Object Detection Combining Recognition and Segmentation

Liming Wang[1], Jianbo Shi[2], Gang Song[2], and I-fan Shen[1]

[1] Fudan University,Shanghai,PRC,200433 {wanglm,yfshen}@fudan.edu.cn
[2] University of Pennsylvania, 3330 Walnut Street, Philadelphia, PA 19104
jshi@cis.upenn.edu,songgang@seas.upenn.edu

Abstract. We develop an object detection method combining top-down recognition with bottom-up image segmentation. There are two main steps in this method: a hypothesis generation step and a verification step. In the top-down hypothesis generation step, we design an improved Shape Context feature, which is more robust to object deformation and background clutter. The improved Shape Context is used to generate a set of hypotheses of object locations and figure-ground masks, which have high recall and low precision rate. In the verification step, we first compute a set of feasible segmentations that are consistent with top-down object hypotheses, then we propose a *False Positive Pruning*(FPP) procedure to prune out false positives. We exploit the fact that false positive regions typically do not align with any feasible image segmentation. Experiments show that this simple framework is capable of achieving both high recall and high precision with only a few positive training examples and that this method can be generalized to many object classes.

1 Introduction

Object detection is an important, yet challenging vision task. It is a critical part in many applications such as image search, image auto-annotation and scene understanding; however it is still an open problem due to the complexity of object classes and images.

Current approaches ([1][2] [3][4][5] [6][7] [8] [9][10]) to object detection can be categorized by top-down, bottom-up or combination of the two. Top-down approaches ([11][2][12]) often include a training stage to obtain class-specific model features or to define object configurations. Hypotheses are found by matching models to the image features. Bottom-up approaches start from low-level or mid-level image features, i.e. edges or segments([8][5][9] [10]). These methods build up hypotheses from such features, extend them by construction rules and then evaluate by certain cost functions.

The third category of approaches combining top-down and bottom-up methods have become prevalent because they take advantage of both aspects. Although top-down approaches can quickly drive attention to promising hypotheses, they are prone to produce many false positives when features are locally extracted and matched. Features within the same hypothesis may not be consistent with respect to low-level image segmentation. On the other hand, bottom-up approaches try to keep consistency in low level image segmentation, but usually need much more efforts in searching and grouping.

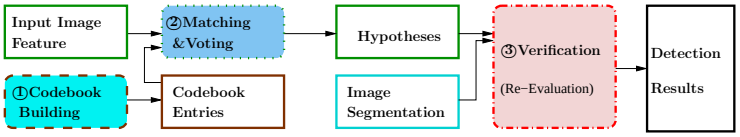


Fig. 1. Method overview. Our method has three parts (shaded rectangles). Codebook building (cyan) is the training stage, which generates codebook entries containing improved SC features and object masks. Top-down recognition (blue) generates multiple hypotheses via improved SC matching and voting in the input image. The verification part (pink) aims to verify these top-down hypotheses using bottom-up segmentation. Round-corner rectangles are processes and ordinary rectangles are input/output data.

Wisely combining these two can avoid exhaustive searching and grouping while maintaining consistency in object hypotheses. For example, Borenstein *et al.* enforce continuity along segmentation boundaries to align matched patches ([2]). Levin *et al.* take into account both bottom-up and top-down cues simultaneously in the framework of CRF([3]).

Our detection method falls into this last category of combining top-down recognition and bottom-up segmentation, with two major improvements over existing approaches. First, we design a new improved *Shape Context* (SC) for the top-down recognition. Our improved SC is more robust to small deformation of object shapes and background clutter. Second, by utilizing bottom-up segmentation, we introduce a novel *False Positive Pruning* (FPP) method to improve detection precision. Our framework can be generalized to many other object classes because we pose no specific constraints on any object class.

The overall structure of the paper is organized as follows. Sec. 2 provides an overview to our framework. Sec.3 describes the improved SCs and the top-down hypothesis generation. Sec.4 describes our FPP method combining image segmentation to verify hypotheses. Experiment results are shown in Sec.5, followed by discussion and conclusion in Sec.6.

2 Method Overview

Our method contains three major parts: codebook building, top-down recognition using matching and voting, and hypothesis verification, as depicted in Fig.1.

