Semantic Image Segmentation Using Region Bank

Wenbin Zou, Kidiyo Kpalma, Joseph Ronsin

Semantic Image Segmentation Using Region Bank

Wenbin Zou, Kidiyo Kpalma and Joseph Ronsin
Université Européenne de Bretagne, INSA/IETR/UMR CNRS 6164, France

Abstract

Semantic image segmentation assigns a predefined class label to each pixel. This paper proposes a unified framework by using region bank to solve this task. Images are hierarchically segmented leading to region banks. Local features and high-level descriptors are extracted on each region of the banks. Discriminative classifiers are learned based the histograms of features descriptors computed from training region bank (TRB). Optimally merging predicted regions of query region bank (QRB) results in semantic labeling. This paper details each algorithmic module used in our system, however, any algorithm fits corresponding modules can be plugged into the proposed framework. Experiments on the challenging Microsoft Research Cambridge (MSRC 21) dataset show that the proposed approach achieves the state-of-the-art performance.

1. Introduction

In recent years, semantic image segmentation, which aims to precisely segmenting objects and assigning a semantic label to each pixel of the image, has attracted considerable attention. This has high practical value in many applications, such as image editing, object retrieval and intelligent image coding.

Several authors have proposed to combine low-level segmentation and high-level knowledge to achieve semantic segmentation. Csurka and Perronnin [6] applied Fisher model to describe over-segmented regions and employed the result of image classification to reduce the number of object classes in an image. Li et al. [7] made use of image tags and scene information to infer the existence of an object in the image. Lempitsky et al. [10] used bounding boxes acquired by object detection as a prior of the segmentation. Some authors also suggested incorporating different cues into a Random Field (RF) model. Verbeek and Triggs [9] combined advantages of probabilistic latent semantic analysis model and Markov Random Field (MRF) model to fuse region-level labels and image-level labels. Jiang and Tu [8] used auto-context model to integrate image appearances with context information learned by a set of classifier. All these methods advise that combining different cues might give a good result.

However, most of existing region-based approaches for semantic segmentation extract local features directly from objects delineated by ground-truth and or single-level regions generated by over-segmentation to train classification models; and at the testing step, the features are extracted on single-level regions. As known that low-level segmentation is unstable and cannot precisely separate objects, while local features are only extracted on the single-level regions for recognition, errors from the low-level segmentation might directly migrate to semantic inference. In this paper, we explore extracting the local features on multi-level regions for both training and testing steps. The region sets used for training and testing are respectively named as training region bank (TRB) and query region bank (QRB). Our motivation is that by fusing multi-level regions one might have more chance to capture objects or discriminative parts of objects; moreover, region hierarchy provides natural spatial constraint for high-level representation. To demonstrate the performance of this combination, we do not use any Random Field model to integrate multiple cues for inference. Experiments on the standard multi-object datasets show that this approach obtains comparable results with the state-of-the-art.

2. Proposed algorithm

Figure 1 shows the framework of our algorithm, which consists of following four algorithmic modules:

(i) Region bank generation: creates multi-level regions for an input image. (ii) Region description: extracts local invariant features on the corresponding region and transfers these features into high-level representation. (iii) Region classification: predicts semantic of region by using a set of discriminative classifiers. (iv) Image labeling: assigns each pixel with an object class label by fusing all regions of QRB.

Any algorithm that fits the above modules can be plugged into our system. The following subsections detail concrete algorithms used in our system.

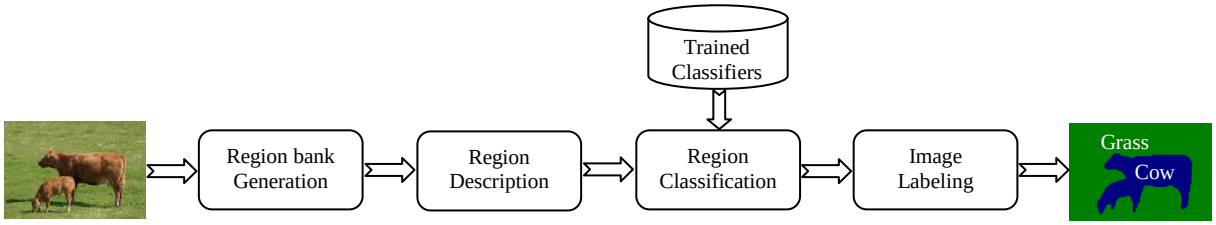

Figure 1. **Unified framework of semantic segmentation**

2.1. Region bank generation

Region bank is a set of multi-level regions. There are mainly two reasons to use region bank for semantic segmentation. On one hand, single-level segmentation or over-segmentation is unstable and far from precisely separating objects. In most cases, objects are segmented into many regions. On the other hand, hierarchical segmentation might capture objects at some levels, but the optimal segmentation level is unpredictable and may change according to components of images. As shown in figure 2, the best segmentation of image-1 is at level 8: cows grass and building are segmented with very few merging; while for image-2 the best is at level 4: face bodies, grass and building are separated. So we take multi-level regions into consideration.

