Multi-Class Image Labeling with Top-Down Segmentation and Generalized Robust P^N Potentials

Georgios Floros[1]
floros@umic.rwth-aachen.de
Konstantinos Rematas[2]
konstantinos.rematas@esat.kuleuven.be
Bastian Leibe[1]
leibe@umic.rwth-aachen.de

[1] UMIC Research Centre
RWTH-Aachen University
Aachen, Germany

[2] ESAT-PSI
KU Leuven
Leuven, Belgium

Recently, there has been increased interest in combining object class detection ("things") and texture segmentation ("stuff") for scene understanding. There is mutual benefit from such a combination. Object detectors can be improved by context ("from stuff"). In return, segmentation can be improved by semantic information provided by the object detector. Ladicky *et al*. [5] propose to obtain the support region for a detected object by applying GrabCut [7] on the detector bounding box. This GrabCut segmentation introduces an additional, separate CRF segmentation step prior to the final image-level CRF segmentation, even though both decisions are based on the same color potentials. We argue that there should be only one segmentation decision made as a result of the joint inference. Furthermore, the GrabCut segmentation step ignores any specific information about the detected object class. In particular, it does not take into account how important a certain pixel was for the initial detection decision. We propose to bring in this information by feeding back soft, class-specific top-down segmentation information from the object detector for optimization in a single CRF. In this paper, this is done in the form of integrating class-specific information in the form of *generalized robust higher order potentials* [4]. These potentials make it possible to specify a per-pixel weight which expresses how important a pixel is for preserving object consistency.

The formulation of the energy function $E(\mathbf{y})$ in a higher order CRF, consisting of unary (ψ_i), pairwise (ψ_{ij}), and robust P^N (ψ_c) potentials takes the following form

$$E(\mathbf{y}) = \sum_{i \in V} \psi_i(y_i) + \sum_{(i,j) \in E} \psi_{ij}(y_i, y_j) + \sum_{c \in \mathscr{S}} \psi_c(\mathbf{y}_c),$$

which has shown state-of-the art performance in the multi-class image labeling problem [4]. While the unary and pairwise potentials are defined on the pixel level, the robust P^N potentials are defined over a set of segments $\mathscr{S}$. In [4], those segments are created by an unsupervised multi-level mean-shift segmentation [2]. The P^N potentials introduce a cost for assigning different label classes to pixels that are part of the same segment, while taking into account the quality of the entire segment. *Generalized* robust P^N potentials provide a structured framework for incorporating the class-specific information provided by an object detector. As introduced in [4], the per-pixel weights provide a nice interface to naturally introduce a per-pixel factor which expresses the importance of each object pixel in the preservation of object consistency.

Top-down segmentations provide output from an object detector in the form of soft decisions on whether an image pixel belongs to a specific object or not. They are obtained from an extended version of the Hough Forest detector [3]. The idea behind Hough Forests is to store for each leaf node the spatial occurrence distribution (relative to the object center) of all patches that were assigned to this node. During testing, those stored locations are then used to cast probabilistic votes for the object center in a Generalized Hough Transform. As shown in [6], the votes corresponding to a local maximum in the Hough space can then be backprojected to the image in order to propagate top-down information to the patches they were originating from. We extend the Hough Forest classifier with this top-down segmentation formalism, using figure-ground labels learned from annotated training examples. Each vote v_j contributing to a Hough space maximum h is backprojected to its originating patch $\mathbf{P}$, augmented with a local figure-ground label $Seg(v_j)$. We can then obtain the *figure* and *ground* probabilities for each pixel $\mathbf{p}$ by averaging over all patches $\mathbf{P}_i$ containing this pixel and summing the backprojected figure-ground labels, weighted by the weight of the corresponding vote w_{v_j}.

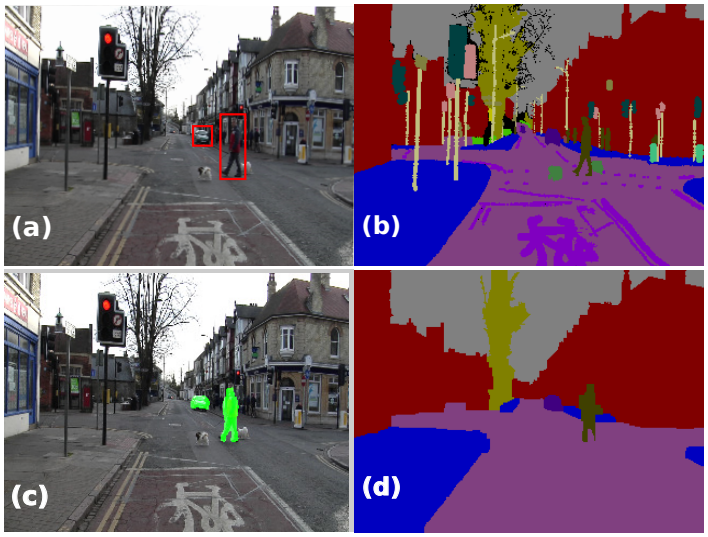

Figure 1: **Top-down segmentations improve multi-class image labeling. (a)** Test image with object detections. **(b)** Ground truth labeled image. Our algorithm uses top-down segmentations **(c)** to produce segmentation results **(d)**. (**Best viewed in color**)

$$p(\mathbf{p} = fig|h) = \frac{1}{\sum_{v_j \in h} w_{v_j}} \sum_{\mathbf{P}_i(\mathbf{p})} \frac{1}{|\mathbf{P}_i|} \sum_{v_j \in votes(\mathbf{P}_i)} w_{v_j} Seg(v_j) \quad (1)$$

$$p(\mathbf{p} = gnd|h) = \frac{1}{\sum_{v_j \in h} w_{v_j}} \sum_{\mathbf{P}_i(\mathbf{p})} \frac{1}{|\mathbf{P}_i|} \sum_{v_j \in votes(\mathbf{P}_i)} w_{v_j} (1 - Seg(v_j))$$

These soft decisions can also be interpreted as weights indicating the importance of each pixel in the preservation of the object's label consistency. It is, therefore, intuitive to propose the use of the foreground probability p_{fig} of each pixel as a weight w_i^k in the generalized robust P^N potentials.

We experimentally evaluate our approach on the CamVid dataset [1]. As our results indicate, we outperform the state-of-the-art systems for the classes that object detections are available and provide similar performance for the rest of the classes using a simpler CRF structure.

[1] G.J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

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

[3] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *CVPR*, 2009.

[4] P. Kohli, L. Ladickỳ, and P.H.S. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3):302–324, 2009.

[5] L'. Ladickỳ, P. Sturgess, K. Alahari, C. Russell, and P. Torr. What, Where and How Many? Combining Object Detectors and CRFs. In *ECCV*, 2010.

[6] B. Leibe, A. Leonardis, and B. Schiele. Robust Object Detection with Interleaved Categorization and Segmentation. *IJCV*, 77(1-3): 259–289, 2008.

[7] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graphics*, 23(3):309–314, 2004.