Reverse Image Segmentation: A High-Level Solution to a Low-Level Task

Jiajun Wu
http://jiajunwu.com
CSAIL
Massachusetts Institute of Technology
Cambridge, MA, USA

Jun-Yan Zhu
http://www.eecs.berkeley.edu/~junyanz
Computer Science Division
University of California, Berkeley
Berkeley, CA, USA

Zhuowen Tu
http://pages.ucsd.edu/~ztu
Department of Cognitive Science
University of California, San Diego
La Jolla, CA, USA

Abstract

Image segmentation is known to be an ambiguous problem whose solution needs an integration of image and shape cues of various levels; using low-level information alone is often not sufficient for a segmentation algorithm to match human capability. Two recent trends are popular in this area: (1) low-level and mid-level cues are combined together in learning-based approaches to localize segmentation boundaries; (2) high-level vision tasks such as image labeling and object recognition are directly performed to obtain object boundaries. In this paper, we present an interesting observation that performs image segmentation in a reverse way, *i.e.*, using a high-level semantic labeling approach to address a low-level segmentation problem, could be a proper solution. We perform semantic labeling on input images and derive segmentations from the labeling results. We adopt graph coloring theory to connect these two tasks and provide theoretical insights to our solution. This seemingly unusual way of doing image segmentation leads to surprisingly encouraging results, superior or comparable to those of the state-of-the-art image segmentation algorithms on multiple publicly available datasets.

1 Introduction

Image segmentation is a fundamental and widely studied problem in computer vision [2, 7, 14, 37]. Continuous efforts have been made to improve the performance of segmentation systems to match human capability [2]; however, it is generally acknowledged that solving the segmentation problem with low-level cues alone might not be possible [27]. There has long been a discussion on solving this seemingly low-level task with high-level knowledge [9], but a clear and concrete solution is not yet available.

Accepting non-perfect segmentation results allows one to use cutting-edge systems [7, 14, 37] to generate over-segmentations (or superpixels) in helping higher-level vision tasks such grouping and labeling. Some representative methods include [18, 30, 34, 39] for object recognition, image labeling, and parsing.

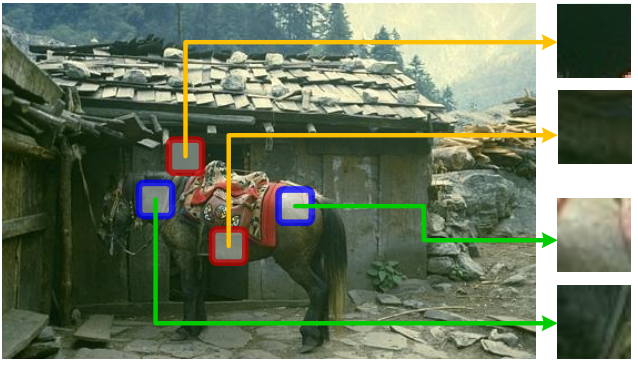

Figure 1: The difficulty in segmentation due to the lack of semantic knowledge: the two patches are from different segments but share similar low-level features, while the bottom two patches belong to the same segment despite of their distinct appearances.

The task of semantic labeling, considered as a high-level one, has been heavily studied recently in computer vision [19, 38]. During training, one is given a set of images, together with their corresponding per-pixel labels where the total number of label category is predefined; during testing, learned models are used to predict per-pixel label in each image. Arguably, research domain along this vein is called "closed universe" [39] in which expanding the labels to include more categories can be a difficult task. Nonparametric models were instead proposed to perform label transferring [25, 39] in which large public datasets such as LabelMe [35] are utilized to break out of the "closed universe". Instead of training class-specific classifiers, these nonparametric models transfer semantic information to a test image from similar training images. The label transferring process is proved effective and also adaptive to new images. However, the problems of these methods include computational burdens in retrieving similar images and the limitation in generalization when processing images of large variations.

