Shape-based Pedestrian Parsing

Yihang Bo
School of Computer and Information Technology
Beijing Jiaotong University, China
yihang.bo@gmail.com

Charless C. Fowlkes
Department of Computer Science
University of California, Irvine
fowlkes@ics.uci.edu

Abstract

We describe a simple model for parsing pedestrians based on shape. Our model assembles candidate parts from an oversegmentation of the image and matches them to a library of exemplars. Our matching uses a hierarchical decomposition into a variable number of parts and computes scores on partial matchings in order to prune the search space of candidate segment. Simple constraints enforce consistent layout of parts. Because our model is shape-based, it generalizes well. We use exemplars from a controlled dataset of poses but achieve good test performance on unconstrained images of pedestrians in street scenes. We demonstrate results of parsing detections returned from a standard scanning-window pedestrian detector and use the resulting parse to perform viewpoint prediction and detection re-scoring.

1. Introduction

A fundamental problem in scene understanding is combining top-down information provided by object detection and recognition with information on object localization provided by bottom-up segmentation. There has been a variety of proposals in the last 10 years for combining segmentation and recognition [32, 16, 29, 2, 15, 17], but perhaps the simplest approach is a feed-forward model in which candidate objects are first detected and then each object is segmented using an object specific model. In order to provide a mechanism for feedback, the resulting segmentations can be used to either rescore detections (see e.g., [23]) and/or combined to yield a consistent interpretation of the entire scene (see e.g., [31]).

In this paper we focus on the problem of segmenting human figures. Segmenting humans poses a particularly good testbed for object specific segmentation since human figures are highly articulated and vary widely in appearance due to clothing. One proposal for bridging the gap between bottom-up segmentation and this high-level task of segmenting a heterogeneous, articulated object, is to search over assemblies of small, bottom-up segments (also known as superpixels) in order to find the human figure [20, 28].

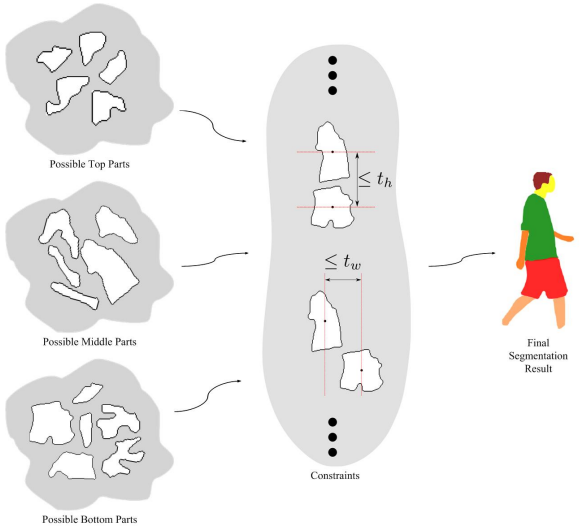

Figure 1. Overview of processing. A large pool of candidate segments are generated by directed aggregation of superpixels. Candidate regions are scored based on shape similarity to a database of shape exemplars. Simple constraints between parts enforce consistent layout (e.g. upper-body must appear above lower-body) in the labeling of regions. Assemblies with variable numbers of parts are scored using a simplified hierarchical model of appearance.

The fundamental challenges to be solved in such an approach are dealing with the combinatorial complexity of assembling a large number of potential superpixels and choosing appropriate scoring functions for evaluating the shape and appearance of a given assembly. We tackle both of these issues using a hierarchical description of segment shapes. This allows us to model both constraints between segments (as are captured by standard approaches to recovering articulated pose [9, 24, 11]) as well as correlations in appearance between parts and sub-parts.

Hierarchical composition is an appealing approach and has been explored in several vision contexts primarily for recognition (see e.g., [13, 14]). Our model is closely related

to the AND/OR graph framework of [7, 33] which was used to parse human body poses in [34]. The primary distinction with the work on AND/OR graphs is that our model is built on segments (rather than edges) and the likelihood of our intermediate parts involve appearance features in addition to part-part and part-subpart interactions. We also use hard constraints in the pairwise interactions which allows for simple pruning of the search space.

Unlike other approaches to image understanding that attempt to perform a "semantic parse" of the image based on high-level concepts, we focus on performing a mid-level "syntactic parsing" of an image into visual components. For example, rather than attempting to divide a human arm into upper and lower arm (kinematic parts which may not be readily apparent in an image) we divide it into clothed and unclothed visual components. This is desirable since the mapping from semantic parts to visual appearance is certainly not one-to-one. Defining an intermediate structure of visual components[1] allows us to focus our modeling on visual features and defer semantic reasoning to a later stage of processing.

