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

Abstract

This paper presents a scalable scene parsing algorithm based on image retrieval and superpixel matching. We focus on rare object classes, which play an important role in achieving richer semantic understanding of visual scenes, compared to common background classes. Towards this end, we make two novel contributions: rare class expansion and semantic context description. First, considering the long-tailed nature of the label distribution, we expand the retrieval set by rare class exemplars and thus achieve more balanced superpixel classification results. Second, we incorporate both global and local semantic context information through a feedback based mechanism to refine image retrieval and superpixel matching. Results on the SIFTflow and LMSun datasets show the superior performance of our algorithm, especially on the rare classes, without sacrificing overall labeling accuracy.

1. Introduction

The goal of scene parsing is to associate a semantic label such as sky, trees, cars, etc. with every pixel in a still image. Such an image description has broad applications, e.g. image editing, image search and autonomous vehicles. Considering potentially hundreds or thousands of semantic labels in common outdoor environments and indoor scenes, it is of great interest to endow the scene parsing system with the ability to operate in a large scale. Large scale scene parsing faces two main challenges. First, 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 human observers. Secondly, it is expensive to optimize image labeling problems with hundreds or thousands of labels. For example, given efficient optimization algorithms such as graph-cut [1], it still takes minutes to solve a pairwise Markov Random Field (MRF) on a 600×800 image with 100 labels. These two challenges make learning based algorithms [19, 11, 5] less applicable.

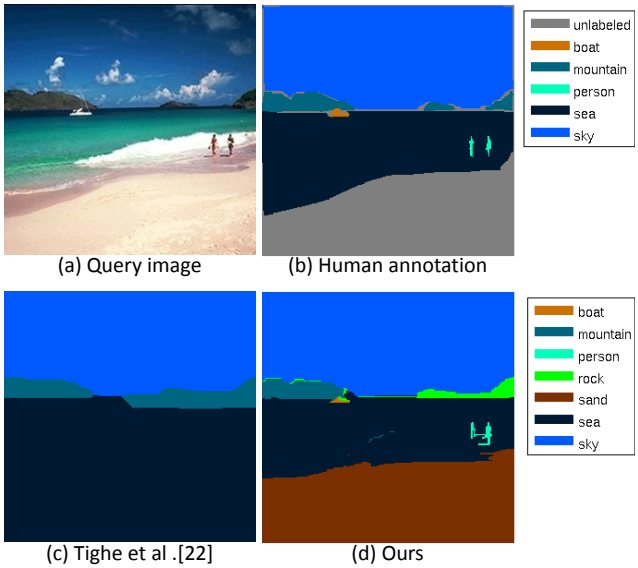

Figure 1. Given a query image (a), our method (d) recognizes small objects of rare classes (people, boat) while state-of-the-art systems (c) tend to miss them. Note that our method also recognizes the sand while the human annotator leaves it unlabeled (b).

An alternative is to use nonparametric approaches [15, 4, 20, 22, 21]. 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 superpxiels. The final semantic labeling is obtained by solving a pairwise MRF model. The core of these nonparametric algorithms is motivated by two important observations. On one hand, a single image usually contains very few labels that are constrained

by the scene category of the image, compared to the hundreds in total. For example, a photo taken at the beach usually contains sand, sea, boat and person (Figure 1(a)). Image retrieval constrains the potentially large number of candidate labels to the ones present in the similar scenes, which greatly reduce the labeling efforts. On the other hand, segments usually capture different partial appearances of objects and are difficult to fit into unified category models. Matching-based algorithms break the category barrier and recognition can be realized by transferring labels from matched segments to query segments. These nonparametric approaches [15, 4, 20, 22, 21] achieve good performance on overall per-pixel labeling accuracy. However, by taking a closer look at their semantic labeling results, we find that the overall performance is in fact dominated by several common background classes, e.g. mountain, building, water, road and wall, while the performance on interesting object classes is still lagging, e.g. people, animals and manmade objects.

In this paper, 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. Visual context plays an important role in scene parsing [17]. 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.

We demonstrate our system on the SIFTflow dataset (2688 images, 33 labels) and the LMSun dataset [22] (45576 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. Related Work

With the same concern on interesting object classes, Tighe and Lazebnik [21] propose to augment their previous superpixel based parsing system [22] with pre-trained exemplar SVMs [16]. Although per-exemplar detectors makes it possible to transfer object shape masks, their training indeed requires considerable amounts of computational resources and their output also needs to be calibrated with superpixel parsing output in a complicated post-classification step. This hybrid system shows state-of-the-art labeling performance in general, but is still constrained by the quality of per-exemplar detectors and the over-smoothing property of post-processing. In Figure 1(c), for example, it misses two interesting object classes (boat and person), although it achieves good overall per-pixel labeling accuracy. Instead of using expensive object detectors, we pay more attention to the context in different levels.

