Multiscale Combinatorial Grouping

Zhang, Liqiang

June 18, 2018

Abstract

This paper propose a unified approach for bottom-up hierarchical image segmentation and object candidate generation for recognition, called Multiscale Combinatorial Grouping (MCG). For this purpose, the author first develop a fast normalized cuts algorithm. They then propose a high-performance hierarchical segmenter that makes effective use of multiscale information. Finally, the author propose a grouping strategy that combines their multiscale regions into highly-accurate object candidates by exploring efficiently their combinatorial space.

1. Introduction

Two paradigms have shaped the field of object recognition in the last decade. The first one, popularized by the Viola-Jones face detection algorithm, formulates object localization as window classification. The basic scanning-window architecture, relying on histograms of gradients and linear support vector machines, was introduced by Dalal and Triggs in the context of pedestrian detection and is still at the core of leading object detectors on the PASCAL challenge such as Deformable Part Models [1].

The second paradigm relies on perceptual grouping to provide a limited number of high-quality and category-independent object candidates, which can then be described with richer representations and used as input to more sophisticated learning methods.

In this paper, the author propose a unified approach to multiscale hierarchical segmentation and object candidate generation called Multiscale Combinatorial Grouping (MCG). Fig. 1 shows an example of the results. Their main contributions are:

- An efficient normalized cuts algorithm, which in practice provides a 20× speed-up to the eigenvector computation required for contour globalization.
- A state-of-the-art hierarchical segmenter that leverages multiscale information.

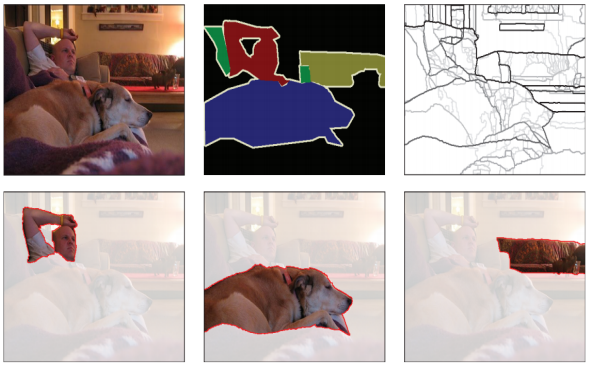

Figure 1. **Top:** original image, instance-level groundtruth from PASCAL and our multiscale hierarchical segmentation. **Bottom:** our best object candidates among 400.

- A grouping algorithm that produces accurate object candidates by efficiently exploring the combinatorial space of their multiscale regions.

2. The Segmentation Algorithm

Spatially transforming an UCM is nontrivial because its boundaries are one-dimensional entities whose topology and strength determine the underlying hierarchy, and an error at a single pixel can have drastic effects. The author therefore opt for sampling uniformly K levels in the hierarchy, transforming them sequentially, and reconstructing from their boundaries a transformed UCM.

They consider two different segmentations $\mathcal{R} = \{R_i\}_i$ and $\mathcal{S} = \{S_j\}_j$. They define the projection of the segmentation $\mathcal{R}$ onto a region $S_j \in \mathcal{S}$ as the majority label

$$\pi(\mathcal{R}, S_j) = \text{argmax} \frac{|S_j \cap R_i|}{|S_j|} \tag{1}$$

And the projection of $\mathcal{R}$ onto $\mathcal{S}$ as

$$\pi(\mathcal{R}, S) = \{\pi(\mathcal{R}, S_j)\}_j. \tag{2}$$

In order to project an UCM onto a target segmentation $\mathcal{S}$, which the author denote $\pi(\text{UCM}, \mathcal{S})$, they project each of the levels in the hierarchy in turn.

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN pool_5	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2
R-CNN fc_6	43.6	50.4	32.2	26.0	9.8	58.5	50.4	30.9	7.9	36.1	18.2	31.7	41.4	52.6	8.8	14.0	26.3	34.0	53.1	58.0	46.2
R-CNN fc_7	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7
R-CNN FT pool_5	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	42.0	47.3	56.6	15.3	12.8	14.6	40.6	53.1	56.4	47.3
R-CNN FT fc_6	53.1	57.1	32.4	12.3	15.8	58.2	56.7	39.6	0.9	44.8	39.9	31.0	54.0	62.4	4.5	20.6	34.2	50.0	57.7	63.0	53.1
R-CNN FT fc_7	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-CNN FT fc_7 BB	**68.1**	**72.8**	**56.8**	**43.0**	**36.8**	**66.3**	**74.2**	**67.6**	**34.4**	**63.5**	**54.5**	**61.2**	**69.1**	**68.6**	**58.7**	**33.4**	**62.9**	**51.1**	**62.5**	**64.8**	**58.5**

Table 1. Detection average precision (%) on VOC 2007 test. Rows 1-3 show R-CNN performance without fine-tuning. Rows 4-6 show results for the CNN pre-trained on ILSVRC 2012 and then fine-tuned (FT) on VOC 2007 trainval. Row 7 includes a simple bounding box regression (BB) stage that reduces localization errors. The first uses only HOG, while the next two use different feature learning approaches to augment or replace HOG.

In the next section, we will iterate this procedure, and project an UCM recursively to a set of target segmentations $\{\mathcal{S}^{1*}, \ldots, \mathcal{S}^{N*}\}$. However, note that the composition of two such projections can be written as :

$$\pi(\pi(\text{UCM}, \mathcal{S}^1), \mathcal{S}^2) = \pi(\text{UCM}, \mathcal{S}^1) \circ \pi(\mathcal{S}^1, \mathcal{S}^2). \quad (3)$$

In practice, this property means that successive projections of the target segmentations can be pre-computed, the UCM has to be projected only to the first target segmentation, and its final labels are obtained by $N - 1$ look-ups.

Table 1 rows 1-3 reveals that features from fc_7 generalize worse than features from fc_7. This means that 29%, or about 16.8 million, of the CNNs parameters can be removed without degrading mAP. More surprising is that removing both fc_7 and fc_6 produces quite good results even though pool_5 features are computed using only 6% of the CNNs parameters.

References

[1] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, pages 3378–3385, 2012. 1