Figure 1 gives a birds-eye view of the proposed algorithm. Pools of candidate parts are generated from bottom-up segmentation and each segment scored based on shape and appearance. These segments are assembled in a bottom-up parse tree which enforces constraints among parts at a given level of the tree and between parents and children. The final result is a hierarchical segmentation of the image into a variable number of parts and background. In the following sections we describe details of the parsing constraints, part segment generation, scoring and assembly procedures. We then describe an experimental quantification of segmentation accuracy on a set of scenes of pedestrians in the wild. We also describe how to leverage the parsing model to perform viewpoint classification and detector rescoring.

2. A Compositional Model for Pedestrians

Figure 2 shows the structure of our parsing model. We decompose the pedestrian into three major parts (head, upper-body, lower-body). Each of these components in turn may consist of sub-parts (hair,face,upper-clothes, arms,lower-clothes,legs). We refer to these components all generically as parts. Unlike the large body of work on human pose analysis, this decomposition into parts is not kinematic. Instead our parts are appearance based. Face and hair are separate parts because for viewpoints (and individuals) these parts are photometrically distinct. Likewise, the division into upper-clothes and arms only applies in those cases where the two are distinct (i.e. short shirt sleeves).

[1]Throughout the rest of the paper we abuse terminology and refer to the visual components in our model generically as "parts" but they should be thought of as visual (parts of the image) rather than semantic (parts of the object).

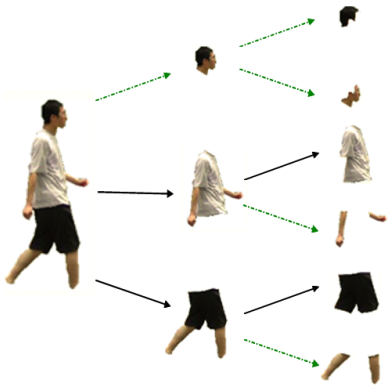

Figure 2. Our hierarchical model for pedestrians includes *head*, *upper-body* and *lower-body*. Each of these regions in turn may be further subdivided depending on the pose of the pedestrian into *hair, face, upper-clothes, arms, lower-clothes* and *legs*. The dashed arrows show "optional" production rules which may not apply to a given image. For example, from a back view only the hair is visible but not the face.

In parsing a candidate pedestrian bounding box, we enforce that each part appears only once and that a part may only appear if its parent part also appears. In addition, we impose the following set of hard geometric constraints on the segments corresponding to parts of the model.

- head should appear above upper-body which should appear above lower-body
- x coordinates of head, upper- and lower-body should be within 25% of the overall bounding box width
- head, upper body and lower body should not overlap by too much or be separated by too great a distance
- arms should be adjacent to, or overlap the upper-clothes
- legs should be adjacent to and below lower-clothes
- face and hair should be adjacent

The thresholds for these constraints are set from training data. In our final parsing, we will only consider labelings of segments which satisfy these constraints.

3. Generating Candidate Part Segments

We derive candidate parts from the superpixels produced by the gPb-OWT-UCM segmentation code of [1]. This segmentation algorithm achieves state-of-the art performance on general purpose segmentation benchmarks and a fast implementation is available which exploits computation on the

GPU to segment an image in a few seconds [6]. The algorithm returns as output a hierarchical segmentation represented by a weighted boundary map termed an Ultrametric Contour Map (UCM). Thresholding the contour map at larger values produces coarser segmentations of the image.

While the segmentation quality is quite good, we found that the hierarchy it produces is seldom aligned with the hierarchy of pedestrian parts. Furthermore, it is seldom possible to choose a threshold that results in the pedestrian or parts of the pedestrian appearing as a single segment (see UCM examples in Figure 5). Markings on clothes and shadows are often higher contrast than those boundaries separating the pedestrian from the background making simple bottom-up segmentation impossible. Instead, we produce a whole collection of segments by selecting a range of possible thresholds of the UCM as well as considering additional assemblies. This approach is grounded in a long line of work starting with Ren and Malik [25] which scores multiple bottom-up segmentation hypotheses [21, 20, 26, 19, 5, 18].

If we start from the collection of segments given by the UCM, there are a large number of candidate assemblies. Without any constraints, on N segments it quickly becomes infeasible to naively consider all 2^N combinations as candidate parts since N is on the order of 100. One route is to use a segment scoring function which decomposes over superpixels (e.g., [8, 35, 10]) in order to utilize efficient combinatorial algorithms. We take a different approach, instead defining arbitrary scoring functions for each part of our model but use these scoring functions along with constraints between parts to direct exploration and greedily prune the space of possible assemblies.