As shown in Figure 1, two main issues (both due to the lack of semantic understanding) contribute to the main difficulty in image segmentation: (1) regions of different appearances might belong to the same segment, (2) and different image segments might have identical local appearances. In this paper, we propose to perform image segmentation in a reverse way. Unlike previous systems where either specific high-level information (*e.g.*, human faces, text, shape) are integrated into segmentation systems [9, 41] or segmentation results are used as a precursor for high-level vision tasks [30, 34], our method takes a path of a high-level segmentation approach: at first per-pixel labeling of semantic categories is performed, followed by a procedure to obtain segmentations with per-pixel labels got discarded in the end. We are inspired from the observation that semantic labels give means of differentiating similar pixels and grouping dissimilar pixels. These labels can be viewed as a quantization of the solution space of segmentation, and the derived segmentations are mostly consistent even when the semantic level labels are not completely correct. For example, in Figure 2, a mammal is classified as a bird because of their similarity in color and texture, but the derived segmentation is mostly correct.

Here we bring a novel top-down view towards segmentation and demonstrate our idea on

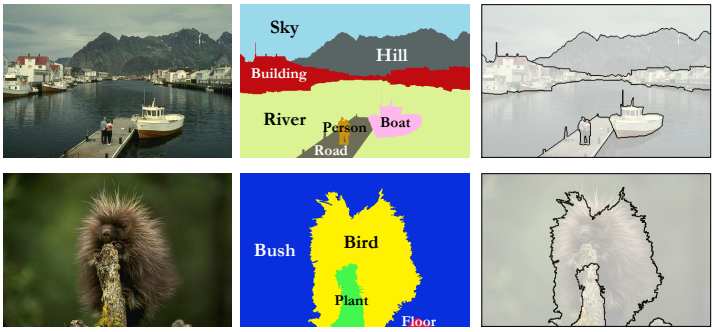

Figure 2: Example images and their semantic labeling and image segmentation results. Even if the semantic labels are not perfect, our pipeline could obtain satisfactory segmentation results.

both parametric and nonparametric formulations, resulting in simple, generic, and efficient algorithms. In particular, we find that parametric image labeling methods might be suitable for the task of image segmentation. Parametric models are usually fast to train and test; a certain level of knowledge abstraction, which is only available in parametric models, is often needed in order to have good generalization capability. Although a parametric approach using a fixed number of categories is considered as a "closed universe" solution, we show, both theoretically and empirically, that the semantic knowledge extracted from this "closed universe" might be a useful direction to obtain accurate segmentation results. Note that graph coloring theory can provide heuristics for the number of semantic categories needed. We validate the effectiveness of our approach on standard benchmarks without retraining, which demonstrates strong resistance to the dataset bias problem [40].

2 Related Work

Image segmentation is an important problem. Some popular algorithms include Mean Shift [7], Normalized Cuts [37], and graph-based methods [14]. Recent efforts can be divided into different categories: methods to learn more reliable affinities for spectral segmentation [23], effective ways of performing cue combination [2], and saliency-guided segmentation [12].

For semantic labeling or class segmentation, popular methods include those estimating pixel-wise class labels with contextual information [19, 38], those using predefined superpixels or segmentation regions [17, 18], and those predicting the boundaries or bounding boxes of the objects in images [8]. Most of these methods learn classification models and are therefore restricted to predefined classes, or so-called "closed universe" [39]. Recently, some nonparametric approaches have also been proposed to break out of the "closed universe" [25].

Some recent approaches [15, 21, 23] learn internal structures, especially learning affinity functions between pairs of pixels or superpixels, for images segmentation. Ren et al. [33] adopt a classification model for image segmentation. Nonetheless, their system focuses on evaluating the *goodness* of image-level segmentations via local features. Cheng et al. [6] learn a classifier to discriminate unstructured objects, or backgrounds, but they do not attempt

to learn foreground objects. Further, none of them utilize the power of semantic knowledge from images.

A recent study [18] investigates holistic image understanding. However, their emphasis is on the correlations among various vision tasks. Neither do they focus on unsupervised segmentation, nor do they discuss the idea of incorporating high-level knowledge into low-level vision tasks. Perhaps the most relevant work is [24], which also exploits high-level semantic information for other vision tasks. However, the semantic knowledge is used there to assist depth estimation, while our work addresses segmentation.

3 From Labeling to Segmentation

A straightforward way of performing image segmentation by labeling methods is to first follow the nonparametric approaches [25, 39] to find the most similar images, transfer semantic information, and then discard the labels. The segmentation results obtained this way are already acceptable, although it takes several minutes to process a single image. In this section, we discuss our approach which also uses high-level semantic knowledge for general purpose segmentation, but in a parametric way. Our algorithm is simple but generic, and produces convincing results superior or comparable to those of multiple state-of-the-art algorithms.