The object models are learned by building a codebook of local features. We extract improved SC as local image features and record the geometrical information together with object figure-ground masks. The improved SC is designed to be robust to shape variances and background clutters. For rigid objects and objects with slight articulation, our experiments show that only a few training examples suffice to encode local shape information of objects.

We generate recognition hypotheses by matching local image SC features to the codebook and use SC features to vote for object centers. A similar top-down voting scheme is described in the work of [4], which uses SIFT point features for pedestrian detection. The voting result might include many false positives due to small context

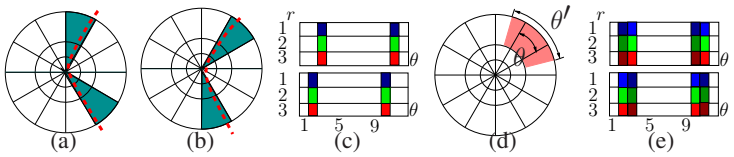


Fig. 2. Angular Blur. (a) and (b) are different bin responses of two similar contours. (c) are their histograms. (d) enlarges angular span θ to θ', letting bins be overlapped in angular direction. (e) are the responses on the overlapped bins, where the histograms are more similar.

of local SC features. Therefore, we combine top-down recognition with bottom-up segmentation in the verification stage to improve the detection precision. We propose a new *False Positive Pruning* (FPP) approach to prune out many false hypotheses generated from top-down recognition. The intuition of this approach is that many false positives are generated due to local mismatches. These local features usually do not have *segmentation consistency*, meaning that pixels in the same segment should belong to the same object. True positives are often composed of several connected segments while false positives tend to break large segments into pieces.

3 Top-down Recognition

In the training stage of top-down recognition, we build up a codebook of improved SC features from training images. For a test image, improved SC features are extracted and matched to codebook entries. A voting scheme then generates object hypotheses from the matching results.

3.1 Codebook Building

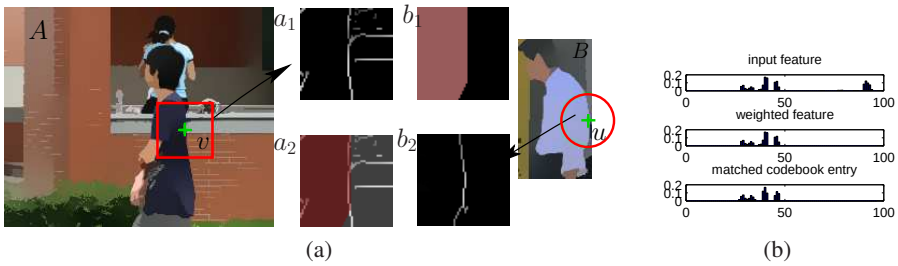
For each object class, we select a few images as training examples. Object masks are manually segmented and only edge map *inside* the mask is counted in shape context histogram to prune out edges due to background clutter.

The Codebook Entries (**CE**) are a repository of example features: $\mathbf{CE} = \{ce_i\}$. Each codebook entry $ce_i = (u_i, \delta_i, m_i, w_i)$ records the feature for a point i in labelled objects of the training images. Here u_i is the shape context vector for point i. δ_i is the position of point i relative to the object center. m_i is a binary mask of figure-ground segmentation for the patch centered at point i. w_i is the weight mask computed on m_i, which will be introduced later.

3.2 Improved Shape Context

The idea of *Shape Context* (SC) was first proposed by Belongie *et al.* ([13]). The basic definition of SC is a local histogram of edge points in a radius-angle polar grid. Following works ([14][15]) improve its distinctive power by considering different edge orientations. Besides SC, other local image features such as wavelets, SIFT and HOG have been used in keypoint based detection approaches ([4],[12]).

Suppose there are n_r (radial) by n_θ (angular) bins and the edge map E is divided into $E_1, \ldots, E_o$ by o orientations (similar to [15]), for a point at p, its SC is defined as $u = \{h_1, \ldots, h_o\}$, where



(a) (b)

Fig. 3. Distance function with mask. In (a), a feature point v has the edge map of a_1 around it. Using object mask b_1, it succeeds to find a good match to u in B (object model patch), whose edge map is b_2. a_2 is the object mask b_1 over a_1. Only the edge points falling into the mask area are counted for SC. In (b), histograms of a_1, a_2 and b_2 are shown. With the mask function, a_2 is much closer to b_2, thus got well matched.

$$h_i(k) = \#\{q \neq p : q \in E_i, \overrightarrow{pq} \in bin(k)\}, \quad k = 1, 2, ..., n_r n_\theta \tag{1}$$

Angular Blur A common problem for the shape context is that when dense bins are used or contours are close to the bin boundaries, similar contours have very different histograms (Fig.2-(c)). This leads to a large distance for two similar shapes if L_2-norm or χ^2 distance function is used. EMD([16]) alleviates this by solving a transportation problem; but it is computationally much more expensive.