4. Scoring Candidate Part Segments

For each candidate segment, we would like to score the likelihood that the segment is part of a pedestrian. Let L_i be the label of segment i where $L_i \in \{head, torso, legs, ..., pedestrian, background\}$. We build an independent model for the shape and appearance of each of these parts.

4.1. Segment Shape and Appearance

Let X_i be the observed shape and appearance features for a given candidate segment i. For modeling parts we use a shape feature built from a spatial histogram of the segment edge orientations (similar to [12]). This shape feature is computed for two different coordinate frames: a relative coordinate frame which is placed tightly around the segment, and an "absolute" coordinate frame which is computed relative to the whole pedestrian. In each case, the bounding box is grided up into 11 by 11 spatial bins and 9 orientation bins. For scoring the clothed parts (upper/lower-body and upper/lower-clothes), only the shape descriptor is used. For unclothed parts (face, hair, arms and legs), we also use color and texture histograms to capture the appearance.

We estimate the label for each segment by comparing it to a library of exemplar segments. Let X_e be the feature vector associated with a exemplar e and S_k be the set of all exemplars of part type k. We model the probability as a logistic function of the feature vector distance

$$P(L_i = k|X_i) = \max_{e \in S_k} \frac{1}{1 + e^{\|X_i - X_e\|^2}}$$

4.2. Segment Area and Location

Let A_i be the area of the candidate segment. We model the area of part k by a Gaussian distribution

$$P(L_i = k|A_i) \propto e^{\frac{(A_i - \mu_k)^2}{2\sigma_k^2}}$$

To model location, we construct an average shape mask $M_k(x)$ which records the frequency with which a pixel at coordinate x relative to the pedestrian centered coordinate system belongs to the part k. Let B_i be the set of pixels associated with the candidate segment i. For a given candidate, we score the location as:

$$P(L_i = k|B_i) = \prod_{x \in B_i} M_k(x)^{(1/A_i)}$$

Finally, we assemble these three cues under a naive Bayes assumption to yield a final probability for a given label of candidate segment i.

$$P(L_i|A_i, B_i, X_i) \propto P(L_i|A_i)P(L_i|B_i)P(L_i|X_i)$$

In practice, we find that only having a single location map per part type is restrictive, particularly for upper and lower body parts which are articulated. To allow for a richer representation, we cluster the training exemplars for a given part and use a mixture model to capture these modes. This can be handled with the above notation by allowing multiple subsets of exemplars $\{S_{k_1}, S_{k_2}, \ldots\}$ for each part type k. Figure 3 shows examples of average shape masks $M_{k_m}(x)$ of the mixture components for upper and lower body parts.

5. Directed and Constrained Region Merging

In order to generate candidate segments for each part k, we start by choosing a seed segment sampled from a high probability location for the given part type. We then carry out a greedy region merging operation by successively appending additional segments to the seed. This merging operation is carried out as a best-first search which only considers adding segments that are adjacent to the candidate assembly and scores them using the features described above. The search terminates when the candidate segment area exceeds a threshold set relative to the largest area for that part

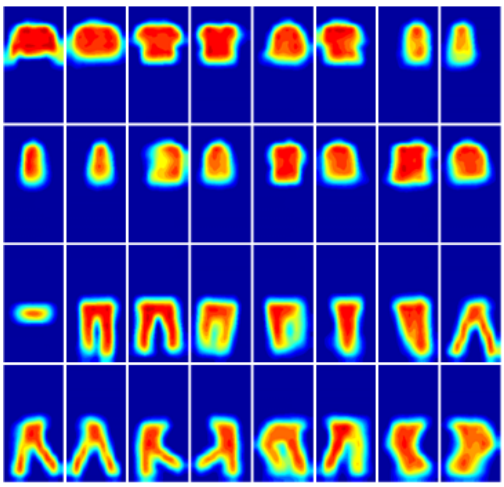

Figure 3. Average shape masks for upper- and lower-body part location used in the part likelihood computation. For each of these parts, we fit a mixture of 16 components to the training data to capture the range of locations relative to the bounding box. The left and right symmetric pairs arise naturally from the training data which includes views from all sides.

seen in the training data set. Each segment generated by this merging process is added to the pool of candidate segments for the given part type.