3.1 Learning Generic Semantic Knowledge

Recently Tighe et al. released LM+SUN dataset as part of the SuperParsing project [39], which combines the SUN dataset [44] with a complete download of LabelMe [35]. The dataset consists of 45,676 images, of which 21,182 are indoor and 24,494 are outdoor. They also used manual synonym correction to obtain 232 semantic labels.

The LM+SUN dataset can serve as a large-scale semantic knowledge base, which provides generic high-level information. To utilize this knowledge, we train a discriminative multi-class classifier on top of the superpixels of the outdoor images in the LM+SUN dataset, which we found to be sufficient for the task of general image segmentation.

Specifically, we first assign each superpixel a semantic label. Following [39], a superpixel is associated with a semantic class if and only if at least half of the superpixel overlaps with a ground truth segment mask with that label. Then, according to the label frequencies on superpixels, 50 most frequent classes are picked out. For each class, 20,000 superpixels of the class are sampled as positive training examples, and another 20,000 superpixels unlabeled or with other class labels are randomly drawn as negative examples; a linear SVM [13] is then trained on the data. These classifiers are generic and applicable to any images including those not in the dataset.

In segmentation, each superpixel is tested by all learned classifiers to obtain a vector of confidence values. Perhaps the easiest way to group the superpixels into different segments based on these values is to assign each superpixel the label whose corresponding classifier reports the highest value, and then directly discard the labeling information. The experiments show that such a simple approach has already achieved encouraging results.

The LM+SUN dataset is designed to be generic enough to break out of the "closed universe" [35], which indicates that the knowledge learned from it should be comprehensive and widely applicable. In other words, the performance of our approach will not be restrained by the limitation of the training data. Further, the models, once learned, could be directly

applied to other datasets including popular segmentation benchmarks. This makes our approach, although seemingly supervised, ould effectively segment natural images without re-training, in this sense resembling unsupervised methods.

3.2 Our Formulation

Bottom-up unsupervised segmentation algorithms with only low-level features may not be able to capture the semantic knowledge which is needed to achieve a satisfactory segmentation [9, 41]. Similarly, purely using top-down semantic knowledge is probably not enough for reasonable segmentation results. For instance, properly incorporating low-level cues helps us to divide multiple connected instances of the same category to different segments. Besides, as it is generally acknowledged that context is necessary for satisfactory parsing results [38], we would also like to enforce contextual constraints to refine the results. To this end, we formulate the problem under the framework of Conditional Random Fields (CRF). Constraints that allow us to reduce over/under segmentations near region boundaries are encoded as pairwise edge potentials.

Denoting $S = \{s_i\}$ as a set of superpixels and $G(S,E)$ as an adjacency graph, the probability of class labels $\boldsymbol{c} = \{c_i\}$, given the set S and weights λ, μ, can be formulated as

$$-\log(\Pr(\boldsymbol{c}|G;\lambda,\mu)) = \sum_{s_i \in S} \Phi(c_i|s_i) + \sum_{(s_i,s_j)\in E} [\lambda \Psi(c_i,c_j) + \mu \Theta(c_i,c_j|s_i,s_j)]. \quad (1)$$

The unary potentials Φ are directly defined as the probability output of our multi-class classifier: $\Phi(c_i|s_i) = -\log(\Pr(c_i|s_i))$. Similar to [39], the first binary potentials Ψ are defined as probabilities of label co-occurrence: $\Psi(c_i,c_j) = -\log[(\Pr(c_i|c_j) + \Pr(c_j|c_i))/2] \cdot \delta[c_i \neq c_j]$, where $\Pr(c_i|c_j)$ is the conditional probability of one superpixel having label c_i given that its neighbor has label c', estimated from the training set, and $\delta[\cdot]$ is the indicator function.

The second pairwise terms Θ are similar to those in [17, 38]

$$\Theta(c_i,c_j|s_i,s_j) = \left(\frac{W(s_i,s_j)}{1+\|s_i - s_j\|}\right) \cdot \delta[c_i \neq c_j], \quad (2)$$

where $\|s_i - s_j\|$ is the L_2 difference between the feature vectors of superpixels s_i and s_j, and $W(s_i,s_j)$ is the normalized shared boundary length. W can be formulated as $W(s_i,s_j) = [L(s_i)^{-1} + L(s_j)^{-1}] \cdot L(s_i,s_j)$, where $L(s_i)$ is the length of boundary of superpixel s_i, and $L(s_i,s_j)$ is the shared boundary length between s_i and s_j.

