Context Driven Scene Parsing with Attention to Rare Classes

Jimei Yang
UC Merced
jyang44@ucmerced.edu

Brian Price
Adobe Research
bprice@adobe.com

Scott Cohen
Adobe Research
scohen@adobe.com

Ming-Hsuan Yang
UC Merced
mhyang@ucmerced.edu

1. Introduction

The distribution of objects in natural images tends to be heavy-tailed, with many pixels in the images coming from common background classes (the sky, water, and sand in Figure 1) and far fewer pixels coming from any given one of the thousands of possible object types. The large number of rare object classes and their relatively small sizes in many images make it difficult for algorithms to accurately segment important objects (the persons and boat in Figure 1). In fact, when evaluating error on a per-pixel basis, the performance of algorithms can often be improved by eliminating the rare classes altogether if their sizes in the images are usually small, despite the rare classes being very important to scene understanding.

Towards an open-ended scene parsing system, many data-driven approaches have been proposed [1, 3, 4, 5, 6]. To parse an input image, these algorithms first retrieve a small set of similar images and their associated semantic labels from the database, and compute classification confidence maps by matching the query with retrieved images in pixels or superpixels. The final semantic labeling is obtained by solving a pairwise MRF model.

In this paper [7], we propose a novel context-driven scene parsing system. Different from previous approaches, we focus more on rare object classes aiming at generating richer and more structured semantic labelings. Beyond the three basic components of nonparametric algorithms, i.e. image retrieval, superpixel matching and MRF labeling, we make two novel contributions:

1. We propose to regularize the retrieval set by a dictionary of rare class superpixels, since the semantic labels of retrieval images usually follow a long-tailed distribution. Therefore, we obtain more balanced classification results.

2. Beyond the traditional co-occurrence statistics, we bring local and global spatial context into superpixel scoring process to refine image retrieval and superpixel classification, which gives us more contextually sensible parsing results.

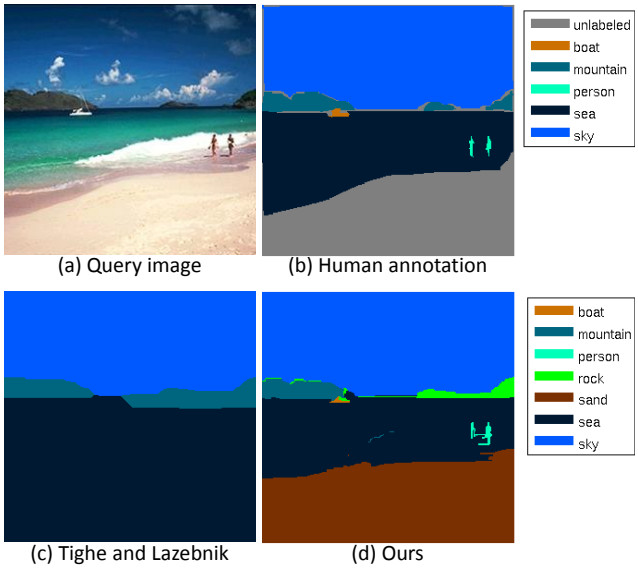

Figure 1. Given a query image (a), our method (d) recognizes small objects of rare classes (people, boat) while state-of-the-art systems (c) tend to miss them. Note that our method also recognizes the sand while the human annotator leaves it unlabeled (b).

We demonstrate our system on the SIFTflow dataset [3] (2688 images, 33 labels) and the LMSun dataset [6] (45676 images, 232 labels). The results show that the proposed algorithm achieves superior labeling performance than the previous state-of-the-art algorithms in terms of per-class accuracy and per-pixel accuracy on the rare classes, while still achieving similar or superior results on all classes.

2. Results

SIFTflow. The SIFTflow dataset consists of 2488 training images and 200 test images. All the images are 256×256 pixels from 33 semantic labels. We retrieve $K = 40$ images for each query. By applying the 80%-20% rule to all the superpixels of training dataset, we identify 5 classes as common while 28 classes as rare. We compare our results with recent work in Table 1. **LMSun.** The LMSun dataset consists of 45176 training images and 500 test images. The size of images ranges from 256×256 pixels to 800×600 pixels. There are 232 semantic labels in total. By using the same