Since the superpixel boundaries are determined primarily based on local edge information, they may not adequately exploit segment level properties such as uniform hue. We carry out an additional refinement step for each candidate segment by using it as initialization for performing a graph-cuts based segmentation [4]. We first create a foreground mask by eroding the candidate segment support. We then use this foreground mask to estimate the foreground color and perform a graph-cut segmentation where the mask pixels are constrained to take on the foreground label and perimeter pixels are constrained to take on background. While this step usually has little effect on the overall candidate part shape, we found that it did improve the precise boundary localization of the final segmentation. After refinement, we remove near duplicate candidates from the pool.

6. Bottom-up Parsing

We assemble a parse of the image by selecting high scoring parts from the candidate pools using a bottom-up parse. Thus, we first generate candidate segments for upper/lower clothes, arms and legs, and then generate candidate segments for upper and lower-body that contain those potential arms and legs respectively, and so on. The final assembly is scored using the whole pedestrian shape model.

Since the depth of our parse tree is limited, this exhaustive approach is possible but still unwieldy. We use two

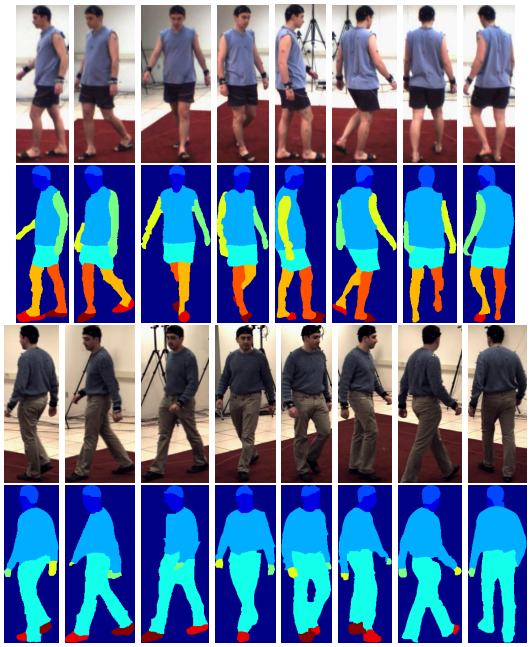

Figure 4. Training exemplars from the walking actions in the HumanEVA dataset [27]. A total of 937 exemplars from 4 individuals were segmented by hand into the 6 body parts. Exemplars were also labeled with one of 8 canonical viewpoints (front, left-front, left, ...) for use in the viewpoint estimation experiment. Note limited range of appearances present. Since our parsing algorithm is based primarily on the measurements of the shape and location of each part, it is able to generalize to the different, richer distribution of appearances in the test dataset.

techniques for pruning the space of parses. First, since our model uses hard part-part and part-subpart constraints (rather than probabilistic versions) we can enforce these constraints at every step to prune away infeasible combinations of parts. For example, we only need consider combinations of lower- and upper-body which satisfy the relative location constraints.

Second, after generating a pool of candidate parts at a given level (either individual parts assembled from segments or intermediates assembled from individual parts), we greedily prune the pool to a fixed size, keeping only the top scoring candidates. The size is selected to keep subsequent running times reasonable. In our experiments, we keep the top 20% of candidates at any given stage of parsing which typically amounts to 75-100 candidate segments for each part.

7. Experiments

For training exemplars, we use images taken from the HumanEva dataset [27]. This dataset consists of several video sequences of four different human actors performing various activities along with motion-capture and other

Part	Accuracy
Hair	44.90 %
Face	60.79 %
Upper Clothes	74.83 %
Arms	26.21 %
Lower Clothes	71.23 %
Legs	42.04 %
Background	81.05 %
Average	57.29 %

Component	Accuracy
Head	51.82 %
Upper Body	73.57 %
Lower Body	71.62 %
Background	81.05 %
Average	69.51 %

Segment	Accuracy
Foreground	73.27 %
Background	81.05 %
Average	77.17 %

Table 1. Per-pixel segmentation accuracy on Penn-Fudan pedestrian database [30]. Accuracies are compiled for individual parts, intermediate parts and the entire pedestrian. Because the parts are nested, the background accuracy is the same for all levels of the hierarchy.

auxiliary data. We use only the color images from walking sequences. For each frame we created a ground-truth segmentation using a simple interface for gluing together UCM superpixels to label the parts of our model. Since the dataset includes actors with both short and long clothes we were able to collect exemplars for all the parts in our parsing model. Figure 4 shows examples of the training images used along with the ground-truth segmentation.

From this dataset we extract 937 total exemplars. This data is split across 16 mixture components each to model the shapes and locations of the upper and lower body and upper and lower clothes.

