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

Georgios Floros, Konstantinos Rematas, Bastian Leibe

UMIC Research Centre, RWTH-Aachen University, Germany

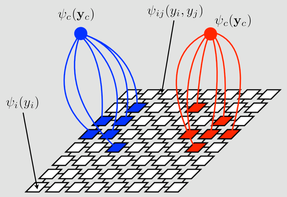
Abstract

We propose a novel formulation for the scene labeling problem which is able to combine object detections with pixel-level information in a Conditional Random Field (CRF) framework. Since object detection and multi-class image labeling are mutually informative problems, pixel-wise segmentation can benefit from powerful object detectors and vice versa. The main contribution of the current work lies in the incorporation of top-down object segmentations as generalized robust $\boldsymbol{P^N}$ potentials into the CRF formulation. These potentials present a principled manner to convey soft object segmentations into a unified energy minimization framework, enabling joint optimization and thus mutual benefit for both problems. As our results show, the proposed approach outperforms the state-of-the-art methods on the categories for which object detections are available. Quantitative and qualitative experiments show the effectiveness of the proposed method.

CRF Formulation for Multi-Class Image Labeling

Energy of the CRF

$$\boldsymbol{E}(\mathbf{y}) = \sum_{i \in \boldsymbol{V}} \boldsymbol{\psi}_i(\boldsymbol{y}_i) + \sum_{(i,j) \in \boldsymbol{E}} \boldsymbol{\psi}_{ij}(\boldsymbol{y}_i, \boldsymbol{y}_j) + \sum_{c \in \boldsymbol{\mathcal{S}}} \boldsymbol{\psi}_c(\mathbf{y}_c)$$



Potentials:

- Unary potentials: $\boldsymbol{\psi}_i(\boldsymbol{y}_i)$.
- Pairwise potentials: $\boldsymbol{\psi}_{ij}(\boldsymbol{y}_i, \boldsymbol{y}_j)$.
- Higher order potentials: $\boldsymbol{\psi}_c(\boldsymbol{y}_c)$.

Higher order potentials come in the form of robust $\boldsymbol{P^N}$ potentials

$$\boldsymbol{\psi}_c(\mathbf{y}_c) = \begin{cases} \boldsymbol{N}_i(\mathbf{y}_c)\frac{1}{\boldsymbol{Q}}\boldsymbol{\gamma}_{max} & \text{if } \boldsymbol{N}_i(\mathbf{y}_c) \leq \boldsymbol{Q}, \\ \boldsymbol{\gamma}_{max} & \text{otherwise .} \end{cases}$$

Segment proposals are created using unsupervised mean-shift segmentations.

Higher order potentials capture the fine contour details better than pairwise potentials yielding impressive results for *stuff* classes.

Recently, [3] proposed to add object detections as a higher order potential

$$\boldsymbol{\psi}_d(\mathbf{y}_d, \boldsymbol{H}_d, \boldsymbol{l}_d) = \min_{\boldsymbol{x}_d \in \{0,1\}} (-\boldsymbol{f}(\mathbf{y}_d, \boldsymbol{H}_d)\boldsymbol{x}_d + \boldsymbol{g}(\boldsymbol{N}_d, \boldsymbol{H}_d)\boldsymbol{x}_d).$$

where

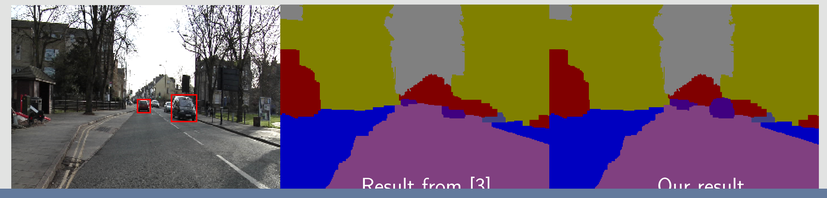
$\boldsymbol{f}(\mathbf{y}_d, \boldsymbol{H}_d) = \boldsymbol{w}_d|\mathbf{y}_d| \max(0, \boldsymbol{H}_d - \boldsymbol{H}_t)$
$\boldsymbol{g}(\boldsymbol{N}_d, \boldsymbol{H}_d) = \frac{\boldsymbol{w}_d}{\boldsymbol{p}_d} \max(0, \boldsymbol{H}_d - \boldsymbol{H}_t)\boldsymbol{N}_d$,
$\boldsymbol{H}_d$ is the detection score and $\boldsymbol{H}_t$ is the detection threshold.