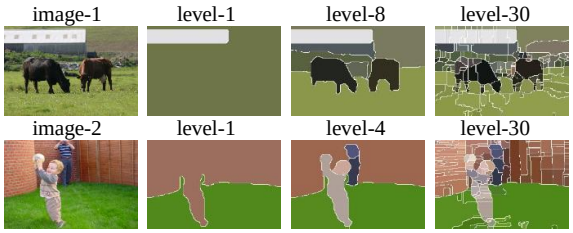

Figure 2. **Results of hierarchical segmentation**

To create region banks, we choose contour-based hierarchical segmentation proposed in [1]. Because it generally preserves object global contour while providing hierarchical regions. The result of this segmentation is a valued ultrametric contour map (UCM), where the contour values reflect contrast between neighboring regions. Hierarchical regions are created by thresholding the UCM with a set of thresholds. For semantic segmentation, too fine regions tend to produce noise labeling. So we design an image self-adapting method to compute thresholds: the minimum and maximum thresholds are computed by multiplying the maximum UCM value of input image by predefined parameters α and β; and UCM values in this range are taken as thresholds to create multi-level regions. In our experiments, α and β are set to be 0.25 and 0.8 respectively. Typically we obtain 5 to 20 thresholds per image. The region set created by hierarchical segmentation for a query image is query region bank (QRB); and that created by hierarchical segmentation and ground-truth segmentation for training images is training region bank (TRB).

2.2. Region description

For regions classification, it is necessary to extract robust feature descriptors for each region. Typically, the feature description consists of two steps: firstly, extracting local descriptors and then transforming these local descriptors into high-level representation.

Local feature descriptors: A good local feature descriptor should possess invariance property, i.e. if there is a transformation (e.g., rotation and scale change) between two instances of an object, the corresponding descriptor values must remain nearly the same. In our algorithm, we use two kinds of local feature descriptors.

The first kind of descriptor is RGB-SIFT. SIFT[2] have been shown to be well-adapted to matching and recognition tasks. The SIFT descriptor is formed by computing the histogram of gradient of 4×4 cells with 8 orientation bins in each cell. This results in a 128-dimensional vector for one SIFT descriptor. We extract SIFT descriptor on R, G and B channels respectively. So this leads to a 384-dimensional vector for one RGB-SIFT descriptor. Many authors extract local descriptors only on keypoints for high efficient image classification. However, it is not adapted to semantic segmentation, because keypoint detectors have difficulties to detect keypoints in uniform regions, such as sky and calm water, and result in unassignment on these areas. We thus prefer to perform on dense grid: SIFT descriptors are extracted respectively at four scales of grid size (8, 16, 24, 32 pixel diameters) with step-size of 6 pixels.

The second kind of descriptor is texton which is used to describe human textural perception. Although the word "texton" remains a vague concept, it has attracted much attention and a lot of methods have been proposed to represent texton for image analysis. We generalize texton descriptors by convolving images with a filter-bank of 17 filters [3] applying it to CIE L*a*b* color space. The L* chan-

nel owns 11 filters, 3 Gaussians ($\sigma = 1,2,4$), 4 Laplacian of Gaussians ($\sigma = 1,2,4,8$), and 4 derivatives of Gausians ($\sigma = 1,2$) along x and y directions. Each color channel a* or b* hold the same 3 Gaussians as L*'s respectively.

High-level representation: One of the popular High-level representations for image classification is bag-of-visual-words (BOV) [6]. We also apply it to semantic segmentation. BOV is visual dictionary-based representation: each local descriptor is quantized into the nearest element of the dictionary. As two kinds of local descriptors are used, we need to construct two visual dictionaries: RGB-SIFT dictionary (Ds), and texton dictionary (Dt).

We use the simplest square-error clustering method, k-means, to generate the Ds and Dt. Note that k-means cannot determine the number of clusters corresponding to the number of visual words in dictionary, therefore, we run k-means several times setting different number of clusters and choose appropriate number. In our experiment, the sizes of Ds and Dt are set to *2000* and *400* respectively. With the dictionary, each local feature descriptor can be represented by its nearest visual word. So each region can be described by the histogram of visual words.