7.1. Segmentation

To test parsing performance we use the Penn-Fudan database [30] which consists of multiple pedestrians in outdoor street scenes. We first consider the ability of the algorithm to correctly segment out each part of the pedestrian. For this purpose, we take tightly cropped image patches containing pedestrians from the dataset. Figure 5 shows example cropped pedestrian images, the UCM from which candidate parts are generated, and the final parsing results. We compare the accuracy of this parsing to ground-truth segmentations in Table 1.

Unfortunately there seems to be a lack of baselines and standard datasets for evaluating body-part segmentation accuracy. Previous work has typically scored accuracy of detection or pose estimation (e.g., [21, 30]) or whole body segmentation accuracy (e.g., [28]) for which our algorithm seems to perform comparably. The work of Bourdev et al. [3] on poselets demonstrates part segmentation but quantitative measures of accuracy on the H3D dataset were not reported. We have made our ground-truth annotations for the Penn-Fudan database available online to allow future comparisons to the results here (http://vision.ics.uci.edu/datasets/).

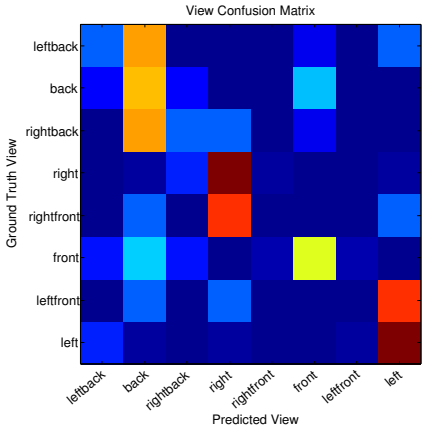

Figure 7. Confusion matrix for predicting viewpoint from the combined shapes of the head, upper and lower-body parts. We achieve an average accuracy of 37% across all viewpoints.

7.2. Viewpoint Detection

We consider the use of our shape comparison to perform viewpoint prediction on the resulting parsed pedestrians. We label each exemplar with one of 8 ground-truth viewpoints (front, left-front, left, etc). This gives a distribution over viewpoints for each exemplar cluster and for each part. Let V denote the viewpoint of a given detected pedestrian. To estimate the viewpoint implied by a given parse we consider each part i in turn and compute:

$$P(V = j|X_i) = \sum_m P(V = j|L_i = k_m, X_i)P(L_i = k_m|X_i)$$

where the sum is over all mixture components associated with part type k. To make a final determination of viewpoint for the whole pedestrian we combine the per-part predictions assuming independence and uniform prior over viewing directions

$$V^* = \arg\max \prod_i P(V = j|X_i)$$

where the product ranges over the set of segments belonging to the final parse.

Figure 7 shows the confusion matrix of predicted versus ground-truth viewpoints for the Penn-Fudan test set. On average we achieve a prediction accuracy of 37% across all viewpoints with the most confusion arising in prediction of diagonal views.

7.3. Detector Rescoring

To test our parsing model in a more realistic setting, we use the multi-resolution HOG-based pedestrian detector of [22] to first detect candidate bounding boxes and then parse