Context has been investigated in various semantic segmentation algorithms from co-occurrence statistics to scene categories [12, 4, 20, 7]. In terms of using semantic context information for nonparametic scene parsing, our paper is closely related to [4, 20]. Similar to ours, both of them learn context information through the feedback mechanism in the parsing process. The key idea is to first label the input image only using appearance information, and then extract context information from initial semantic labels. Eigen and Fergus [4] investigate context information in superpixel neighborhoods. By observing that the initial labeling usually produces reliable results on background classes, they build a context index for each superpixel using background labels in its four-directional neighborhood. Therefore, in additional to image retrieval, their method is able to find more relevant superpixels for matching. However, in large scale, even similar background (indoor) can encapsulate many object categories, which will result in the uncontrollable size of retrieved superpixels. Singh and Kosecka [20] instead focus on global semantic context. They construct a semantic label descriptor for each image in a three-layer spatial pyramid to refine image retrieval. Compared to appearance image descriptors, semantic label descriptors are much lower dimensional and allow fast image matching, but also more vulnerable when the initial labeling fails. The major difference of our method from [4, 20] lies in that we incorporate richer context information and show its effectiveness in large scale. Instead of using semantic labels for context description, we build our global and local context descriptors based on classification likelihood maps, in a way similar to the Object Bank representation [9]. Our context representations are more tolerant to labeling mistakes. Furthermore, we use our context descriptors together with appearance features for superpixel classification. This combined feature representation has been shown to be effective for semantic segmentation at the patch level [23].

3. The Baseline System

Our base system consists of three components: image retrieval, superpixel matching and MRF labeling.

3.1. Image Retrieval

Image retrieval is a critical step for our system. It determines the labels we use to parse the input image. If the algorithm fails to retrieve relevant images and true labels, we are not able to recover them in later steps. In this paper, we use the method in [18] to compute the spatially constrained image similarity $m(I_q, I_d)$ between the query image I_q and database images I_d, and retrieve top-K most similar images $\{I_d^1, I_d^2, ..., I_d^K\}$,where $m(I_q, I_d^k) > m(I_q, I_d^{k+1})$. Based on the Bag-of-Words image matching algorithm, this method incorporates spatial voting of local features (SIFT and RGB color) into image scoring. Therefore, the retrieved images usually have similar scene layout to the query, which is desirable for our parsing system. We use a SIFT vocabulary of 10,000 words and a RGB color vocabulary of 1,000 words for local feature quantization. The retrieval top-K images also determine a subset of candidate labels $\mathcal{L}' \subset \mathcal{L}$ for the query image, where $\mathcal{L}$ is the overall label set. This method shares a similar spirit with commonly used spatial pyramid matching [13], but favors scene retrieval more than scene classification in terms of implementation.

3.2. Superpixel Matching

We intend to assign semantic labels to every pixel of the query image, based on retrieval images and their corresponding ground truth semantic labels (annotated by human). As a single pixel alone does not contain sufficient information for recognition, we thus choose to recognize pixels in their proper neighboring regions, i.e. superpixels. We use the fast graph-based segmentation algorithm in [6] for producing superpixels for both query and retrieved images. Different from traditional methods, we harvest superpixels of retrieved images from multiple scales. This increases the chance to find good matches for the query superpixel at a controllable computational cost. In experiments, we segment the retrieved images in four scales by varying the k value $k = 50, 100, 200, 400$ in [6]. The smaller k means fine-scale segmentation while the larger k means coarse-scale segmentation. Note that many superpixels from multi-scale segmentation may include labels from different classes, and cannot be assigned a single category label. We thus screen the superpixels by checking their label purity, which is defined as the percent of label majority. We assign a semantic label y_i to a superpixel s_i if its label purity is greater than 95%; otherwise, we remove it from retrieval set. For the query image, we segment it in the finest scale by setting $k = 50$ to control their purity.

We represent each superpixel by four kinds of features, SIFT histogram, RGB histogram, location histogram and PHOG histogram. We extract SIFT descriptors of four scales per 4 pixels by using VLFeat package [24] and encode them by 5 words from a vocabulary of size 1024 using the LLC algorithm [25]. For each superpixel, we compute a 128-dimensional color histogram by quantizing the RGB features from a vocabulary of 128 color words, and a 36-dimensional location histogram by quantizing the (x,y)-locations into a 6×6 grid. In addition, the 168-dimensional PHOG histogram is extracted from the bounding box of each superpixel in a $1 \times 1, 2 \times 2, 4 \times 4$ pyramid. To incorporate the contextual features into the superpixel representation, we also dilate the superpixel masks by 10 pixels and extract the same four kinds of features in the dilated superpixel regions. We thus obtain a 2712-dimensional $((1024+128+36+168) \times 2)$ feature vector x_i for each superpixel s_i.