Suboptimal Object Detection Integration

The integration of the object detections in [3] is not optimal:

- GrabCut provides hard segmentations without considering **object-specific** nor **context** information.
- CRF optimization cannot compensate for errors made during GrabCut segmentation, since only the number of consistent pixel matters and not their **importance**.

GrabCut failure case



Main Ideas

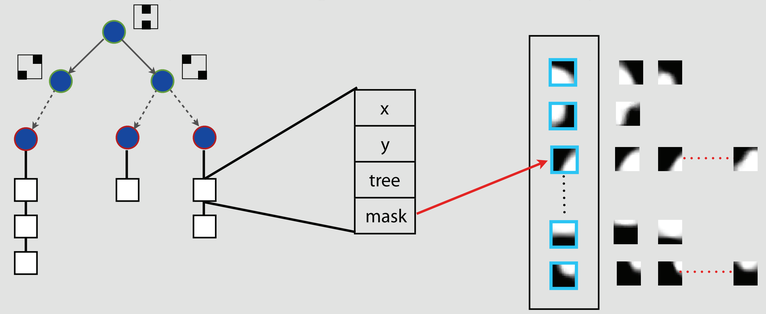
Integrate object detections as soft class-specific top-down segmentations.

Joint optimization of the multi-class image labeling problem and the object existence in a single CRF framework.

The importance of each pixel inside the object segmentation should be considered.

Class-specific Top-down Segmentation

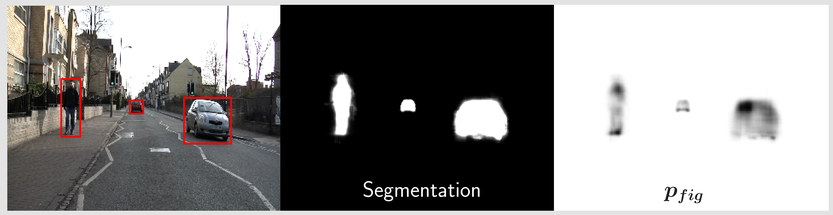
ISM formalism on top of a dense Hough Forest classifier to obtain top-down information.



Each vote $\boldsymbol{v}_j$ contributing to a Hough space maximum $\boldsymbol{h}$ is backprojected to its originating patch $\mathbf{P}$ augmented with a local figure-ground label patch $\boldsymbol{Seg}(\boldsymbol{v}_j)$.

$$p(\mathrm{p} = \boldsymbol{fig}|\boldsymbol{h}) = \frac{1}{\sum_{v_j \in h} \boldsymbol{w}_{v_j}} \sum_{\mathrm{P}_i(\mathrm{p})} \frac{1}{|\mathrm{P}_i|} \sum_{v_j \in votes(\mathrm{P}_i)} \boldsymbol{w}_{v_j} \boldsymbol{Seg}(\boldsymbol{v}_j)$$

$$p(\mathrm{p} = \boldsymbol{gnd}|\boldsymbol{h}) = \frac{1}{\sum_{v_j \in h} \boldsymbol{w}_{v_j}} \sum_{\mathrm{P}_i(\mathrm{p})} \frac{1}{|\mathrm{P}_i|} \sum_{v_j \in votes(\mathrm{P}_i)} \boldsymbol{w}_{v_j}(1 - \boldsymbol{Seg}(\boldsymbol{v}_j))$$



Generalized Robust P^N Potentials

Generalized robust $\boldsymbol{P^N}$ potentials take the form:

$$\boldsymbol{\psi}_c(\mathbf{y}_c) = \min\left\{ \min_{k \in \boldsymbol{\mathcal{L}}} ((\boldsymbol{P} - \boldsymbol{f}_k(\mathbf{y}_c))\boldsymbol{\theta}_k + \boldsymbol{\gamma}_k), \boldsymbol{\gamma}_{max} \right\}.$$

