Finding Things: Image Parsing with Regions and Per-Exemplar Detectors

Joseph Tighe
University of North Carolina at Chapel Hill
jtighe@cs.unc.edu

Svetlana Lazebnik
University of Illinois at Urbana-Champaign
slazebni@illinois.edu

1. Introduction

This paper addresses the problem of image parsing, or labeling each pixel in an image with its semantic category. Our goal is achieving *broad coverage* – the ability to recognize hundreds or thousands of object classes that commonly occur in everyday street scenes and indoor environments. A major challenge in doing this is posed by the non-uniform statistics of these classes in realistic scene images. A small number of classes – mainly ones associated with large regions or "stuff," such as road, sky, trees, buildings, etc. – constitute the majority of all image pixels and object instances in the dataset. But a much larger number of "thing" classes – people, cars, dogs, mailboxes, vases, stop signs – occupy a small percentage of image pixels and have relatively few instances each.

Our proposed method is outlined in Figure 2. It combines the region-based parser from our earlier work [6] with a novel parser based on per-exemplar detectors. Each parser produces a score or *data term* for each possible label at each pixel location, and the data terms are combined using a support vector machine (SVM) to generate the final labeling. This scheme produces state-of-the-art results on three challenging datasets: SIFT Flow [4], LM+SUN [6], and CamVid [1]. In particular, the LM+SUN dataset, with 45,676 images and 232 labels, has the broadest coverage of any image parsing benchmark to date.

Complete code and results for our system can be found at `http://www.cs.unc.edu/SuperParsing`.

2. Method

This section presents our hybrid image parsing method as illustrated in Figure 2. We start with our region-based parser from our earlier work [6], which gives a log likelihood score for each class at each pixel (E_{R}). Section 2.1 describes our detector-based component, and Section 2.2 outlines our proposed method for combining our region- and detector-based components.

2.1. Detector-Based Parsing

Following the per-exemplar framework of [5], we train a per-exemplar detector for each labeled object instance in our dataset. At test time, given an image that needs to be parsed, we first obtain a retrieval set of globally similar training images as in [6]. Then we run the detectors associated with the first k instances of each class in that retrieval set (the instances are ranked in decreasing order of the similarity of their image to the test image, and different instances in the same image are ranked arbitrarily). We restrict k purely to reduce computation; all our experiments use $k = 100$. Next, we take all detections that are above a given threshold t_d (we use the negative margin or $t_d = -1$ as suggested in [5]). For each detection we project the associated object mask into the detected bounding box (Figure 1). To compute the *detector-based data term* E_{D} for a class c and pixel p, we simply take the sum of all detection masks from that class weighted by their detection scores:

$$E_{\mathrm{D}}(p, c) = \sum_{d \in D_{p,c}} (w_d - t_d)\,, \qquad (1)$$

where $D_{p,c}$ is the set of all detections for class c whose transferred mask overlaps pixel p and w_d is the detection score of d. Figure 2(e) shows some detector-based data terms for the test image of Figure 2(a).

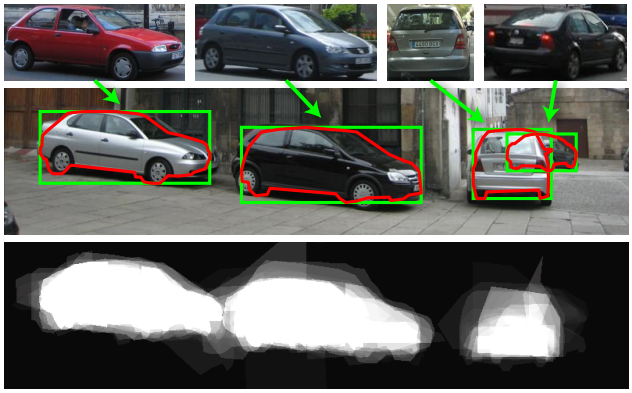

Figure 1. Computation of the detector-based data term. For each positive detection (green bounding box) in the test image (middle row) we transfer the mask (red polygon) from the associated exemplar (top) into the test image. The data term for "car" (bottom) is obtained by summing all the masks weighted by their detector responses.

2.2. SVM Combination and MRF Smoothing

Once we run the parsing systems of [6] and 2.1 on a test image, for each pixel p and each class c we end