We compute the classification cost of each input superpixel $s_i \in \mathcal{Q}$ by its k-nearest neighbors $\mathcal{N}_k(i)$ in retrieval set $\mathcal{R} = \{s_j, x_j, y_j\}$,

$$U(y_i = c|s_i) = 1 - \frac{\sum_{j \in \mathcal{N}_k(i), y_j = c} \mathcal{K}(x_i, x_j)}{\sum_{j \in \mathcal{N}^k(i)} \mathcal{K}(x_i, x_j)}, \quad (1)$$

where $\mathcal{K}(x_i, x_j)$ denotes the intersection kernel between two histogram feature vectors x_i and x_j.

To reduce the computational complexity, we further map feature vectors into a high-dimensional space $\phi(x_i)$ where the inner product approximates the intersection kernel [24],

$$\mathcal{K}(x_i, x_j) \approx < \phi(x_i), \phi(x_j) > \quad (2)$$

3.3. MRF Labeling

We build a four-connected pairwise MRF for semantic labeling. The energy function is given by

$$E(Y) = \sum_p U(y_p = c) + \lambda \sum_{pq} V(y_p = c, y_q = c'), \quad (3)$$

where p, q are pixel indices, c, c' are candidate labels that belong to retrieval label subset $\mathcal{C}'$ and λ is the weight of pairwise energy. The unary energy of one pixel is given by its superpixel,

$$U(y_p = c) = U(y_i = c|s_i), p \in s_i. \quad (4)$$

The pairwise energy on edges is given by spatially variant label cost,

$$V(c, c') = d(p, q) \cdot \mu(c, c'), \quad (5)$$

where $d(p, q) = \exp(-\|I(p) - I(q)\|^2 / 2\sigma^2)$ is the color dissimilarity between two adjacent pixels and $\mu(c, c')$ is the penalty of assigning label c and c' to two adjacent pixels. We define $\mu(c, c')$ by the log-likelihood of label co-occurrence statistics,

$$\mu(c, c') = -\log[(P(c|c') + P(c'|c))/2] \times \delta[c' \neq c] \quad (6)$$

which we estimate from the training images by calculating conditional probabilities $P(c|c')$ of adjacent superpixels. We obtain the semantic labeling by performing MAP inference on $E(Y)$ by alpha-beta swap algorithm [1].

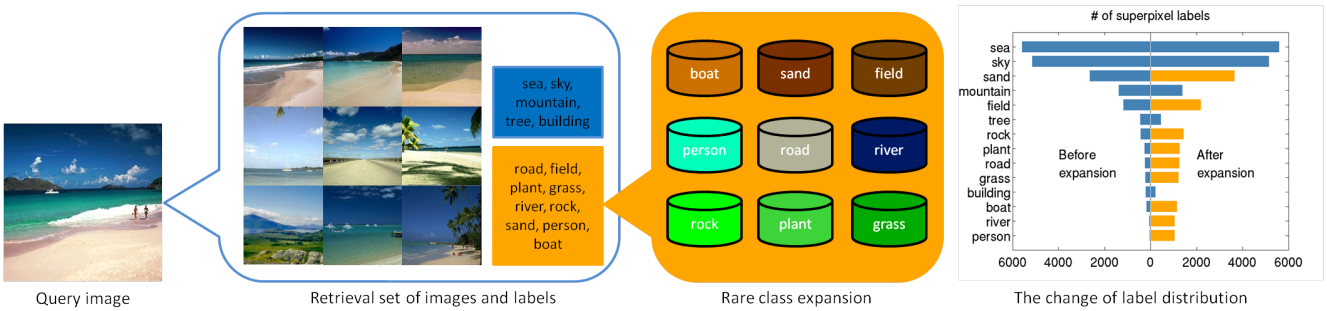

Figure 2. Rare class expansion. The orange bars denotes the rare classes while the blue bars denote the common classes. In this example, we enrich 9 rare classes.

4. Rare Class Expansion

In scene images, the salient regions usually capture the attention of human observers [8], as they provide more information than generic background for scene understanding. It is thus crucial to recognize these objects of interest for generating rich semantic description of images. The saliency property of interesting objects also result in their insufficient representations in the retrieval set. For example, in the "Before expansion" portion of the plot in Figure 2, the superpixels of retrieved images are dominated by sky, sea and sand while boat and people classes are in the very tail of the distribution. This fact brings challenges to recognizing those objects of rare classes.