The way we overcome this problem is to overlap spans of adjacent angular bins: $bin(k) \cap bin(k+1) \neq \emptyset$ (Fig.2-(d)). This amounts to blurring the original histogram along the angular direction. We call such an extension *Angular Blur*. One edge point in the overlapped regions are counted in both of the adjacent bins. So the two contours close to the original bin boundary will have similar histograms for the overlapping bins(Fig.2-(e)). With angular blur, even simple L_2-norm can tolerate slight shape deformation. It improves the basic SC without the expensive computation of EMD.

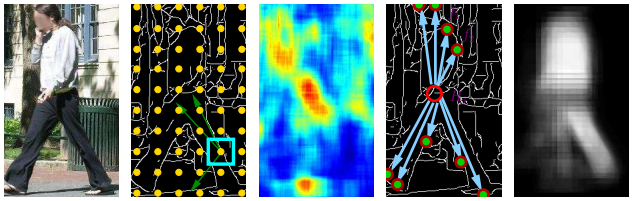
Mask Function on Shape Context In real images, objects SCs always contain background clutter. This is a common problem for matching local features. Unlike learning methods ([1][12]) which use a large number of labeled examples to train a classifier, we propose to use a mask function to focus only on the parts inside object while ignoring background in matching.

For $ce = (u, \delta, m, w)$ and a SC feature f in the test image, each bin of f is masked by figure-ground patch mask m of ce to remove the background clutter. Formally, we compute the weight w for bin k and distance function with mask as:

$$w(k) = Area(bin(k) \cap m)/Area(bin(k)), \quad k = 1, 2, ..., n_r n_\theta \tag{2}$$

$$D_m(ce, f) = D(u, w \cdot v) = ||u - w \cdot v||^2 \tag{3}$$

where $(\cdot)$ is the element-wise product. D can be any distance function computing the dissimilarity between histograms (We simply use L_2-norm). Figure 3 gives an example for the advantage of using mask function.



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

Fig. 4. Top-down recognition. (a) An input image;(b) A matched point feature votes for 3 possible positions; (c) The vote map V. (d) The hypothesis H_j traces back find its voters $\{f_i\}$. (d) Each f_i predicts the figure-ground configration using Eq. (5).

3.3 Hypothesis Generation

The goal of hypothesis generation is to predict possible object locations as well as to estimate the figure-ground segmentation for each hypothesis. Our hypothesis generation is based on a voting scheme similar to [4]. Each SC feature is compared with every codebook entry and makes a prediction of the possible object center. The matching scores are accumulated over the whole image and the predictions with the maximum scores are the possible object centers. Given a set of detected features $\{f_i\}$ at location $\{l_i\}$, we define the probability of matching codebook entry ce_k to f_i as $P(ce_k|l_i) \propto \exp(-D_m(ce_k, f_i))$. Given the match of ce_k to f_i, the probability of an object o with center located at c is defined as $P(o, c|ce_k, l_i) \propto \exp(-||c + \delta_k - l_i||^2)$. Now the probability of the hypothesis of object o with center c is computed as:

$$P(o, c) = \sum_{i,k} P(o, c|ce_k, l_i)P(ce_k|l_i)P(l_i) \tag{4}$$

$P(o, c)$ gives a voting map V of different locations c for the object class o. Extracting local maxima in V gives a set of hypotheses $\{H_j\} = \{(o_j, c_j)\}$.

