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

Jiajun Wu
http://jiajunwu.com
Jun-Yan Zhu
http://www.eecs.berkeley.edu/~junyanz
Zhuowen Tu
http://pages.ucsd.edu/~ztu

CSAIL
Massachusetts Institute of Technology
Computer Science Division
University of California, Berkeley
Department of Cognitive Science
University of California, San Diego

Image segmentation is a fundamental and widely studied problem in computer vision [1, 2, 4]. Continuous efforts have been made to improve the performance of segmentation systems to match human capability [1]; however, it is generally acknowledged that solving the segmentation problem with low-level cues alone might not be possible. There has long been a discussion on solving this seemingly low-level task with high-level knowledge [3], but a clear and concrete solution is not yet available.

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. 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 1, a mammal is classified as a bird because of their similarity in color and texture, but the derived segmentation is mostly correct.

The LM+SUN dataset [5] 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.

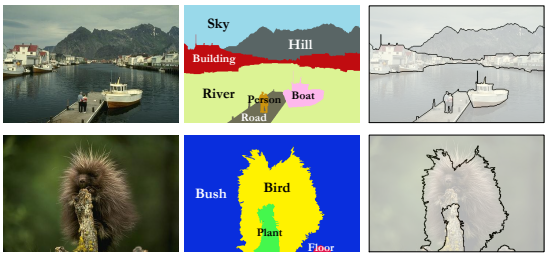
Specifically, we first assign each superpixel a semantic label. Following [5], a superpixel is associated with a semantic class if and only if at least half of the superpixel overlaps with a ground truth segment mask with that label. Then, according to the label frequencies on superpixels, 50 most frequent 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 is then trained on the data. These classifiers are generic and applicable to any images including those not in the dataset. For segmentation, each superpixel is tested by all learned classifiers to obtain a vector of confidence values.

We then 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 [5], 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 defined as $\Theta(c_i,c_j|s_i,s_j) = W(s_i,s_j)/(1+\|s_i - s_j\|)\cdot\delta[c_i \neq c_j]$, 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.

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. Given λ and μ, we adopt MCMC methods for inference. Because the CRF is built on superpixels, the inference is highly efficient, taking approximately 0.1 second per image on average.



(a) (b) (c)

Figure 1: 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.

	BSDS300					
	Covering ↑		PRI ↓		VoI ↑	
	ODS	OIS	ODS	OIS	ODS	OIS
Human	0.73	0.73	0.87	0.87	1.16	1.16
RIS+HL	**0.59**	**0.65**	**0.82**	**0.86**	1.71	1.53
RIS+H	0.55	0.60	0.80	0.84	1.82	1.63
RIS+L	0.57	0.63	0.79	0.82	1.80	1.60
RIS	0.52	—	0.77	—	1.99	—
SuperParsing	0.48	—	0.74	—	2.07	—
gPb-owt-ucm	**0.59**	**0.65**	0.81	0.85	**1.65**	1.47
fPb-owt-ucm	0.57	0.63	0.80	0.84	1.69	1.49
cPb-owt-ucm	**0.59**	**0.65**	0.81	0.85	1.66	**1.46**
MShift	0.54	0.58	0.78	0.80	1.83	1.63
FH	0.51	0.58	0.77	0.82	2.15	1.79
Canny	0.48	0.56	0.77	0.82	2.11	1.81
MNCuts	0.44	0.53	0.75	0.79	2.18	1.84
SWA	0.47	0.55	0.75	0.80	2.06	1.75
Quad-Tree	0.33	0.39	0.71	0.75	2.34	2.22

Table 1: Comparison on the test sets of BSDS300 and BSDS500 with both supervised and unsupervised methods. For each measure, the best algorithm is highlighted.

We finally discard the semantic labels produced by CRF to obtain segmentations. 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) [5].

As shown in Table 1, our solution yields highly competitive results on the famous Berkeley Segmentation Benchmark (BSDS300) [1]. When methods based purely on the ambiguous low-level features [1] tend to merge patches of similar appearances but different semantics, high-level semantic knowledge could help to figure out a correct segmentation. We also conduct experiments on multiple other datasets and obtain consistent results. Detailed illustrations and comparisons can be found in our paper and supplementary material.

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

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

[3] 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.

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

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