In Eq. (2), $W(s_i,s_j)$ is used instead of $L(s_i,s_j)$ as the regularization term which discourages small isolated regions. This is because for superpixels with different sizes, using $L(s_i,s_j)$ leads to an overemphasis on the connectivity between large neighboring superpixels, *e.g.*, superpixels of the classes "sky" and "tree", and consequently make the algorithm merge superpixels that should not be merged.

There are two parameters λ and μ in our formulation, which represent the effects of high-level contextual information and low-level spatial regularization, respectively. Both of them can be chosen either empirically or by cross validation. Given λ and μ, we adopt Markov Chain Monte Carlo methods for inference. Because the CRF is built on superpixels, the inference is highly efficient, taking approximately 0.1 second per image on average. We finally discard the semantic labels produced by CRF to obtain segmentations.

4 A View in Graph Theory

Here we study the tasks of semantic labeling and image segmentation from the perspective of graph coloring. We demonstrate that results in graph coloring offer theoretical insights and practical heuristics for our solution.

We first consider a planar graph G_1 whose vertices represent the segments in the ground truth and whose edges connect all adjacent segments, as shown in Figure 6. Regarding each color as a semantic label, a proper coloring of G_1, in which all adjacent vertices have different colors, inherently corresponds to a segmentation of the original image.

The Four Color Theorem [1] shows that four categories are enough to derive all possible segmentations. However, in our problem, the following constraint needs to be considered:

The same concept, e.g. *tree, may appear in different places in the image, which indicates that separate vertices in G_1 are required to share the same color.*

Incorporating this constraint leads to a popular research topics in graph coloring named the *empire problem*, and Heawood [20] has proved that if the size of any set that share the same color is no larger than m, G_1 can be colored by at most

$$H = \left\lfloor \frac{1}{2} \cdot \left(6m + 1 + \sqrt{(6m+1)^2 - 24\chi} \right) \right\rfloor \tag{3}$$

colors, where $\lfloor \cdot \rfloor$ is the floor function and χ is the Euler characteristic of the surface of G_1. For planar graphs, $\chi = 2$.

The other way of building a graph from an image is to regard each superpixel as a vertex. We name the graph derived in this way G_2. Because adjacent superpixels may belong to the same semantic class, and thus share the same color, it is no longer appropriate to add edges between all adjacent superpixels. Instead, we assume that any two vertices are directly connected with probability $c(n)/n$, where n is the number of vertices and $c(n)$ is a function of n. For edge probabilities $c(n)/n$ where $c(n) = np$ is a linear function of n, Bollobás et al. [2] showed that with probability one the graph $G(n,p)$ satisfies

$$\left(\frac{1}{2} - o(1)\right) \log \frac{1}{1-p} \frac{n}{\log n} \leq H(G_2(n,p)) \leq \left(\frac{1}{2} + o(1)\right) \log \frac{1}{1-p} \frac{n}{\log n} \tag{4}$$

as $n \to \infty$, where $o(1)$ denotes a function of n converging to zero as $n \to \infty$.

These results in coloring theory loosely justify our solution of deriving segmentations from labelings, and also, as shown in Section 5.4, they provide heuristic values for the number of categories needed.

5 Experiments

5.1 Setup

The proposed image segmentation framework is tested both with and without the high/low-level pairwise potentials, resulting in four variants (RIS, RIS+H, RIS+L, RIS+HL). For completeness, we also evaluate the segmentations derived from the outputs of a state-of-the-art nonparametric semantic labeling system (SuperParsing) [39].

Our method use superpixels. There are many algorithms to obtain a superpixel initialization, including FH [14], Mean Shift [7], and those from Ren et al. [30, 33]. Here we use

Methods	PRI ↑	VoI ↓	GCE ↓	BDE ↓
RIS+HL	**0.8137**	1.8232	0.1805	**13.07**
RIS+H	0.8052	1.9233	0.1952	13.16
RIS+L	0.8003	1.9054	0.2012	13.37
RIS	0.7871	2.0597	0.2199	13.78
SuperParsing	0.7628	2.0387	0.2178	15.05
MShift	0.7958	1.9725	0.1888	14.41
NCuts	0.7242	2.9061	0.2232	17.15
JSEG	0.7756	2.3217	0.1989	14.40
FH	0.7139	3.3949	**0.1746**	16.67
MNCuts	0.7559	2.4701	0.1925	15.10
NTP	0.7521	2.4954	0.2373	16.30
Saliency	0.7758	**1.8165**	0.1769	16.24
SpectClust	0.7357	2.6336	0.2469	15.40
ROI-Seg	0.7599	2.0072	0.1846	22.45
Affin	—	—	0.2140	—
Seed	—	—	0.2090	—

Table 1: Evaluations on BSDS. We highlight the best algorithm for each measure.