Furthermore, figure-ground segmentation for each H_j can be estimated by backtracing the matching results. For those f_i giving the correct prediction, the patch mask m in the codebook is "pasted" to the corresponding image location as the figure-ground segmentation. Formally, for a point p in image at location p_l, we define $P(p = fig|ce_k, l_i)$ as the probability of point p belonging to the foreground when the feature at location l_i is matched to the codebook ce_k: $P(p = fig|ce_k, l_i) \propto \exp(-||p_l - l_i||)m_k(\overrightarrow{p_l l_i})$. And we assume that $P(ce_k, l_i|H_j) \propto P(o_j, c_j|ce_k, l_i)$ and $P(f_i|ce_k) \propto P(ce_k|f_i)$. The figure-ground probability for hypothesis H_j is estimated as

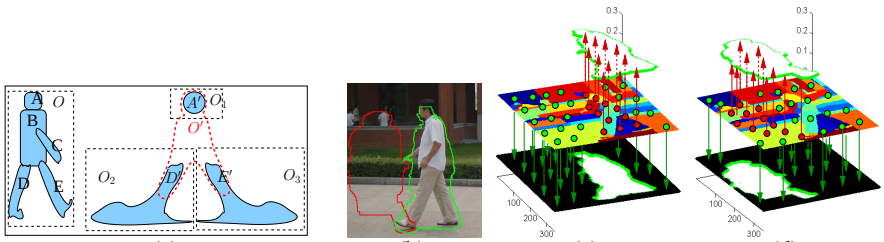
$$P(p = fig|H_j) \propto \sum_k \exp(-||p_l - l_i||)m_k(\overrightarrow{p_l l_i})P(f_i|ce_k)P(ce_k, l_i|Hj) \tag{5}$$

Eq. (5) gives the estimation of top-down segmentation. The whole process of top-down recognition is shown in Fig. 4. The binary top-down segmentation (F, B) of figure(F) and background (B) is the obtained by thresholding $P(p = fig|H_j)$.

4 Verification: Combining Recognition and Segmentation

From our experiments, the top-down recognition using voting scheme will produce many *False Positives* (FPs). In this section, we propose a two-step procedure of *False Positive Pruning* (FPP) to prune out FPs. In the first step we refine the top-down hypothesis mask by checking its consistency with bottom-up segmention. Second the final score on the refined mask is recomputed by considering spatial constraints.

Combining Bottom-up Segmentation The basic idea for local feature voting is to make global decision by the consensus of local predictions. However, these incorrect local predictions using a small context can accumulate and confuse the global decision. For example, in pedestrian detection, two trunks will probably be locally taken as human legs and produce a human hypothesis (in Fig. 5-(a)); another case is the silhouettes from two standing-by pedestrians.



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

Fig. 5. Combining bottom-up segmentation. FPs tend to spread out as multiple regions from different objects. In example of (a). an object O consists of five parts (A, B, C, D, E). $(A' \in O_1, D' \in O_2, E' \in O_3)$ are matched to (A, D, E) because locally they are similar. The hypothesis of $O' = (A', D', E')$ is generated. (b) shows boundaries of a FP (in red) and a TP (in green) in a real image. (c) is the layered view of the TP in (b). The top layer is the top-down segmentation, which forms a force (red arrows) to pull the mask out from the image. The bottom layer is the background force (green arrows). The middle layer is the top-down segmentation (we threshold it to binary mask) over the segmentation results.(d) is the case for the FP.

In pedestrian detection, the top-down figure-ground segmentation masks of the FPs usually look similar to a pedestrian. However we notice that such top-down mask is not *consistent* with the bottom-up segmentation for most FPs. The bottom-up segments share bigger contextual information than the local features in the top-down recognition and are homogenous in the sense of low-level image feature. The pixels in the same segment should belong to the same object. Imagine that the top-down hypothesis mask(F, B) tries to pull the object F out of the whole image. TPs generally consists of several well-separated segments from the background so that they are easy to be pulled out (Fig. 5-(c)). However FPs often contain only part of the segments. In the example of tree trunks, only *part* of the tree trunk is recognized as foreground while the whole tree trunk forms one bottom-up segment. This makes pulling out FPs more difficult because they have to break the homogenous segments (Fig. 5-(d)).