The parameters $\boldsymbol{P}$ and functions $\boldsymbol{f}_k(\mathbf{y}_c)$ are defined as:

$$\boldsymbol{P} = \sum_{i \in c} \boldsymbol{w}_i^k, \quad \forall \boldsymbol{k} \in \boldsymbol{\mathcal{L}}$$

$$\boldsymbol{f}_k(\mathbf{y}_c) = \sum_{i \in c} \boldsymbol{w}_i^k \boldsymbol{\delta}_k(\boldsymbol{y}_i)$$

$$\boldsymbol{\delta}_k(\boldsymbol{y}_i) = \begin{cases} 1 & \text{if } \boldsymbol{y}_i = \boldsymbol{k}, \\ 0 & \text{otherwise,} \end{cases}$$

We propose the use of $\boldsymbol{p}_{fig}$ as a weight $\boldsymbol{w}_i^k$ in the above formulation.

A high figure probability for a pixel will yield a higher cost in the case that it takes a label different from the object's.

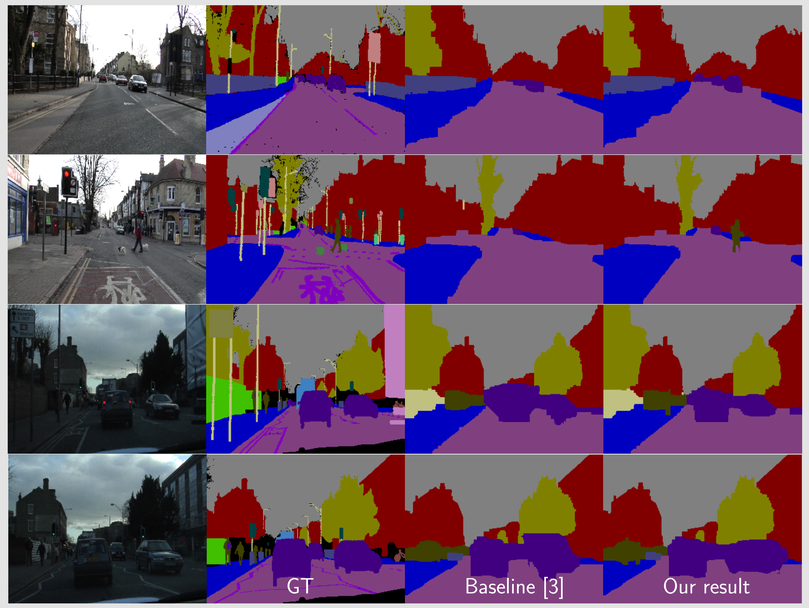
Results

Camvid dataset

- 10 min of high quality 30 Hz footage at 960x720 pixel resolution.
- 367 training and 233 test images with 11 semantic classes.

Quantitative results

	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Sidewalk	Bicyclist	**Global**	**Average**
Recall													
[1] (SfM)	*46.2*	*61.9*	*89.7*	*68.6*	*42.9*	*89.5*	**53.6**	*46.6*	*0.7*	*60.5*	*22.5*	*69.1*	*53.0*
[2] (SfM)	**84.5**	*72.6*	**97.5**	*72.7*	*34.1*	**95.3**	*34.2*	*45.7*	*8.1*	*77.6*	*28.5*	**83.8**	*59.2*
[3] (no det.)	79.3	76.0	96.2	74.6	**43.2**	94.0	40.4	47.0	**14.6**	81.2	31.1	83.1	61.6
[3] (det.)	**81.5**	**76.6**	**96.2**	78.7	40.2	93.9	43.0	**47.6**	14.3	**81.5**	**33.9**	**83.8**	**62.5**
Baseline [3]	79.0	75.2	95.7	74.1	20.2	93.7	39.7	46.7	8.3	78.1	18.3	82.3	57.2
Ours	80.4	76.1	96.1	**86.7**	20.4	**95.1**	**47.1**	47.3	8.3	79.1	19.5	83.2	59.6



Conclusion

Novel framework for combining object detections with a CRF framework.

Top-down soft object segmentations on the pixel-level enable a single energy optimization.

Improved performance on object classes with available detections.

References

[1] G.J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In ECCV, 2008.

[2] P. Sturgess, K. Alahari, L. Ladicky, and P. Torr. Combining appearance and structure from motion features for road scene understanding. In BMVC, 2009.

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

http://www.mmp.rwth-aachen.de
floros@umic.rwth-aachen.de