Methods	PRI ↑	VoI ↓
RIS+HL	**0.78**	**1.29**
RIS+H	0.75	1.35
RIS+L	0.76	1.36
RIS	0.73	1.42
SuperParsing	0.71	1.40
gPb-owt-ucm	**0.78**	1.68
C-Cluster-P	0.77	1.65
C-Cluster-H	**0.78**	1.59
Joint-kernel	**0.78**	1.62
TBES	0.76	1.49
MNCuts	0.63	2.77

Table 2: Evaluations on MSRC with both supervised and unsupervised methods. We highlight the best algorithm for each measure.

Mean Shift with multiple sets of parameters to generate superpixels with various granularities. The numbers of superpixels for each image approximately range from 50 to 2,000.

We use the same visual descriptors as those in [39] on superpixels. These features generally encode the shape, location, texture/SIFT, and color of the superpixels.

Following [2, 23, 32], we conduct experiments on different datasets with multiple widely adopted measures including Probabilistic Rand Index (PRI) [31, 42], Variation of Information (VoI) [28], Global Consistency Error (GCE) [27], Boundary Displacement Error (BDE) [16], and Segmentation Covering (Covering) [2]. Higher values of PRI and Covering and lower values of VoI, GCE, and BDE correspond to more accurate segmentations.

5.2 Berkeley Segmentation Dataset

The Berkeley Segmentation Dataset (BSDS300) [27] consists of 300 natural images, each of which has been manually segmented by multiple human subjects. The dataset has found wide acceptance as a benchmark for image segmentation [2, 12].

In this experiment, we set $\lambda = 0.05$ and $\mu = 40$ for RIS variants if necessary. Four RIS variants and SuperParsing [39] are compared on the BSDS300 with eleven other segmentation algorithms including Mean Shift (MShift) [7], Normalized Cuts (NCuts) [37], JSEG [10], Graph-based Segmentation (FH) [14], Multiscale Normalized Cuts (MNCuts) [8], Normalized Tree Partitioning (NTP) [43], Saliency-based Segmentation (Saliency) [12], Spectral Clustering (SpectClust) [45], MSER-based segmentation (ROI-Seg) [11], pixel affinity based method (Affin) [15], and the seeded graph cuts (Seed) [29]. Following [12, 43], we use PRI, VoI, GCE and BDE as the metrics. The results of the others are taken from [12].

Table 1 shows that the idea of solving low-level segmentation by high-level labeling produces convincing performance both parametrically and nonparametrically, outperforming state-of-the-art algorithms. In particular, the performance of RIS+HL is better than the others by a large margin in both PRI and BDE. Also, SuperParsing [39], as an off-the-