2.3. Region classification

Once regions have been represented by histograms of visual words, the problem of region classification is transformed to that one of multi-class supervised classification. To predict the classes of unlabeled region, the classifier performs two steps: training and testing. Theoretically, any discriminative classifier may be performed within this task. We choose Support Vector Machine with Multiple Kernel Learning [4] (SVM-MKL) since it is convenient to integrate multiple features and it generally produces good results on high dimensional classification.

Suppose h_i^s and h_i^t are visual word histograms of SIFT and texton respectively of region *i;* and their combination is denoted as $h_i^c = \{h_i^s, h_i^t\}$. The classification function of a SVM in kernel formulation is expressed as:

$$SVM(h^c) = \sum_{i=1}^{I} y_i \, \alpha_i K(h^c, h_i^c) + b \qquad (1)$$

where h^c is feature histogram of a test region*; I* is the number of regions in TRB *;* $y_i \in \{+1, -1\}$ indicates their class label; and *K* is positive definite kernel, which is a linear combination of feature histogram kernels

$$K(h^c, h_i^c) = d^s K(h^s, h_i^s) + d^t K(h^t, h_i^t) \qquad (2)$$

where d^s and d^t denote nonnegative weights of kernels. Radial basis function (RBF) kernel is applied here to map the feature histograms into high dimension spaces. SVM-MKL learns parameters α_i, b, d^s, d^t for each classifier. Noted that most elements of α_i are zero, in other words, SVM-MKL only chooses the most import regions of TRB for classification. If a dataset contains *n* object classes, SVM-MKL trains *n* classifiers and a query region obtains *n* SVM scores.

2.3. Image labeling

Image labeling is to fuse all predicted regions to produce a semantic labeled image. In this module, SVM scores, regions sizes and common sense are taken into consideration. Specifically, the most likely object classes that have the maximum SVM scores are used to pre-label each region. Then, these regions are sorted by their increasing SVM scores and gradually merged to form a complete labeled image by observing their sizes, their SVM scores and considering common sense. For example, when two regions overlap, if a large region predicted as sky has a lightly larger SVM score than a small region predicted as bird; we preserve the small one and label it as bird.

3. Experiments

In this section, we report results on the MSRC 21 dataset [5]. This is one of the most challenging data set for semantic image segmentation which consists of 591 color images of 21 object classes. We use the same splitting protocol as [6] and [9]: 276 images for training and the remainder for testing.

Pixel-wise global accuracy, per-class accuracy and average accuracy are used to evaluate performance of the system. The global accuracy is computed as

$$\bar{g} = \frac{1}{\sum_i N_i} \sum_i \sum_{p \in T_i} 1(\psi(p) = s(p),\ s(p) > 0) \qquad (3)$$

where, T_i is the image lattice for test image *i*; N_i is the number of ground-truth labeled pixels of image *i;* For pixel p in image *i*, the output label of the system is $\psi(p)$ and the ground-truth label is $s(p)$*;* for unlabeled pixels, $s(p) = 0$. We also compute per-class accuracy as

$$\bar{c_l} = \frac{\sum_i \sum_{p \in T_i} 1\,(\psi(p) = s(p),\ s(p) = l)}{\sum_i \sum_{p \in T_i} 1\,(s(p) = l)},\ l = 1, \ldots, L. \qquad (4)$$

Then the average accuracy is the mean of all classes' accuracies.

Figure 3 shows the qualitative performance of our algorithm. The inputs images are displayed in figures 3(a)(d)(g), and their corresponding inferred labels and ground-truth labels are in figures 3(b)(e)(h) and 3(c)(f)(i) respectively. Each object class is labeled by

a unique color. Those black pixels of ground-truth image are unlabeled, but our algorithm does not produce unlabeled output. In most cases, our algorithm provides reasonable prediction in those unlabeled regions, such as in figure 3.2(h), the grass under chair is correctly inferred. Some good segmentation results are shown in figure 3.1 and 3.2. There are also typical failure examples shown in figure 3.3, where objects in the images are ambiguous and or occluded leading to failure labeling.

In table 1, we compare our results to the state-of-the-art. Our approach provides highest segmentation accuracy for 5 classes that are "tree", "water", "car" "book", and "dog". All approaches produce low accuracy for "boat", because it has very few examples and dramatic intra variance (rowboat, sailship, steamship, etc.). The average accuracy we obtained is 70% and ranks second. However, our method provides 80% global accuracy which is higher than others.