Based on these observations we combine the bottom-up segmentation to update the top-down figure-ground mask. Incorrect local predictions are removed from the mask if they are not consistent with the bottom-up segmentation. We give each bottom-up segment S_i a binary label. Unlike the work in [17] which uses graph cut to propose the optimized hypothesis mask, we simply define the ratio $\frac{Area(S_i \bigcap F)}{Area(S_i \bigcap B)}$ as a criteria to assign S_i to F or B. We try further segmentation when such assignment is uncertain to avoid the case of under-segmentation in a large area. The Normalized Cut (NCut) cost ([18]) is used to determine if such further segmentation is reasonable. The procedure to refine hypothesis mask is formulated as follows:

Input: top-down mask (F, B) and bottom-up segments $\{S_i, i = 1, \ldots, N\}$.
Output: refined object mask (F, B).
Set $i = 0$.
1) If $i > N$,exit; else, $i = i + 1$.
2) If $\Lambda = \frac{Area(S_i \bigcap F)}{Area(S_i \bigcap B)} > \kappa_{up}$, then $F = F \cup S_i$,goto 1);
elseif $\Lambda < \kappa_{down}$, then $F = F - (F \cap S_i)$, goto 1). Otherwise, go to 3).
3) Segment S_i to (S_i^1, S_i^2). If $\zeta = \text{NCut}(S_i) > \Upsilon_{up}$, $F = F - (F \cap S_i)$, goto 1);
else $S_{N+1} = S_i^1$,$S_{N+2} = S_i^2$, $S = S \cup \{S_{N+1}, S_{N+2}\}$, $N = N + 2$, goto 1).

Re-Evaluation There are two advantages with the updated masks. The first is that we can recompute more accurate local features by masking out the background edges. The second is that the shapes of updated FPs masks will change much more than those of TPs, because FPs are usually generated by locally similar parts of other objects, which will probably be taken away through the above process. We require TPs must have voters from all the different locations around the hypothesis center. This will eliminates those TPs with less region support or with certain partial matching score.

The final score is the summation of the average scores over the different spatial bins in the mask. The shape of the spatial bins are predefined. For pedestrians we use the radius-angle polar ellipse bins; for other objects we use rectangular grid bins. For each hypothesis, SC features are re-computed over the masked edge map by F and feature f_i is only allowed to be matched to ce_k in the same bin location. For each bin j, we compute an average matching score $E_j = \frac{\sum P(ce_k|f_i)}{\#(ce_k, f_i)}$, where both ce_k and f_i come from bin j. The final score of this hypothesis is defined as:

$$E = \sum_j E'_j \text{ ,where } E'_j = \begin{cases} E_j & \text{, if } E_j > \alpha; \\ -\alpha & \text{, if } E_j = 0 \text{ and } \#\{ce_k, ce_k \in \text{bin}(j)\} > 0. \end{cases} \quad (6)$$

The term α is used to penalize the bins which have no matching with the codebook. This decreases the scores of FPs with only part of true objects, i.e. bike hypothesis with one wheel. Experiments show that our FPP procedure can prune out FPs effectively.

5 Results

Our experiments test different object classes including pedestrian, bike, human riding bike, umbrella and car (Table. 1). These pictures were taken from scenes around campus

Table 1. Dataset for detection task

#Object	Pedestrian	Bike	Human on bike	Umbrella	Car
Training	15	3	2	4	4
Testing	345	67	19	16	60

and urban streets. Objects in the images are roughly at the same scale. For pedestrians, the range of the heights is from 186 to 390 pixels.

For our evaluation criteria, a hypothesis whose center falls into an ellipse region around ground truth center is classified as true positive. The radii for ellipse are typically chosen as 20% of the mean width / height of the objects. Multiple detections for one ground truth object are only counted once.

Angular Blur and Mask Function Evaluation We compare the detection algorithm on images w/ and w/o Angular Blur (AB) or mask function. The PR curves are plotted in Fig.6. For pedestrian and umbrella detection, it is very clear that adding Angular Blur and mask function can improve the detection results. For other object classes, AB+Mask outperforms at high-precision/low-recall part of the curve, but gets no significant improvement at high-recall/low-precision part. The reason is that AB+Mask can improve the cases where objects have deformation and complex background clutter. For bikes, the inner edges dominate the SC histogram; so adding mask function makes only a little difference.

Pedestrian Detection Compared with HOG We also compare with HOG.using the implementation of the authors of ([12]) Figure 6-(a) shows that our method with FPP procedure are better than the results of HOG. Note that we only use a very limited number of training examples as shown in Table. 1 and we did not utilize any negative training examples.