In this paper, we propose to enrich the retrieved set of superpixels with exemplars of rare classes from the entire database. The label distribution of the retrieval set could be noisy due to the possibly irrelevant images. It is thus difficult to tell if a class in the tail part are interesting objects (boat and people) or simply outliers (building). We instead define "rare" classes by examining the superpixel label distribution of the entire training set. We partition this distribution into head and tail parts based on the 80%-20% Pareto rule, and define the classes in the tail as "rare" $\mathcal{L}_r$ while the other classes in the head as "common" $\mathcal{L}_c$. In Figure 3, we present the superpixel distribution of the SIFT flow training set [15], and our definition of rare and common classes. Given this definition, we can partition the label subset of retrieval superpixels into two parts $\mathcal{L}' = \mathcal{L}'_r \bigcup \mathcal{L}'_c$ and populate the superpixels in the classes of $\mathcal{L}'_r$ with exemplars. Note that the reduced label set remains the same after expansion. In Figure 2, we expand 9 classes (road, field, plant, grass, river, rock, sand, person, boat) and obtain a more balanced, noise resistant superpixel distribution as shown in the "After expansion" portion of the plot.

4.1. Building a Dictionary of Exemplar Superpixels

We build a dictionary $\mathcal{D}_c$ of exemplar superpixels for each class $c \in \mathcal{L}$. We project superpixel feature vectors x_i into a low-dimensional space by PCA and cluster them into 1000 centers by using k-means. We select those superpixels

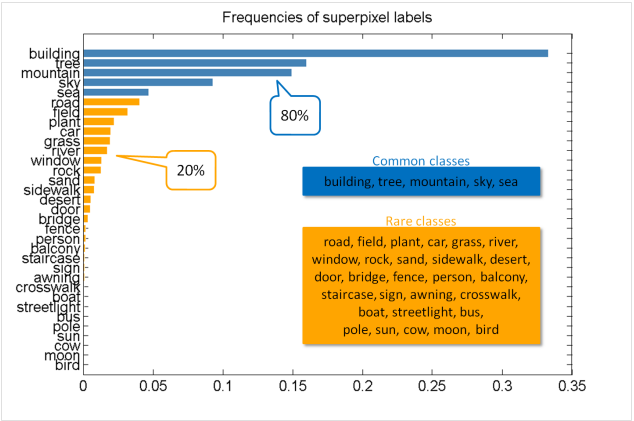

Figure 3. The long tailed superpixel label distribution on the SIFT-flow training set. The orange bar denotes the rare classes while the red bars denote the common classes.

which are closest to the centers as exemplars of particular class. Note that we use this method to build dictionaries for its simplicity, although a more sophisticated algorithm such as in [14] could help to mine more discriminative and diverse exemplars.

4.2. Superpixel Classification

We classify the superpixels of query image $s_i \in \mathcal{Q}$ by using both the retrieval set $\mathcal{R}$ and auxiliary set of rare class exemplars $\mathcal{D}_c, c \in \mathcal{L}'_r$. Similar to our base system in Section 3.2, we compute the classification cost of one query superpixel s_i by its κ-nearest neighbors $\mathcal{N}'_k(i) \in \{\mathcal{R} \bigcup \mathcal{D}_c, c \in \mathcal{L}'_r\}$,

$$U_1(y_i = c|s_i) = 1 - \frac{\sum_{j \in \mathcal{N}'_k(i), y_j = c} \mathcal{K}(x_i, x_j)}{\sum_{j \in \mathcal{N}'^k(i)} \mathcal{K}(x_i, x_j)}. \quad (7)$$

We set $\kappa = 37$ through all the experiments. To increase the classification accuracy of κ-NN, systems in [4, 20] learn weights for superpixels feature vectors. In this work, we choose to hybridize the κ-NN classifier with an SVM classifier [3]. We train the SVM classifier only using the exemplars in our dictionary. Given this small and balanced set of training samples, we can train linear SVM classifiers

$\{w_c, b_c\}, c \in \mathcal{L}$ in a very efficient way [2]. The SVM classification cost for the query superpixel $s_i \in \mathcal{Q}$ is given by $U_2(y_i = c|s_i) = - < w_c, \phi(x_i) > -b_c$. The total superpixel classification cost is thus computed by combining the κ-NN cost and the SVM cost,

$$U(y_i = c|s_i) = \alpha U_1(y_i = c|s_i) + (1-\alpha) U_2(y_i = c|s_i), \quad (8)$$

where α is the combination coefficient.

5. Semantic Context

Context is an important source of information for scene labeling. Although there are many ways to explore this information, we choose to a simply, yet effective feedback mechanism based approach, inspired by [4, 20, 23]. In the base system with rare class expansion, we transfer semantic labeling information from the database to the input image through image retrieval and superpixel classification. In this feedforward process, we obtain the initial semantic knowledge of the input image that are represented by the pixel-wise classification likelihood maps.

$$\ell(p, c) = \frac{1}{1 + \exp(U(y_p = c))}, c \in \mathcal{L}', \quad (9)$$

where $U(y_p = c)$ is the cost of assigning label c to pixel p in (4) and $\mathcal{L}' \subset \mathcal{L}$ is the candidate label subset. These classification maps in the reduced label set grant us naturally sparse representation of semantic information without an extra sparse coding step [9]. The question is how we can use this initial result as a feedback to reinforce the labeling process, in particular the two key components, image retrieval and superpixel classification. First, the likelihood maps in (9) can serve as global context, which has the potential to improve the appearance based image retrieval with semantic scene description. Second, we can generate local semantic descriptors from the likelihood maps for superpixels. Together with local appearance descriptors, we can achieve more contextually consistent classification results. We introduce the algorithms to construct both global and local context descriptors from the likelihood maps in (9) below. Note that to compare semantic context between the query and the database images, we compute the classification likelihood maps for all the training images in a leave-one-out fashion.