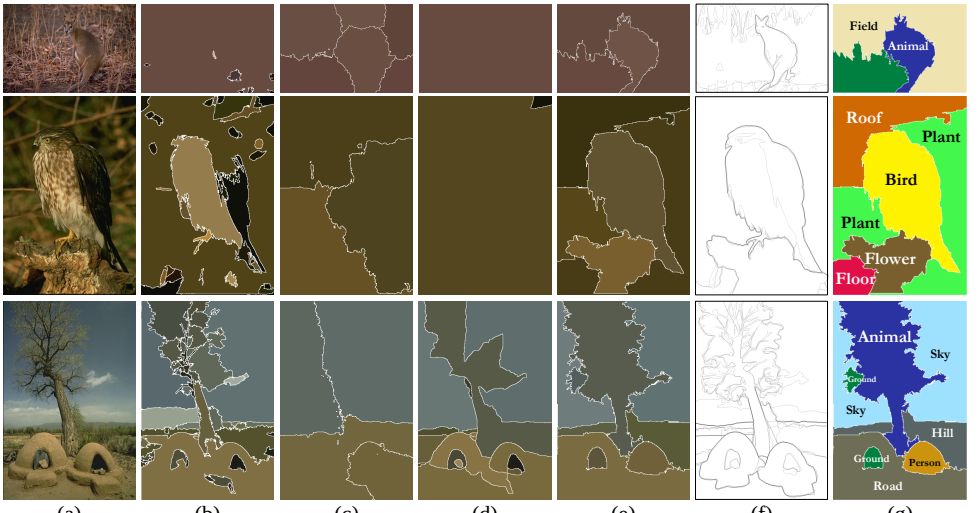

Figure 3: Segmentation results on the Berkeley Segmentation Dataset. From left to right: (a) original image, (b) results of Mean Shift [7], (c) Multiscale Normalized Cuts [8], (d) gPb-owt-ucm [2], and (e) our framework, (f) ground truth segmentations, and (g) our labeling results. As we can see, methods based purely on low-level features like (d) tend to merge patches of similar appearances but different semantics.

shelf image labeling algorithm, achieves comparable performance with traditional methods like Normalized Cut. These results justify the effectiveness of adopting high-level semantic knowledge.

When the BSDS300 was proposed in [27], the 300 images were divided into two groups, 200 for training and 100 for testing. Recently, Arbeláez et al. [2] released the BSDS500 superset, which added 200 images to BSDS300. Because some algorithms report hierarchical segmentations or sets of segmentations, obtaining a single segmentation involves a choice of scale. To have an objective evaluation for these methods, they calculated, for each measure, both *Optimal Dataset Scale (ODS)*, which uses a fixed threshold for the entire dataset, and *Optimal Image Scale (OIS)*, which selects the optimal threshold on a per-image basis.

Here we evaluate our algorithm under the same setting. To obtain multiple segmentations, λ is set from 0 to 0.1 with a step of 0.01 and μ is set from 0 to 200 with a step of 20, which produce $11 \times 11 = 121$ segmentation maps per image. Larger values of λ and μ emphasize more on the smoothness between neighboring superpixels, and typically generate fewer segments. We compare our algorithm with both supervised methods including gPb-owt-ucm [2], fPb-owt-ucm [23], cPb-owt-ucm [23], Correlation Clustering (C-Cluster) [22], and joint kernel learning (Joint-kernel) [21], and unsupervised methods like Canny-owt-ucm (Canny) [2], Segmentation by Weighted Aggregation (SWA) [36], Texture and Boundary Encoding (TBES) [32], Average Dissimilarity (AvDis) [3], Chan-Vese model (ChanVese) [3], a fixed hierarchy of regions (Quad-Tree), and others listed in Table 1.

Again, as shown in Table 3, our high-level solution yields highly competitive results. Specifically, on both datasets, RIS+HL is comparable with [2] and [23] in most measures and consistently better than the others in all measures. As shown in Figure 3, when methods

	BSDS300						BSDS500					
	Covering ↑		PRI ↑		VoI ↓		Covering ↑		PRI ↑		VoI ↓	
	ODS	OIS	ODS	OIS	ODS	OIS	ODS	OIS	ODS	OIS	ODS	OIS
Human	0.73	0.73	0.87	0.87	1.16	1.16	0.88	0.88	1.17	1.17	0.72	0.72
RIS+HL	**0.59**	**0.65**	**0.82**	**0.86**	1.71	1.53	0.57	**0.66**	**0.84**	**0.86**	1.73	1.55
RIS+H	0.55	0.60	0.80	0.84	1.82	1.63	0.53	0.61	0.82	0.84	1.91	1.70
RIS+L	0.57	0.63	0.79	0.82	1.80	1.60	0.55	0.61	0.81	0.84	1.85	1.68
RIS	0.52	—	0.77	—	1.99	—	0.50	—	0.78	—	2.05	—
SuperParsing	0.48	—	0.74	—	2.07	—	0.47	—	0.75	—	2.19	—
gPb-owt-ucm	**0.59**	**0.65**	0.81	0.85	**1.65**	1.47	**0.59**	0.65	0.83	**0.86**	1.69	1.48
fPb-owt-ucm	0.57	0.63	0.80	0.84	1.69	1.49	0.58	0.63	0.82	0.85	1.70	1.50
cPb-owt-ucm	**0.59**	**0.65**	0.81	0.85	1.66	**1.46**	**0.59**	0.65	0.83	**0.86**	**1.65**	**1.45**
C-Cluster-P	—	—	0.81	—	1.83	—	—	—	—	—	—	—
C-Cluster-H	—	—	0.81	—	1.74	—	—	—	—	—	—	—
Joint-kernel	—	—	0.79	—	1.90	—	—	—	—	—	—	—
MShift	0.54	0.58	0.78	0.80	1.83	1.63	0.54	0.58	0.79	0.81	1.85	1.64
FH	0.51	0.58	0.77	0.82	2.15	1.79	0.52	0.57	0.80	0.82	2.21	1.87
Canny	0.48	0.56	0.77	0.82	2.11	1.81	0.49	0.55	0.79	0.83	2.19	1.89
MNCuts	0.44	0.53	0.75	0.79	2.18	1.84	0.45	0.53	0.78	0.80	2.23	1.89
SWA	0.47	0.55	0.75	0.80	2.06	1.75	—	—	—	—	—	—
Saliency	0.57	—	0.78	—	1.81	—	—	—	—	—	—	—
TBES	0.54	—	0.78	—	1.86	—	—	—	—	—	—	—
AvDis	0.47	—	0.76	—	2.62	—	—	—	—	—	—	—
ChanVese	0.49	—	0.75	—	2.54	—	—	—	—	—	—	—
Quad-Tree	0.33	0.39	0.71	0.75	2.34	2.22	0.32	0.37	0.73	0.74	2.46	2.32