6 Conclusion and Discussion

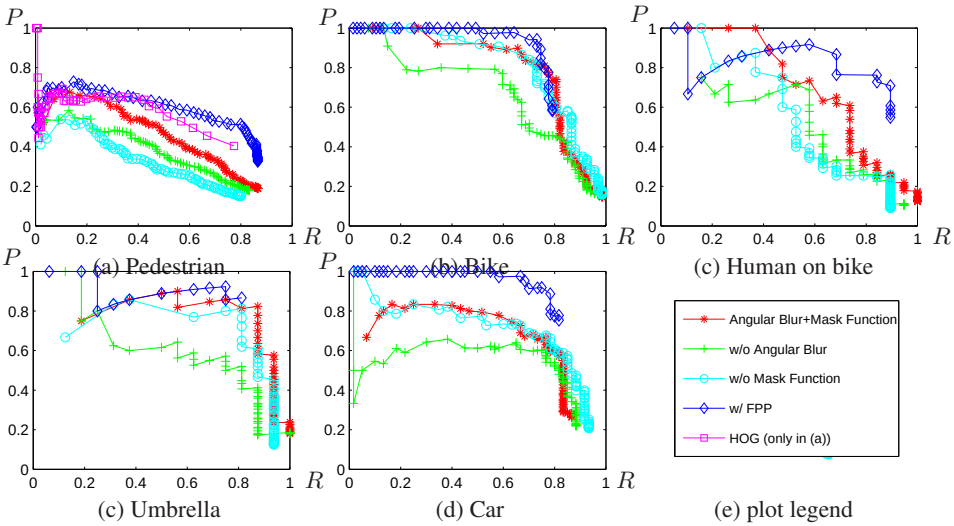
In this paper, we developed an object detection method of combining top-down model-based recognition with bottom-up image segmentation. Our method not only detects object positions but also gives the figure-ground segmentation mask. We designed an improved Shape Context feature for recognition and proposed a novel FPP procedure to verify hypotheses. This method can be generalized to many object classes.

Results show that our detection algorithm can achieve both high recall and precision rates. However there are still some FPs hypotheses that cannot be pruned. They are typically very similar to objects, like a human-shape rock, or some tree trunks. More information like color or texture should be explored to prune out these FPs. Another failure case of SC detector is for very small scale object. These objects have very few edges points thus are not suitable for SC. Also our method does no work for severe occlusion where most local information is corrupted.

Acknowledgment This work is partially supported by National Science Foundation through grants NSF-IIS-04-47953(CAREER) and NSF-IIS-03-33036(IDLP). We thank Qihui Zhu and Jeffrey Byrne for polishing the paper.

References

1. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: CVPR. (2001)



(a) Pedestrian (b) Bike (c) Human on bike

(c) Umbrella (d) Car (e) plot legend

Fig. 6. PR-Curves of object detection results.

2. Borenstein, E., Ullman, S.: Class-specific, top-down segmentation. In: ECCV (2). (2002)
3. Levin, A., Weiss, Y.: Learning to combine bottom-up and top-down segmentation. In: ECCV. (2006)
4. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: CVPR. (2005)
5. Ferrari, V., Tuytelaars, T., Gool, L.J.V.: Object detection by contour segment networks. In: ECCV. (2006)
6. Kokkinos, I., Maragos, P., Yuille, A.L.: Bottom-up & top-down object detection using primal sketch features and graphical models. In: CVPR. (2006)
7. Zhao, L., Davis, L.S.: Closely coupled object detection and segmentation. In: ICCV. (2005)
8. Ren, X., Berg, A.C., Malik, J.: Recovering human body configurations using pairwise constraints between parts. In: ICCV. (2005)
9. Mori, G., Ren, X., Efros, A.A., Malik, J.: Recovering human body configurations: Combining segmentation and recognition. In: CVPR. (2004)
10. Srinivasan, P., Shi, J.: Bottom-up recognition and parsing of the human body. In: CVPR. (2007)
11. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision **61**(1) (2005)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
13. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4) (2002)
14. Mori, G., Belongie, S.J., Malik, J.: Efficient shape matching using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **27**(11) (2005)
15. Thayananthan, A., Stenger, B., Torr, P.H.S., Cipolla, R.: Shape context and chamfer matching in cluttered scenes. In: CVPR. (2003)
16. Rubner, Y., Tomasi, C., Guibas, L.J.: A metric for distributions with applications to image databases. In: ICCV. (1998)
17. Ramanan, D.: Using segmentation to verify object hypotheses. In: CVPR. (2007)
18. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: CVPR. (1997)

