 27.2% using the full IIP method.

Combination of cues. Figure 6 offers some insight into the source of improved results obtained by combining gradient and color segmentation features. Shown are the individual responses of the HOG and segmentation classifiers, and the combined response. It can be seen that both detectors give a strong response in the correct location but also give false responses in many areas of clutter, and other areas of the body. Crucially the responses of the two detectors prove complementary, with false responses lying in differing image regions. When the outputs are combined these false responses are successfully suppressed while the correct responses are enhanced.

Comparison to PSR [1]. As shown in Table 1, our overall results combining HOG and segmentation cues are slightly better than those of the recent "PSR" method (56.4% vs.



(a) Test image (b) HOG (c) Segmentation (d) Combined

Figure 6. Individual and combined detector responses for the left lower leg on the image shown in (a). The HOG detector gives a strong response in the correct location but also responds noisily in other areas which will be amplified across multiple orientations. The response from the segmentation detector shows similar peaks in the correct location, along with high responses in other areas. Combining the two detector outputs effectively removes the majority of the false responses while amplifying the correct peak, leaving a correctly localized lower left leg. (Note: Red indicates high probability through to blue indicating low, figure best viewed in color.)

55.2%). Considering the localization accuracy of individual limbs (Table 1), our method improves accuracy notably for the arms and one of the lower legs, suggesting that our pose estimates are more accurate, while the "detection" abilities of PSR based on torso and head (the latter being particularly strong) are better. The PSR approach uses boosted classifiers operating on shape context descriptors, somewhat similar to the HOG descriptors we employ. Without the use of segmentation cues the PSR method slightly outperforms ours (55.2% vs. 51.8%). This may be due to differences in the appearance descriptor, use of a boosted classifier, or the use of max marginal inference. A salient observation is that the PSR method does not exploit any color information, relying on edge features alone. As shown here, incorporating color information improves results, and this suggests that even better results could be obtained by incorporating our proposed color segmentation cues into the PSR framework.

Failure modes. Some cases of incorrectly localized limbs can be seen in Figure 7, and further examples of failures are shown in Figure 8. A number of images (Figures 7(a), 7(h), 8(d), and 8(e)) exhibit "classic" pictorial structure model failures where both left and right limbs are positioned in the same location. This is caused by the over-counting of image evidence due to the assumption of independence between part appearance (Eqn. 5). Figure 8(d) shows a failure to account for self-occlusion, with the right arm being detected incorrectly as lying over the head. Although the lower part of the arm is visible, the upper arm is completely occluded by the torso. The image in Figure 8(a) shows a failure due to an extreme pose. The torso is oriented towards the camera position such that it is significantly foreshortened, not accounted for in our current approach. Also the lower left leg is at a particularly unusual angle which has low probability according to the learnt prior model. Extreme failures of the proposed method are shown in Figures 8(b) and 8(c). Here the large amount of background clutter has led to the MAP pose being in completely the wrong location. Figure 8(b) in particular shows how regions with strong edges and highly plausible limb shapes can lead to incorrect pose estimates.

5. Conclusions and future work

We have proposed extensions to the established PSM approach to articulated pose estimation, comprising improved discriminatively-learnt models for part appearance which exploit both gradient and color segmentation cues. We showed significantly improved results compared to the "Iterative Image Parsing" approach. The method is computationally efficient and can be incorporated in other models for pose estimation and part-based object localization. We hope in particular that combination with the method of Andriluka *et al*. [1] would further improve results.

Limitations of the proposed method include the simple color model, which results in imprecise segmentation cues. We are working on methods to incorporate more comprehensive color models without compromising computational efficiency. Many of the failures of our method are due to limitations of the PSM approach itself *e.g*. no modeling of self-occlusion. In future work we will combine the method with sampling-based global verification, using the improved prediction of poses offered by our method to obtain more efficient sampling.

Acknowledgements. We are grateful for financial support for this work from an EPSRC doctoral training grant (Sam Johnson) and an RCUK Academic Fellowship (Mark Everingham).

(a) 9/10 (b) 8/10 (c) 9/10 (d) 9/10 (e) 10/10

(f) 8/10 (g) 7/10 (h) 6/10 (i) 9/10 (j) 6/10

Figure 7. Estimated poses on images from the "Iterative Image Parsing" dataset [11] using our proposed method. The numbers below each image show the number of correctly localized body parts.

(a) 1/10 (b) 0/10 (c) 0/10 (d) 4/10 (e) 5/10

Figure 8. Failure modes of our proposed method on images from the "Iterative Image Parsing" dataset [11]. The estimated poses in (d) and (e) exhibit over-counting problems common to Pictorial Structure Model approaches, while images (a), (b), and (c) show total failures due to extreme pose and background clutter. The numbers below each image show the number of correctly localized body parts.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. CVPR*, 2009.

[2] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *IJCV*, pages 105–112, 2001.

[3] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proc. BMVC*, 2008.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley-Interscience, 2001.

[6] P. Felzenszwalb. Learning models for object recognition. In *Proc. CVPR*, 2001.

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

[8] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proc. CVPR*, 2008.

[9] H. Lee and Z. Chen. Determination of 3D human body postures from a single view. *CVGIP*, pages 148–168, 1985.

[10] G. Mori. Guiding model search using segmentation. In *Proc. ICCV*, pages 1417–1423, 2005.

[11] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006.

[12] D. Ramanan. Using segmentation to verify object hypotheses. In *Proc. CVPR*, 2007.

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

[14] D. Ramanan and C. Sminchisescu. Training deformable models for localization. In *Proc. CVPR*, 2006.

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

[16] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *Proc. CVPR*, 2004.

[17] A. Thayananthan, B. Stenger, P. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *Proc. CVPR*, pages 127–133, 2003.

[18] D. Tran and D. Forsyth. Configuration estimates improve pedestrian finding. In *NIPS*, pages 1529–1536, 2007.

[19] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 2001.