5.1. Global Context Descriptor

We define the global context of an image as the spatial layout of semantic content in multi-scale. To this end, we partition an image into a three-layer spatial pyramid $\{I = \bigcup_i \Omega_i^l, l = 1, 2, 3, i = 1, ..., 2^{l-1}\}$, and for each cell Ω_i^l, we compute its $|\mathcal{L}| \times 1$ sparse context vector $\mathbf{z}_i^l = [z_{ic}^l]_{c=1,2,...,|\mathcal{L}|}$ by max pooling of classification likelihood maps,

$$z_{ic}^l = \begin{cases} \max_{p \in \Omega_i^l} \ell(p, c) & \text{if } c \in \mathcal{L}'; \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

The global context descriptor is thus formed by concatenating the sparse vectors from all the cells $\mathbf{z} = [\mathbf{z}_i^l]_{l=0,1,2, i=1,...,2^{l-1}}$. Figure 4 demonstrates the process of computing the global context descriptor of one query image. We use global context descriptor $\mathbf{z}$ to update the image similarity between the query image and database images, $m'(I_q, I_d) = m(I_q, I_d) + < \mathbf{z}_q, \mathbf{z}_d >$ and obtain a new set of retrieved images.

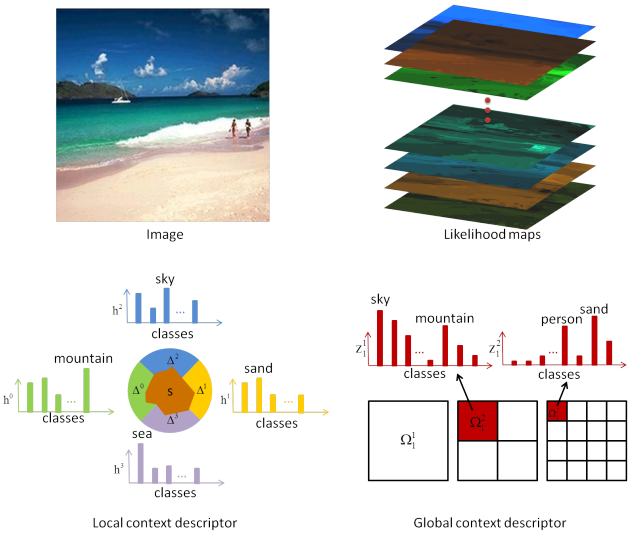

Figure 4. Computing global and local context descriptors from the likelihood maps. The bottom-right figure shows the method of constructing global context descriptor using a three-layer spatial pyramid. In the bottom-left figure, the local context descriptor of one superpixel s (shown in orange) is computed from four-directional neighborhoods: top (blue), bottom (purple), left (green) and right (yellow).

5.2. Local Context Descriptor

We describe superpixels by their local context for robust matching. For each superpixel s_i, we divide its neighborhood into left, right, top, bottom four cells $\{\Delta_i^0, \Delta_i^1, \Delta_i^2, \Delta_i^3\}$ as illustrated in Figure 4, and for each cell Δ_i^j, we compute its $|\mathcal{L}| \times 1$ sparse context vector $\mathbf{h}_i^j = [h_{ic}^j], c = 1, 2, ..., |\mathcal{L}|$ by the same operation as for global context descriptors,

$$h_{ic}^j = \begin{cases} \max_{p \in \Delta_i^j} \ell(p, c) & \text{if } c \in \mathcal{L}'; \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

We represent the superpixel s_i by concatenating the visual feature vector $\mathbf{x}_i$ and spatial context descriptor $\mathbf{h}_i = [\mathbf{h}_i^0; \mathbf{h}_i^1; \mathbf{h}_i^2; \mathbf{h}_i^3]$. Therefore, we can classify superpixels of the query image using the same procedure described in Section 4, but with new feature vectors $\phi([\mathbf{x}_i; \mathbf{h}_i])$.