Table 3: Comparison on the test sets of BSDS300 and BSDS500 with both supervised and unsupervised methods. For each measure, the best algorithm is highlighted.

Figure 4: Segmentation results on MSRC.

based purely on the ambiguous low-level features like [2] tend to merge patches of similar appearances but different semantics, high-level semantic knowledge could help to figure out a correct segmentation. More results are available in the supplementary material.

We demonstrate the capacity of our high-level solution on a simple model and achieve highly competitive results. The performance of our algorithm is subject to further improvement in many aspects, *e.g.*, incorporating recent developments in affinity learning [23].

5.3 MSRC Database

Most of our experiments are done on the BSDS, hitherto the most complete and widely used segmentation benchmark. To prove our generality, or the ability to break out of "closed universe", we also report results on the MSRC Database.

The MSRC Database [33] consists of 591 images with objects grouped into 23 categories. We use the cleaned up segmentations provided in [26] as the ground truth, which are more reasonable and precise than the original data. Following [32], PRI and VoI are used to evaluate segmentation results.

Table 2 compares our solution with TBES [32], gPb-owt-ucm (gPb-owt-ucm) [2], Correlation Clustering (C-Cluster) [22], joint kernel learning (Joint-kernel) [21], and Multiscare Normalized Cuts [8]. Our method outperforms all other methods.

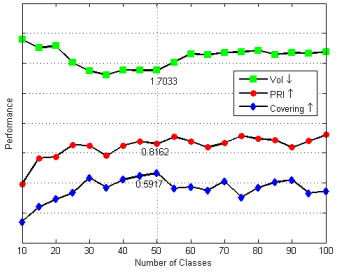

Figure 5: Performance with a varying number of classes.

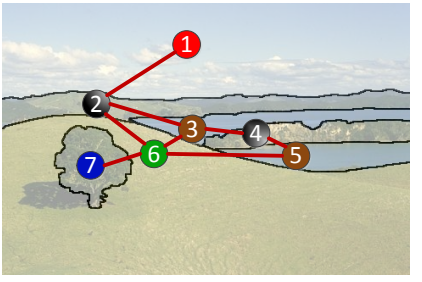

Figure 6: A view in graph coloring theory. Adjacent regions are assigned different colors for their different semantic labels.

5.4 Evaluation on the Number of Categories

Following the theory discussed earlier, we evaluate the effect of the number of categories in the framework on the test set of BSDS300, measured in PRI, VoI, and Covering.

We use the most frequent classes in the evaluation. Figure 5 shows that the performance reaches a peak in VoI and Covering when the number of classes is between 30 and 60. In terms of PRI, the performance tends to be stable when the number of classes is over 40, with a slight tendency of increase. Note that the bound given by the empire problem [20] is 60 when the maximum size of empire $m = 10$, and the bound given by the random graph formulation [4] is 36.46 when the number of superpixels $n = 300$ and the edge probability $p = 0.75$. These correspondences demonstrate that theoretical results do provide close heuristics.