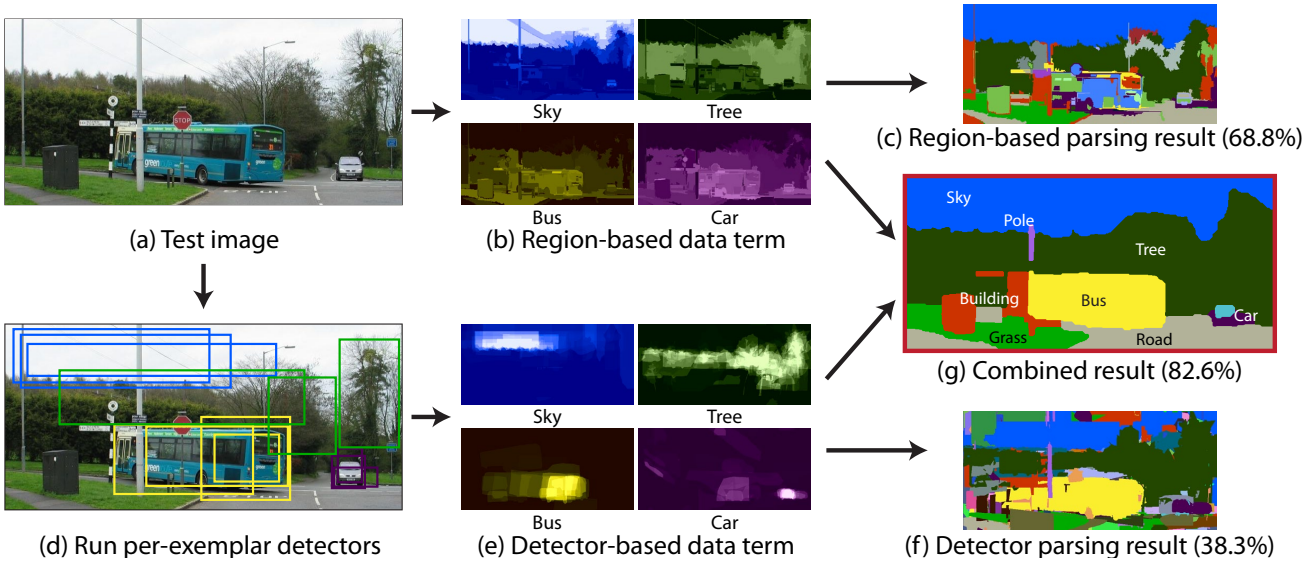

Figure 2. Overview and sample result of our approach. The test image (a) contains a bus – a relatively rare "thing" class. Our region-based parsing system [6] computes class likelihoods (b) based on superpixel features, and it correctly identifies "stuff" regions like sky, road, and trees, but is not able to get the bus (c). To find "things" like bus and car, we run per-exemplar detectors [5] on the test image (d) and transfer masks corresponding to detected training exemplars (e). Since the detectors are not well suited for "stuff," the result of detector-based parsing (f) is poor. However, combining region-based and detection-based data terms (g) gives the highest accuracy of all and correctly labels most of the bus and part of the car.

SIFT Flow	Per-Pixel	Per-Class
Ours: Combined MRF	78.6	39.2
Tighe and Lazebnik [6]	77.0	30.1
Liu et al. [4]	76.7	N/A
Farabet et al. [3]	78.5	29.6
Farabet et al. [3] balanced	74.2	46.0
Eigen and Fergus [2]	77.1	32.5

Table 1. Comparison to state-of-the-art on the SIFT Flow dataset.

up with two data terms, $E_{\mathrm{R}}(p, c)$ and $E_{\mathrm{D}}(p, c)$. For a dataset with C classes, concatenating these values gives us a $2C$-dimensional feature vector at each pixel. Next, we train C one-vs-all SVMs, each of which takes as input the $2C$-dimensional feature vectors and returns final per-pixel scores for a given class c.

Training data for each SVM is generated by running region- and detector-based parsing on the entire training set using a leave-one-out method: for each training image a retrieval set of similar training images is obtained, regions are matched to generate $E_{\mathrm{R}}(p, c)$, and the per-exemplar detectors from the retrieval set are run to generate $E_{\mathrm{D}}(p, c)$. The resulting SVMs produce C responses at each pixel. Let $E_{\mathrm{SVM}}(p_i, c_i)$ denote the response of the SVM for class c_i at pixel p_i. To obtain the final labeling, we can simply take the highest-scoring label at each pixel, but this produces noisy results. We smooth the labels with an MRF energy function similar to [4].

2.3. Experiments

Table 1 compares our combined system to a number of state-of-the-art approaches on the SIFT Flow dataset. We outperform them, in many cases beating the average per-class rate by up to 10% while maintaining or exceeding the per-pixel rates. The one exception is the system of Farabet et al. [3] when tuned for balanced per-class rates, but their per-pixel rate is much lower than ours in this case.

LM+SUN	Per-Pixel	Per-Class
Ours: Combined MRF	61.4	15.2
Tighe and Lazebnik [6]	54.9	7.1

Table 2. Comparison to [6] on the LM+SUN dataset with results broken down by outdoor and indoor test images.

On LM+SUN, which has an order of magnitude more images and labels than SIFT Flow, the only previously reported results are from our earlier region-based system [6]. As Table 2 shows, by augmenting the region-based term with a novel detector-based data term and SVM inference, we are able to raise the per-pixel rate from 54.9% to 61.4% and the per-class rate from 7.1% to 15.2%.

References

[1] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008. 1

[2] D. Eigen and R. Fergus. Nonparametric image parsing using adaptive neighbor sets. In *CVPR*, 2012. 2

[3] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Scene parsing with multiscale feature learning, purity trees, and optimal covers. *Arxiv preprint arXiv:1202.2160 [cs.CV]*, 2012. 2

[4] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 33(12):2368–2382, June 2011. 1, 2

[5] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *ICCV*, 2011. 1, 2

[6] J. Tighe and S. Lazebnik. SuperParsing: Scalable nonparametric image parsing with superpixels. *IJCV*, 101(2):329–349, Jan 2013. 1, 2