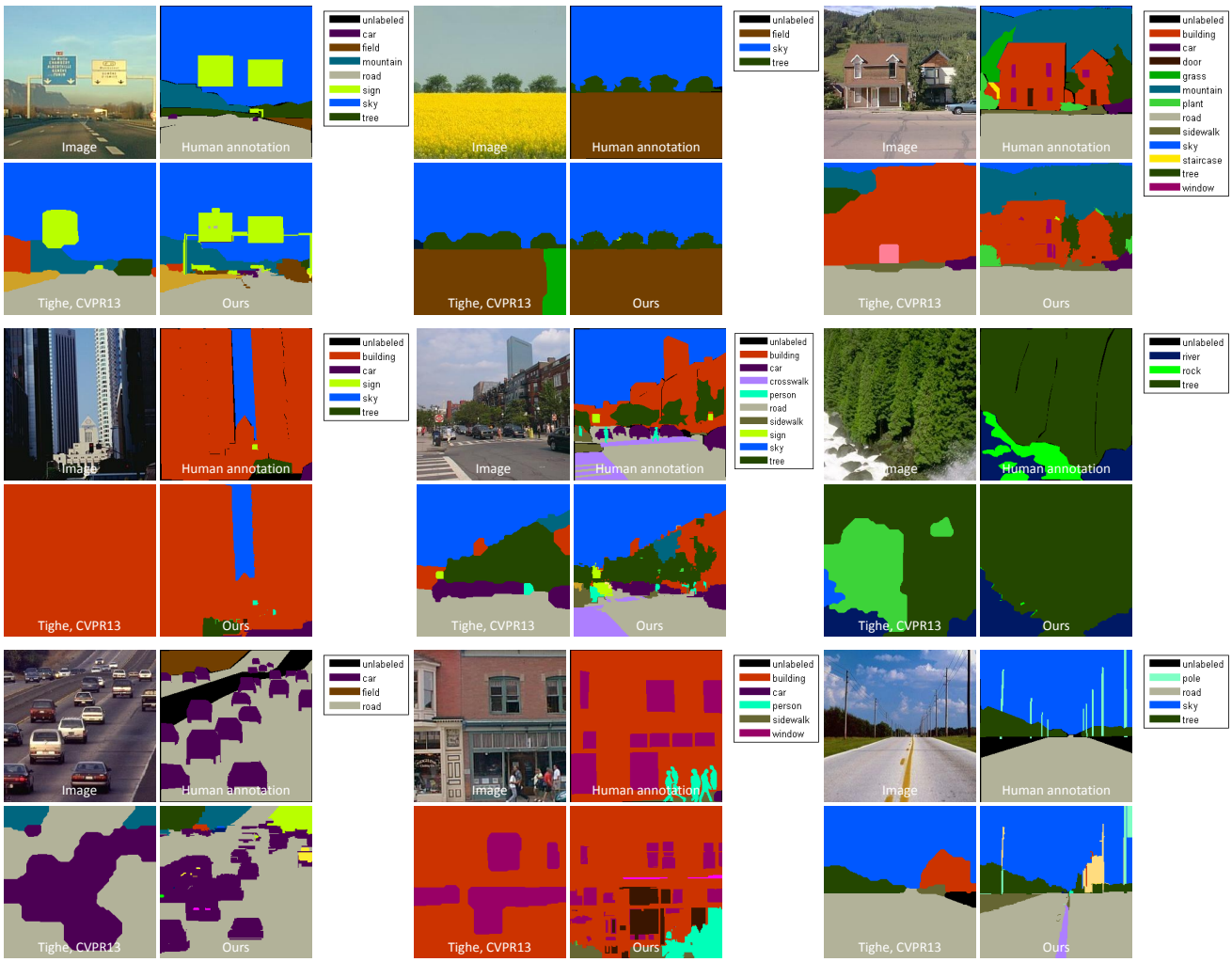

Figure 5. Some representative scene parsing results on the SIFTflow dataset

6. Experimental Results

6.1. 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 (Fig. 3). We set $\alpha = 0.7$ to combine the κ-NN and SVM classifiers in (8), and set $\lambda = 6$ for the pairwise term of MRF energy function in (3). We compare our results with recent work in Table 1. Compared to nonparametric methods, our method (79.8%) outperforms the state-of-the-art per-pixel rate (79.2%) in [20], and per-class rate (39.2%) in [21] by a large margin (8.5%). The state-of-the-art learning based method in [5] can achieve close per-class rate (46.0%) to our system at a cost of more than 4% performance drop on per-pixel rate (74.2%). In contrast, we achieve overall performance improvement using the proposed rare class expansion and semantic context descriptors. We present some qualitative results in Fig. 5. In the bottom of Table 1, we evaluate the contributions of important components to our system: rare class expansion

Table 1. Comparing accuracy (%) on the SIFTflow dataset. Note that in our results, Full=baseline+RCE+LCD+GCD.