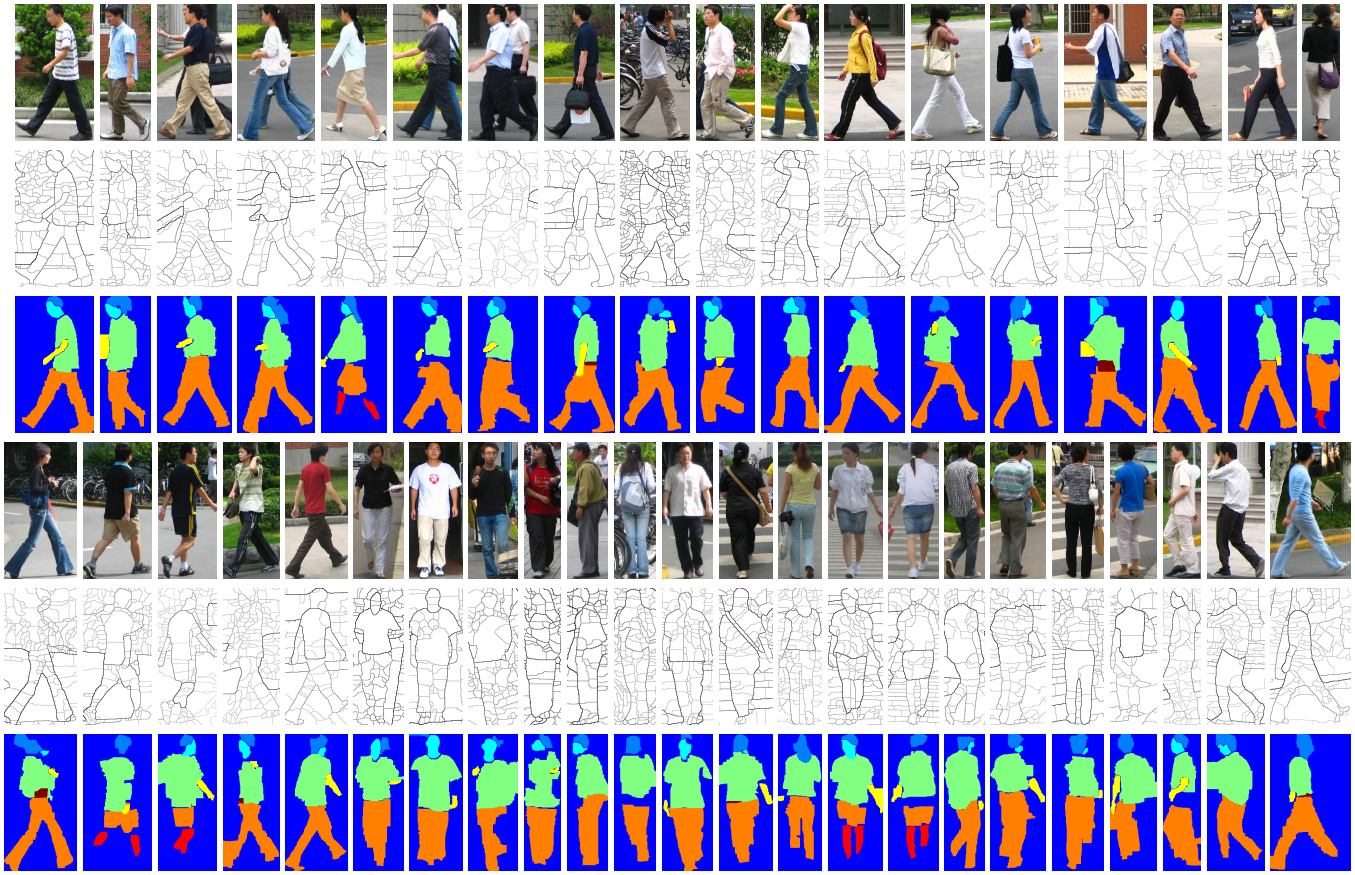

Figure 5. Example parses of pedestrians from tightly cropped bounding boxes. Rows show original image, Ultrametric Contour Map (UCM) from which candidate segments are assembled, and final top scoring segmentation result.

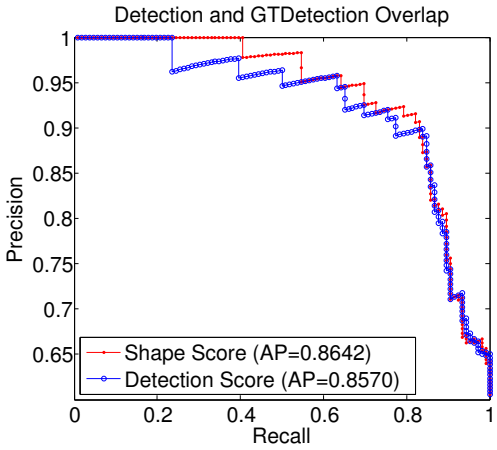

Figure 8. Detection rescoring using whole pedestrian shape score boosts performance in the high precision regime.

each detection returned. For the Penn-Fudan dataset, we set the detector to search over a range of scales where the resulting bounding boxes were 200 pixels or taller. This detector returns a list of candidate bounding boxes ordered by a detector confidence score. On the dataset of 100 test images the detector returned 175 candidate windows. Of these candidate windows, 60.5% are correct detections which overlap with a ground-truth bounding box by more than 50%.

Figure 6 shows example results of running the parser on bounding boxes returned by the template based pedestrian detector. In each case we report the top scoring parse for that subwindow. When the detector works well and returns a close-cropped bounding box, the parsing system also performs reasonably. However, the detector also returns false-positives which are not centered on a pedestrian (see examples in the bottom-row of Figure 6). For these subwindows the best parsing typically looks far less pedestrian-like.

