Object Detection by Labeling Superpixels

Junjie Yan[1,2], Yinan Yu[3], Xiangyu Zhu[1], Zhen Lei[1], Stan Z. Li[1]
[1]National Laboratory of Pattern Recognition, Chinese Academy of Sciences. [2]Institute of Data Science and Technology, Alibaba Group. [3]Institute of Deep Learning, Baidu Research.

Object detection is a computer vision task to automatically localize objects in categories of interest from images, such as the the twenty categories in Pascal VOC and two hundred categories in ImageNet. While numerous works have been proposed for object detection, most of them actually transform the object detection to image classification. They first generate object proposals and then classify each proposal independently by the image classification techniques. The traditional paradigm to get proposal is to use the sliding window to extensively sample about 100, 000 bounding boxes in various scales and locations. The recently popular paradigm is to generate about 2, 000 proposals by clustering or segmentation according to low-level image cues. After that, image classification techniques are used to classify each proposal. The classification has achieved great advances recently, due to the robust low level features, sophisticated models and convolutional neural networks.

Through the transformation, the detection performance can always benefit from the advances in image classification. However, it also results in two problems. The first is that if an object is missed in object proposal step, such as an object with partially occlusion or unusual aspect ratio, the detection system would definitely miss the object. The second is that the independent classification of proposals cannot incorporate the global image context, which is very important to detect overlapped objects and distinguish object part and object itself.

To alleviate the two problems, in this paper, we move the focus in detection from proposals to superpixels. The pixels in one superpixel can be safely assumed to belong to the same object and superpixels can be grouped together flexibly to form objects. The interaction between objects, which is hard to model in object level, also becomes easier in superpixel level. If we know the label of each superpixel, then the object detection problem becomes trivial. To this end, we conduct object detection by labeling superpixels.

For each superpixel generation setting, we can get a superpixel partition of an image and denote it as $\mathcal{P} = \{p_1, p_2, \cdots, p_N\}$, where p_i is the i-th superpixel and N is the superpixel number. Based on the partition, we also have a neighborhood system $\mathcal{N}$, where $(p_i, p_j) \in \mathcal{N}$ if p_i and p_j are spatially connected. The detection is conducted by finding a label configuration for each superpixel $\mathcal{L} = \{l_1, l_2, \cdots, l_N\}$, where the label $l_i \in \{0, 1, 2, \cdots, \infty\}$. Here $l_i = 0$ means p_i belongs to the background, $l_i = j$ means p_i belongs to the j-th object and the object number can be any non-negative integer. For the simplicity, we handle each category independently at the labeling step.

For each labeling configuration, we define an energy function $E(\mathcal{L})$ to measure its cost and can find the best label configuration $\mathcal{L}^*$ with the smallest cost by minimizing $E(\mathcal{L})$. Now let us think what an appropriate label configuration should be. When considering each superpixel independently, its label should be based on the fitness between its appearance and the appearance model learned from the training data of this category. Considering the smoothness nature of objects in image, the labels of neighborhood superpixels should be correlated and punished for varying labels. If two neighborhood superpixels have the same label and thus be taken as the same object, their appearance should also be correlated. Finally, the label configuration should favor fewer labels for compact detection. To this end, we use the following energy function,

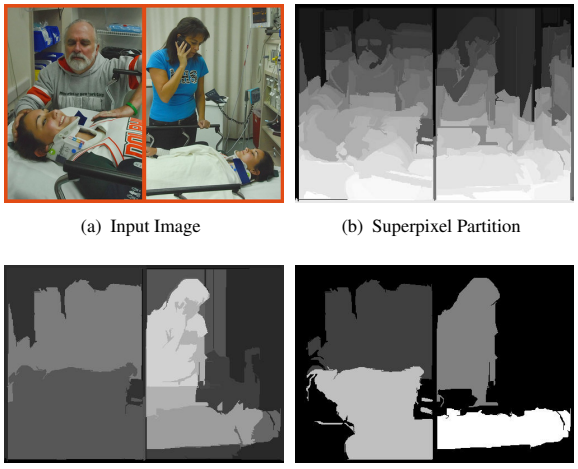
$$E(\mathcal{L}) = \sum_{p_i \in \mathcal{P}} D(l_i, p_i) + \sum_{(p_i, p_j) \in \mathcal{N}} V(l_i, l_j, p_i, p_j) + C(\mathcal{L}), \qquad (1)$$

where we always ignore the image notation I to simplify the notation. $D(l_i, p_i)$ is the data cost to capture the appearance of p_i and assign a cost based on the conflict between the appearance model and the label l_i. $V(l_i, l_j, p_i, p_j)$ is the pairwise smooth cost defined on the neighborhood system $\mathcal{N}$. $C(\mathcal{L})$ is the label cost term, which is defined on the label configurations $\mathcal{L}$ and is image invariant. It is motivated by the MDL prior and plays an important role to get objects in detection instead of object parts.

In this way, the detection becomes a multi-label labeling problem with label cost, and α-expansion based method can be used for approximate inference. To learn the parameters in the energy function, such as the weight of different terms, a structural SVM is conducted to maximize the detection performance. We use the region CNN (RCNN) [1] as a special case of our method by taking it as the data cost term.

Method	single model	# CNNs	Combined
Deep Insight	40.2	3	40.5
CUHK-DeepID-Net	37.7	10	40.7
GoogLeNet	38.0	7	43.9
Superpixel Labeling	**42.5**	4	**45.0**

Table 1: Results on the testing set of ILSVRC2014 detection task.



(a) Input Image (b) Superpixel Partition

(c) Data Cost (d) Superpixel Labeling

Figure 1: Example of the proposed superpixel labeling approach.

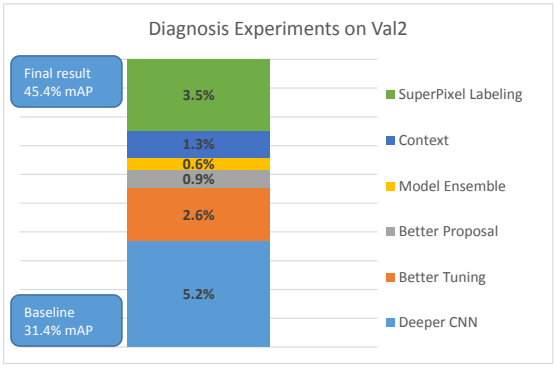


Figure 2: Diagnosis Experiments on val2 of ImageNet Detection.

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.