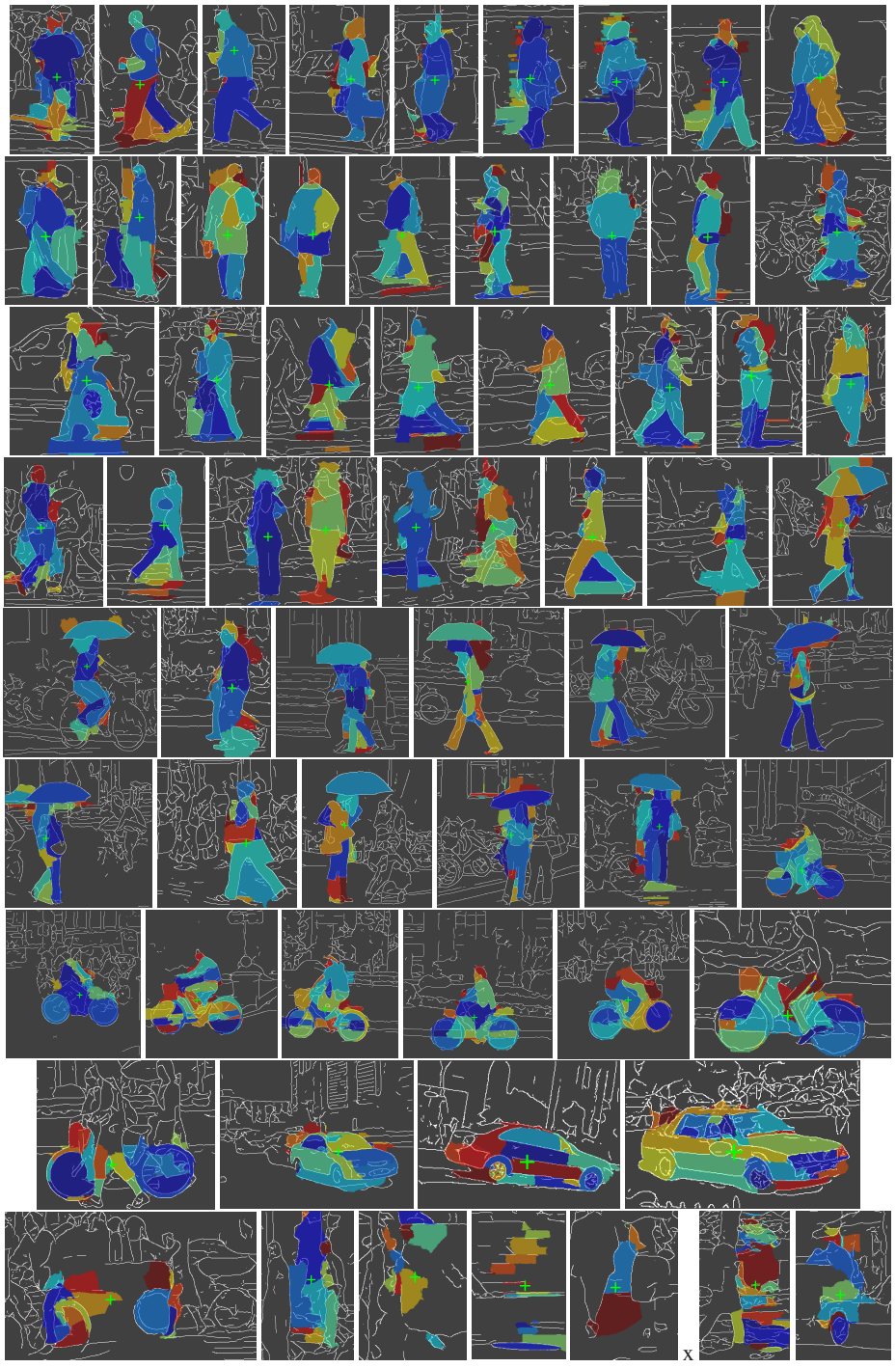


Fig. 7. Detection result on real images.The color indicates different segments. The last row contains cases of FPs for bikes and pedestrians.