We can exploit this fact in order to rescore the detection, similar to the work of [23]. We combine the detector score with the best match score computed for the whole pedestrian segmentation using a sigmoid function to convert the SVM output into a probability. Figure 8 shows the resulting

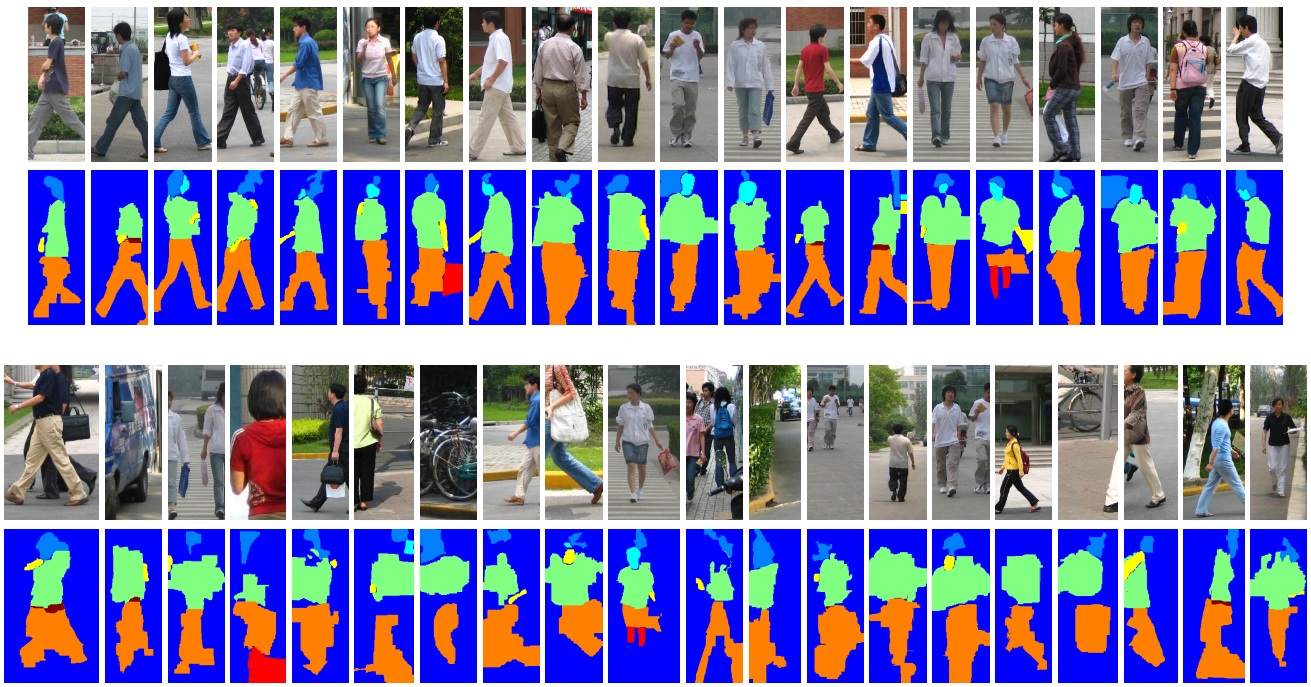

Figure 6. Example parses of pedestrians output from 175 bounding boxes produced by a multi-scale template based pedestrian detector[22]. Top row shows 20 highest ranked pedestrian parses based on combined detector and segmentation score. Bottom row shows the 20 lowest ranked pedestrian parses.

classification performance for this rescoring on the set of 175 candidate windows. Here, a bounding box which overlaps with a ground-truth by more than 50% is considered a true detection. The precision-recall curve shows that including the shape score gives a boost in performance in the high-precision regime.

8. Conclusion

We have described a simple compositional model for parsing images of pedestrians into visual parts. We are able to get good segmentation performance on realistic street scenes based largely on segment shapes learned from an entirely different dataset. While we have no doubt that a specialized, discriminatively trained classifier could outperform this approach in task of viewpoint and detector rescoring, we find it promising that simply using probabilistic estimates based on the shape training dataset yields decent results.

Much of this generalization ability rests on having good quality bottom-up segmentation. In cluttered, low-contrast images where shapes of candidate segments are small and indistinct, greedy assembly can often fail to discover good candidate parses. For example, Figure 9 shows examples of difficult images in which the top-scoring parse quality is poor. We expect that solving such cases will require having stronger models of shape and utilizing top-down feedback to refine segment boundaries where local image contrast is insufficient.

9. Acknowledgements

This work was supported by the UC Labs Research Program, a Google Research Award, and by a Development Grant for Computer Application Technology jointly sponsored by Beijing Municipal Commission of Education and Beijing Jiaotong University (XK100040519), National Nature Science Foundation of China (60975078,60902058,60805041,60872082), Beijing Natural Science Foundation (4092033,4112047) and Doctoral Foundations Ministry of Education of China (200800041049).