6 Conclusion

In this paper, we observe that a high-level solution performs well in the task of low-level image segmentation. Extensive experiments show that our simple framework achieves convincing results, highly competitive with the state-of-the-art algorithms. We believe that the idea of using high-level knowledge for low-level tasks deserves future attention.

Acknowledgement

This work is supported by NSF IIS-1216528 (IIS-1360566) and NSF IIS-0844566 (IIS-1360568).

References

[1] K. Appel and W. Haken. Every planar map is four colorable. *Illinois Journal of Mathematics*, 21(3):429–567, 1977.

[2] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, 2011.

[3] L. Bertelli, B. Sumengen, BS Manjunath, and F. Gibou. A variational framework for multiregion pairwise-similarity-based image segmentation. *IEEE TPAMI*, 30(8):1400–1414, 2008.

[4] B. Bollobás. The chromatic number of random graphs. *Combinatorica*, 8(1):49–55, 1988.

[5] J. Carreira, F. Li, and C. Sminchisescu. Object recognition by sequential figure-ground ranking. *IJCV*, 98(3):243–262, 2012.

[6] C. Cheng, A. Koschan, C.H. Chen, D.L. Page, and M.A. Abidi. Outdoor scene image segmentation based on background recognition and perceptual organization. *IEEE TIP*, 21(3):1007–1019, 2012.

[7] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE TPAMI*, 24(5):603–619, 2002.

[8] T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In *CVPR*, 2005.

[9] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *IJCV*, 72(2):195–215, 2007.

[10] Y. Deng and BS Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE TPAMI*, 23(8):800–810, 2001.

[11] M. Donoser and H. Bischof. Roi-seg: Unsupervised color segmentation by combining differently focused sub results. In *CVPR*, 2007.

[12] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof. Saliency driven total variation segmentation. In *ICCV*, 2009.

[13] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[14] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[15] C. Fowlkes, D. Martin, and J. Malik. Learning affinity functions for image segmentation: Combining patch-based and gradient-based approaches. In *CVPR*, 2003.

[16] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí. Yet another survey on image segmentation: Region and boundary information integration. In *ECCV*, 2002.

[17] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[18] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009.

[19] X. He, R.S. Zemel, and M.A. Carreira-Perpinán. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

[20] P. J. Heawood. Map colour theorem. *J. Pure Appl. Math*, 24:332–338, 1890.

[21] Jongmin Kim, Youngjoo Seo, Sanghyuk Park, Sungrack Yun, and Chang D Yoo. Joint kernel learning for supervised image segmentation. In *ACCV*. 2013.

[22] Sungwoong Kim, Sebastian Nowozin, Pushmeet Kohli, and Chang Dong Yoo. Higher-order correlation clustering for image segmentation. In *NIPS*, 2011.

[23] Tae Hoon Kim, Kyoung Mu Lee, and Sang Uk Lee. Learning full pairwise affinities for spectral segmentation. *TPAMI*, 35(7):1690–1703, 2013.

[24] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, 2010.

[25] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE TPAMI*, 33(12):2368–2382, 2011.

[26] Tomasz Malisiewicz and Alexei A. Efros. Improving spatial support for objects via multiple segmentations. In *BMVC*, 2007.

[27] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.

[28] M. Meilă. Comparing clusterings: an axiomatic view. In *ICML*, 2005.

[29] B. Mičušík and A. Hanbury. Automatic image segmentation by positioning a seed. In *ECCV*, 2006.

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

[31] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[32] S.R. Rao, H. Mobahi, A.Y. Yang, S.S. Sastry, and Y. Ma. Natural image segmentation with adaptive texture and boundary encoding. In *ACCV*, 2009.

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

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

[35] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1):157–173, 2008.

[36] E. Sharon, M. Galun, D. Sharon, R. Basri, and A. Brandt. Hierarchy and adaptivity in segmenting visual scenes. *Nature*, 442(7104):810–813, 2006.

[37] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8): 888–905, 2000.

[38] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

[39] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*, 2010.

[40] A. Torralba and A.A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

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

[42] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *IEEE TPAMI*, 29(6):929–944, 2007.

[43] J. Wang, Y. Jia, X.S. Hua, C. Zhang, and L. Quan. Normalized tree partitioning for image segmentation. In *CVPR*, 2008.

[44] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[45] S.X. Yu and J. Shi. Multiclass spectral clustering. In *ICCV*, 2003.