SIFTflow	Per-pixel	Per-class
Liu et al. [15]	76.7	N/A
Farabet et al. [5]	78.5	29.5
Farabet et al. [5] balanced	74.2	46.0
Eigen et al. [4]	77.1	32.5
Singh and Kosecka [20]	79.2	33.8
Tighe and Lazebnik [22]	77.0	30.1
Tighe and Lazebnik [21]	78.6	39.2
Full	79.8	48.7
baseline + RCE + LCD	79.4	46.9
baseline + RCE	78.4	45.4
baseline	78.0	27.5

(RCE), local context descriptors (LCD) and global context descriptors (GCD). The results show that rare class expansion plays a central role on per-class rates while semantic context boosts the system performance in general. To further investigate the influence of the rare classes, we compare the performance of our full system with [21] only on the 28 rare classes. Table 2 shows that our system outperforms the state-of-the-art by more than 10% for both per-pixel and per-class rates on those rare classes.

Table 2. Accuracy (%) on the 28 rare classes of SIFTflow dataset.

Rare classes	Per-pixel	Per-class
Tighe and Lazebnik [21]	48.8	29.9
Our full system	59.4	41.9

Runtime. For this dataset, it takes <1 sec. to retrieve relevant images, ~ 5 sec. for feature loading, ~ 5 sec for superpixel classification, and <1 sec. to solve the MRF. The use of context descriptors doubles the classification time.

6.2. 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 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 set $\alpha = 0.9$ to trade-off the KNN and SVM classifiers, and set $\lambda = 6$ for the pairwise term in MRF energy function. We compare our results with recent work in Table 3. As we focus on the rare classes, our method indeed produces the superior per-class result 18.0% to the previous one 15.2%, while remaining competitive on the per-pixel rate 60.6% vs. 61.4% in [21]. By looking at the accuracy on outdoor (65.4 per-pixel, 17.7 per-class) and indoor (41.8 per-pixel, 16.1 per-class) separately, we observe our system loses the per-pixel performance mainly on indoor images (4.5% lower than [21] vs. 0.1% lower than [21] for outdoor). In Table 4, we present the results on the 185 rare classes. It turns out that our system outperforms the state-of-the-art for both per-pixel and per-class rates, which further demonstrates our contributions to rare class boosting. We present some qualitative results in Figure 6.

Table 3. Comparing accuracy (%) on the LMSun dataset.Note that in our results, Full=baseline+RCE+LCD+GCD.

	Per-pixel	Per-class
Tighe and Lazebnik [22]	54.9	7.1
Tighe and Lazebnik [21]	61.4	15.2
Full	60.6	18.0
baseline + RCE + LCD	59.4	17.8
baseline + RCE	57.1	14.5
baseline	58.5	9.0

Table 4. Accuracy (%) on the 185 rare classes of LMSun dataset.

Rare classes	Per-pixel	Per-class
Tighe and Lazebnik [21]	19.0	12.9
Our full system	26.4	14.4

Runtime. Scene parsing is more expensive on the LMSun dataset than the SIFTflow dataset. It takes <20 sec to retrieve relevant images, ~ 60 seconds for feature loading, ~ 60 sec for superpixel classification, and ~ 60 sec to solve MRF for an 600×800 image with 50-100 labels. Using context descriptors doubles the superpixel classification time.

6.3. Discussion

Image retrieval has significant influence on our system. Superpixel matching and MRF inference becomes much easier to solve within a compact set of relevant images to the query; on the contrary, we notice most of the failure cases are caused by incorrect retrieval. We plan to investigate more effective image retrieval techniques, such as convolutional neural networks [10].

Our system faces challenges in indoor scenes. Indoor scenes are usually composed of many man-made 3D objects (bed, cabinet, table, chairs) and thus have more line structures than textures. Our SIFT based superpixel representation becomes less applicable in this scenario, compared to the HOG feature used in object detectors [21]. We plan to develop better indoor object representations by exploring their 3D geometric structures.

The runtime efficiency is one of the most important factors in large scale problems. On one hand, we plan to incorporate hashing algorithms to accelerate superpixel feature loading and matching; on the other hand, we plan to investigate faster MRF inference algorithms.

7. Conclusions

We have presented a novel scene parsing algorithm, which can operate in large scale. By investigating the roles of rare classes in the database, we have proposed two novel techniques: rare class expansion and local/global semantic context descriptors, which are able to significantly boost the per-class performance of our system. Based on that, we have achieved the state-of-the-art results on the SIFTflow and the large scale LMSun datasets.

Acknowledgements

This work is done partially when the first author was an intern at Adobe. The work is supported in part by NSF CAREER Grant #1149783 and NSF IIS Grant #1152576, and a gift from Adobe. We thank Zhe Lin for helpful discussions and the image retrieval code, Joseph Tighe for providing us his results for comparisons, and the CVPR reviewers for their feedback on this work.