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *CVPR*, 2009. 2266

[2] E. Borenstein and S. Ullman. Combined top-down/bottom-up segmentation. *PAMI*, 2008. 2265

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 2269

[4] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. *ICCV*, 2001. 2268

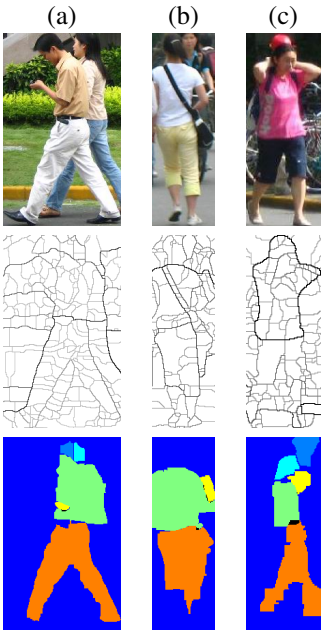

Figure 9. Examples of poor segmentations showing difficulty in correctly finding good candidate segments from pool of superpixels due to (a) lack of contrast between torsos (b) clutter around small parts (head and hair), (c) arms in unusual position.

[5] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010. 2267

[6] B. Catanzaro, B.-Y. Su, N. Sundaram, Y. Lee, M. Murphy, and K. Keutzer. Efficient, high-quality image contour detection. *ICCV*, 2009. 2267

[7] H. Chen, Z. Xu, Z. Liu, and S. C. Zhu. Composite templates for cloth modeling and sketching. In *CVPR*, 2006. 2266

[8] T. Cour and J. Shi. Recognizing objects by piecing together the segmentation puzzle. In *CVPR*, 2007. 2267

[9] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 2265

[10] P. Felzenszwalb and D. McAllester. A min-cover approach for finding salient curves. *POCV*, 2006. 2267

[11] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 2265

[12] C. Gu, J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, 2009. 2267

[13] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. In *CVPR*, 2006. 2265

[14] I. Kokkinos and A. Yuille. Hop: Hierarchical object parsing. In *CVPR*, 2009. 2265

[15] M. Kumar, P. Ton, and A. Zisserman. Obj cut. In *CVPR*, volume 1, 2005. 2265

[16] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV 04 workshop on statistical learning in computer vision*, pages 17–32, 2004. 2265

[17] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. *International Journal of Computer Vision*, 81(1):105–118, 2009. 2265

[18] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, 2010. 2267

[19] T. Malisiewicz and A. A. Efros. Improving spatial support for objects via multiple segmentations. *BMVC*, 2007. 2267

[20] G. Mori. Guiding model search using segmentation. In *ICCV*, 2005. 2265, 2267

[21] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, 2004. 2267, 2269

[22] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *ECCV*, 2010. 2269, 2271

[23] D. Ramanan. Using segmentation to verify object hypotheses. *CVPR*, 2006. 2265, 2270

[24] D. Ramanan. Learning to parse images of articulated bodies. *Advances in Neural Information Processing Systems*, 19:1129, 2007. 2265

[25] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, 2003. 2267

[26] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 2267

[27] L. Sigal and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Techniacl Report CS-06-08, Brown University*, 2006. 2268

[28] P. Srinivasan and J. Shi. Bottom-up recognition and parsing of the human body. In *CVPR*, 2007. 2265, 2269

[29] Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *IJCV*, 63(2):113–140, 2005. 2265

[30] L. Wang, J. Shi, G. Song, and I.-F. Shen. Object detection combining recognition and segmentation. In *ACCV*, 2007. 2269

[31] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object detection for multi-class segmentation. In *CVPR*, 2010. 2265

[32] S. Yu, R. Gross, and J. Shi. Concurrent object recognition and segmentation by graph partitioning. *NIPS*, pages 1407–1414, 2003. 2265

[33] L. Zhu, Y. Chen, C. Lin, and A. Yuille. Rapid inference on a novel and/or graph: Detection, segmentation and parsing of articulated deformable objects in cluttered backgrounds. In *NIPS*, 2007. 2266

[34] L. Zhu, Y. Chen, Y. Lu, C. Lin, and A. Yuille. Max margin and/or graph learning for parsing the human body. In *CVPR*, 2008. 2266

[35] Q. Zhu, L. Wang, Y. Wu, and J. Shi. Contour context selection for object detection: A set-to-set contour matching approach. In *ECCV*, 2008. 2267