# Second-Order Constrained Parametric Proposals and Sequential Search-Based Structured Prediction for Semantic Segmentation in RGB-D Images

Dan Banica[1], Cristian Sminchisescu[2,1]
[1]Institute of Mathematics of the Romanian Academy. [2]Department of Mathematics, Faculty of Engineering, Lund University.

We focus on the problem of semantic segmentation based on RGB-D data, with emphasis on analyzing cluttered indoor scenes containing many visual categories and instances. Our approach is based on a parametric figure-ground intensity and depth-constrained proposal process that generates spatial layout hypotheses at multiple locations and scales in the image followed by a sequential inference algorithm that produces a complete scene estimate.

Our contributions can be summarized as follows: (1) a generalization of parametric max flow figure-ground proposal methodology to take advantage of intensity and depth information, in order to systematically and efficiently generate the breakpoints of an underlying spatial model in polynomial time, (2) new region description methods based on second-order pooling over multiple features constructed using both intensity and depth channels, (3) a principled search-based structured prediction inference and learning process that resolves conflicts in overlapping spatial partitions and selects regions sequentially towards complete scene estimates, and (4) extensive evaluation of the impact of depth, as well as the effectiveness of a large number of descriptors, both pre-designed and automatically obtained using deep learning, in a difficult RGB-D semantic segmentation problem.

We achieve state of the art results in the challenging NYU Depth v2 dataset [4], extended for the RMRC 2013 and RMRC 2014 Indoor Segmentation Challenges, where currently the proposed model ranks first. Moreover, we show that by combining second-order and deep learning features, over 15% relative accuracy improvements can be additionally achieved. In a scene classification benchmark, our methodology further improves the state of the art by 24%.

In contrast to methodologies that compute hierarchical, non-overlapping partitions of the image into multiple regions, our approach relies on generating multiple overlapping figure-ground segmentations, systematically, based on parametric max-flow solvers. We focus on constrained parametric min cuts models CPMC[1] generalized to take advantage of intensity and depth information (CPMC-RGBD). The figure-ground segmentation proposals are generated by solving a family of optimization problems for spatial energies of the form:

$$E^{\lambda}(L) = \sum_{x} D_{\lambda}(l_x) + \sum_{x,y \in \mathcal{N}(x)} V_{xy}(l_x, l_y) \qquad (1)$$

In order to incorporate depth data, our CPMC-RGBD model relies on two boundary probability maps, one based on the RGB information in the image and the other using the depth image. We then modulate the pairwise term $V_{xy}$ of the spatial model (eq. 1) to account for both intensity and depth discontinuities, which has the following form when two neighboring pixels $x, y$ are assigned different labels (and zero otherwise):

$$V_{xy}(l_x, l_y) = \exp\left[-\frac{\max(Gb_{\mathcal{I}}(x), Gb_{\mathcal{I}}(y), Gb_{\mathcal{D}}(x), Gb_{\mathcal{D}}(y))}{\sigma^2}\right], \qquad (2)$$

where $Gb_{\mathcal{I}}$ is the output of a generalized, trained contour detector computed for the image $\mathcal{I}$ at a given pixel and $Gb_{\mathcal{D}}$ is the output of the global contour detector applied on the depth image.

The region proposals generated by CPMC-RGBD are characterized using local descriptors that capture both the appearance and the depth information available in the RGB-D images. Local descriptors extracted inside the region are aggregated using Second Order Pooling (O2P) [2]. O2P introduces multiplicative second-order analogues of average pooling that together with additional non-linearities (matrix logarithm, power normalization) produce good predictors without the need of going through a feature coding step. We pool local features characterized by say, $M$ descriptors, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_M)$, $\mathbf{x} \in \mathbb{R}^n$, extracted over patches centered at image locations sampled inside a region proposal $R$, to form global descriptors based on second-order statistics.

$$\mathbf{G}^{log}_{avg}(R) = \log\left(\frac{1}{M}\sum_i \mathbf{x}_i \cdot \mathbf{x}_i^{\top}\right) \qquad (3)$$

The pooling process operates over descriptors that capture both appearance (*e.g.* SIFT) and geometry (*e.g.* spin images).

Besides the pooled local descriptors we also extracted features from a large convolutional neural network trained for image classification on ImageNet. In the experiments we present the performance of the above features individually and also show the benefits of using them together to jointly describe each region.

A class label is assigned to each segment by learning linear category models, one per class, trained to predict the overlap (IoU) between the segment and the best matching object of that class.

An inference procedure is defined in order to resolve conflicts between overlapping segments and to generate a final per-pixel segmentation. We formulate the solution in terms of a principled sequential framework which involves learning a policy to optimally select among a set of actions until a final state is reached. This procedure is derived from Search-based Structured Prediction (SEARN) [3] and shares connections with reinforcement learning. We are not aware of an application of such principles for the task of semantic segmentation.

Formally, we learn a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$, where $\mathcal{S}$ represents the set of states (partial semantic segmentations of an image in our case) and $\mathcal{A}$ represents the set of actions (labeled segments to be added next, along with a special 'stop action'). In order to train a policy to optimally select an action available in a given state, we gather multiple cost-sensitive training examples. Ideally, these examples should be generated from intermediary partial segmentations which are similar in nature to those which will be encountered at test time – to this end, training proceeds in an iterative fashion, starting with an initial policy. This framework allows to consider the long-term effect of actions and to optimize decisions under the true cost we are interested in. The behavior at test-time is illustrated in fig. 1. Our method selects one labeled region at a time, using the learned policy to take into account interactions between candidate regions and regions already included.


Figure 1: Sequential search-based structured prediction procedure at test-time. We sequentially apply the learned policy and select the action predicted as most suitable given the current state.

[1] João Carreira and Cristian Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *PAMI*, 34(7), 2012.

[2] João Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*. Springer, 2012.

[3] Hal Daumé III, John Langford, and Daniel Marcu. Search-based structured prediction. *Machine learning*, 75(3):297–325, 2009.

[4] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

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