4. Conclusions

We have proposed a novel approach for semantic image segmentation by using region bank. Hierarchical regions are used for both training and testing. Experiments show that our approach obtains comparable results with the state-of-the-art on the standard dataset for semantic image segmentation. It is worthwhile to note that our approach has not employed any Random Field models which are used in most existing approaches to incorporate context information, and only used two types of local feature. Taking more features and considering the context information would increase the segmentation accuracy.

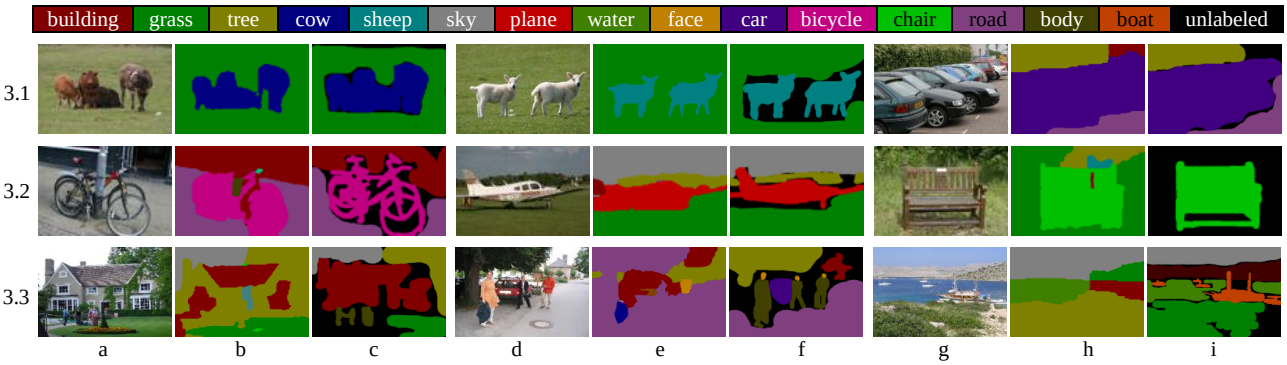

Figure 3. **Examples of semantic segmentation.** (a)(d)(g) original images; (b)(e)(h) segmented result of our system; (c)(f)(i) ground-truth segmentation.

Table 1. **Segmentation accuracies on the MSRC 21 dataset**

	Building	Grass	Tree	Cow	Sheep	Sky	Plane	Water	Face	Car	Bike	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat	**Average**	**Global**
[5]	49	88	79	**97**	**97**	78	82	54	**87**	74	72	74	36	24	93	51	78	75	35	**66**	18	67	72
[6]	**84**	95	81	67	78	89	72	77	**87**	71	86	66	59	28	85	19	68	59	47	35	9	65	77
[8]	53	**97**	83	70	71	**98**	75	64	74	64	88	67	46	32	92	61	**89**	59	66	64	13	68	78
[11]	60	78	77	91	68	88	**87**	76	73	77	**93**	**97**	**73**	**57**	95	**81**	76	**81**	46	56	**46**	**75**	77
Ours	66	92	**86**	67	85	91	72	**84**	79	**79**	83	87	39	38	**96**	42	74	60	**77**	50	19	70	**80**

References

[1] P. Arbelaez, M. Maire, C. Fowlkes and J. Malik. "Contour Detection and Hierarchical Image Segmentation," *PAMI*, 2011.

[2] D. G. Lowe, "Object recognition from local scale-invariant features," *ICCV*, 1999.

[3] J. Winn, A. Criminisi and T. Minka "Object Categorization by Learned Universal Visual Dictionary," *ICCV*, 2005.

[4] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," ICCV, 2007.

[5] J. Shotton, M. Johnson, R. Cipolla, "Semantic Texton Forests for Image Categorization and Segmentation," *CVPR*, 2008.

[6] G. Csurka, F. Perronnin, "An efficient approach to semantic segmentation," *IJCV*, 2010.

[7] L.-J. Li, R. Socher and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," *CVPR*, 2009

[8] J. Jiang, Z. Tu, "Efficient scale space auto-context for image segmentation and labeling," *CVPR*, 2009

[9] J. Verbeek and B. Triggs, "Region classification with Markov field aspect models," *CVPR*, 2007.

[10] V. Lempitsky, P. Kohli, C. Rother, T. Sharp, "Image segmentation with a bounding box prior," *ICCV*, 2009.

[11] J.M. Gonfaus, X. Boix, J. van de Weijer, A.D. Bagdanov, J. Serrat, and J. Gonzàlez, "Harmony potentials for joint classification and segmentation," *CVPR*, 2010.