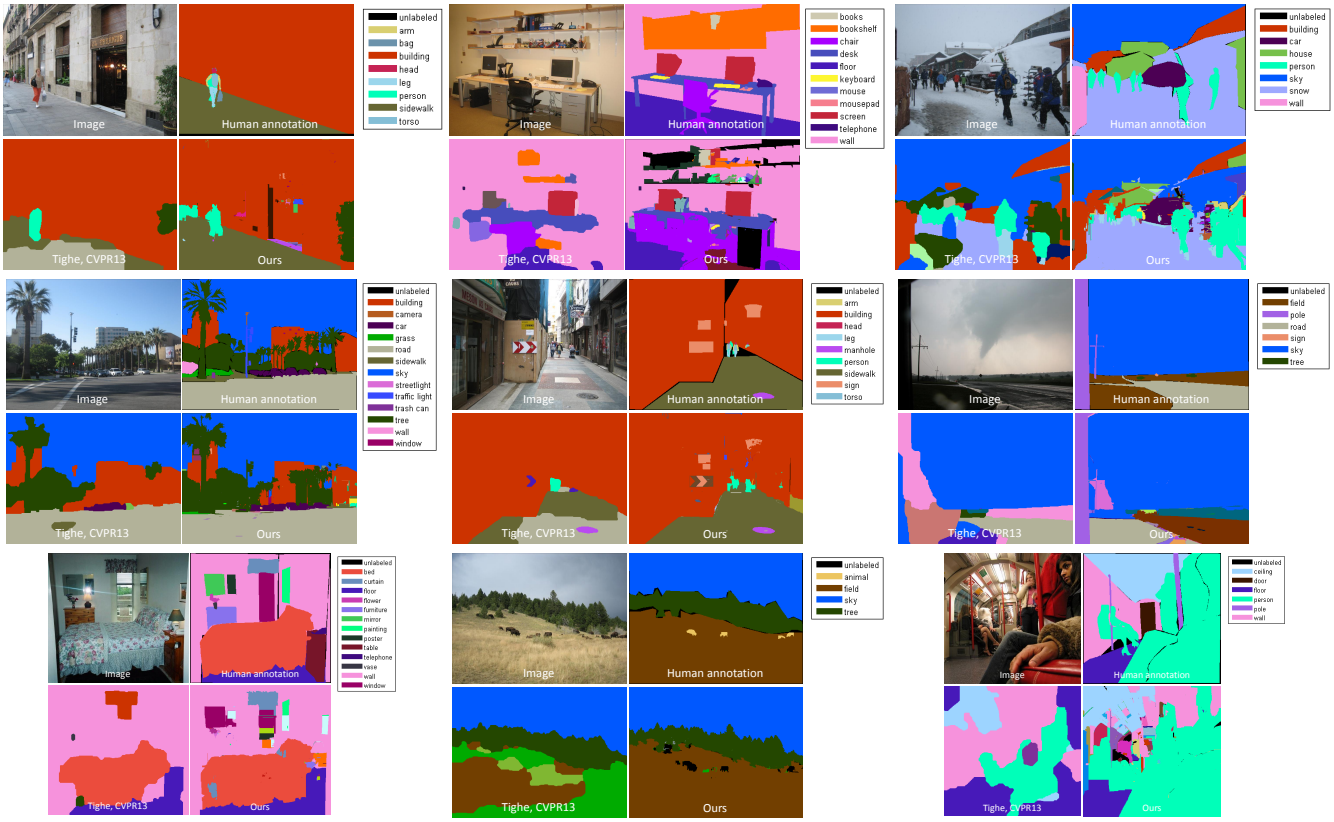

Figure 6. Some representative scene parsing results on the LMSun dataset

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222 – 1239, 2001.

[2] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[3] P. Chaudhuri, A. K. Ghosh, and H. Oja. Classification based on hybridization of parametric and nonparametric classifiers. *PAMI*, 31(7):1153 – 1164, July 2009.

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

[5] 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.

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Graph-Based Image Segmentation. *IJCV*, 59(2), 2004.

[7] Q. Huang, M. Han, B. Wu, and S. Ioffe. A hierarchical conditional random field model for labeling and images of street scenes. In *CVPR*, 2011.

[8] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene anaysis. *PAMI*, 20(11):1254–1259, 1998.

[9] L. jia Li, H. Su, E. P. Xing, and L. Fei-fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[11] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative Hierarchical CRFs for Object Class Image Segmentation. In *ICCV*, 2009.

[12] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[14] Z. Liao, A. Farhadi, Y. Wang, I. Endres, and D. Forsyth. Building a dictionary of image fragments. In *CVPR*, 2012.

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

[16] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[17] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *CVPR*, 2007.

[18] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn reranking. In *CVPR*, 2012.

[19] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *IJCV*, 81(1):2–23, 2009.

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

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

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

[23] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *PAMI*, 32(10):1744 – 1757, 2010.

[24] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[25] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.