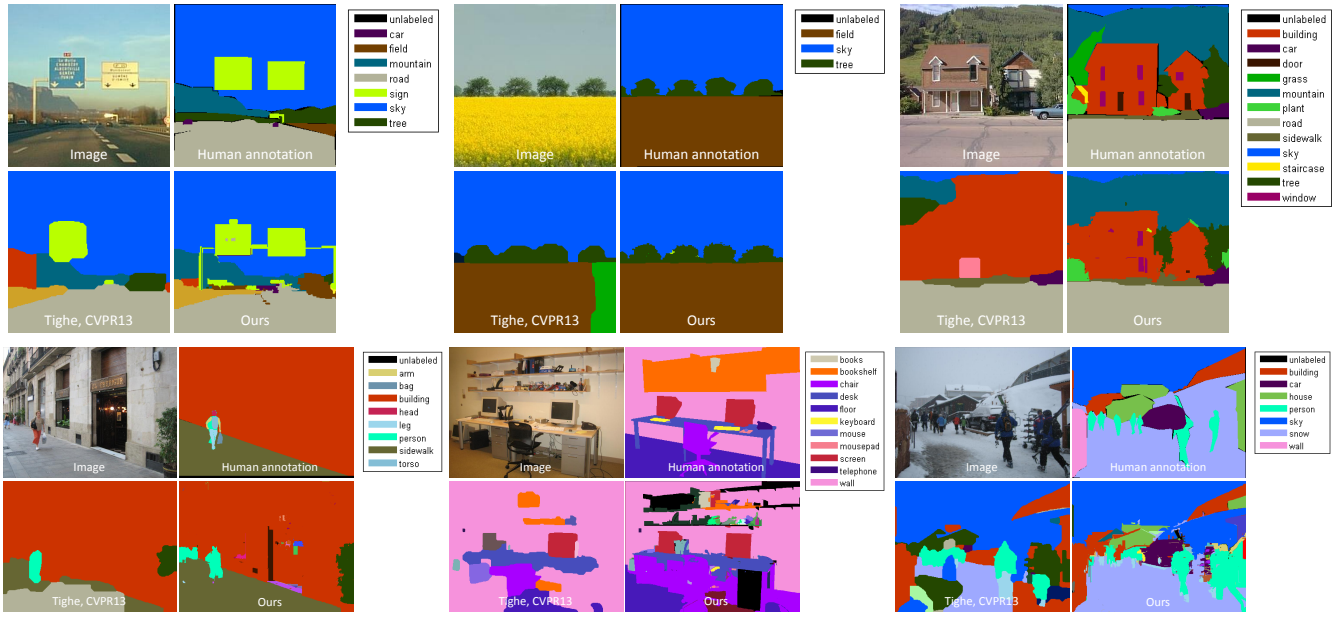

Figure 2. Some representative scene parsing results on the SIFTflow and LMSun datasets

Table 1. Comparing accuracy (%) on the SIFTflow dataset.

SIFTflow	Per-pixel	Per-class
Liu et al. [3]	76.7	N/A
Farabet et al. [2]	78.5	29.5
Farabet et al. [2] balanced	74.2	46.0
Eigen et al. [1]	77.1	32.5
Singh and Kosecka [4]	79.2	33.8
Tighe and Lazebnik [6]	77.0	30.1
Tighe and Lazebnik [5]	78.6	39.2
baseline	78.0	27.5
Ours	79.8	48.7
28 rare classes	Per-pixel	Per-class
Tighe and Lazebnik [5]	48.8	29.9
Ours	59.4	41.9

80%-20% rule on all the superpixels in the training set, we identify 47 common classes and 185 rare classes. Since this is more complex dataset, we retrieve $K = 120$ images to cover large appearance variations. We compare our results with recent work in Table 2. It turns out that our system

Table 2. Comparing accuracy (%) on the LMSun dataset.

LMSun	Per-pixel	Per-class
Tighe and Lazebnik [6]	54.9	7.1
Tighe and Lazebnik [5]	61.4	15.2
Our	60.6	18.0
baseline	58.5	9.0
185 rare classes	Per-pixel	Per-class
Tighe and Lazebnik [5]	19.0	12.9
Our	26.4	14.4

outperforms the state-of-the-art for both per-pixel and per-class rates, which further demonstrates our contributions to rare class boosting.

References

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

[2] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Scene parsing with multiscale feature learning, scene parsing with multiscale feature learning, purity trees, and optimal covers. In *ICML*, 2012.

[3] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 33:2368 – 2382, 2011.

[4] G. Singh and J. Kosecka. Nonparametric scene parsing with adaptive feature relevance and semantic context. In *CVPR*, 2013.

[5] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[6] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. *IJCV*, 101:329–349, 2013.

[7] J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *CVPR*, 2014.