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

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 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.

1 Introduction

Recently, there has been increased interest in combining object class detection ("things") and texture segmentation ("stuff") for scene understanding (see Fig. 1). There is mutual benefit from such a combination. Object detectors can be improved by context ("from stuff") [12]. In return, segmentation can be improved by semantic information provided by the object detector. Moreover, the appearance of many object classes is so complex that segmentation decisions are only well-defined on an object level and not on a pixel or patch level.

In particular, the CRF based approach by Ladicky *et al*. [17] has shown very good performance in practice. The particular appeal of this approach lies in its property of seamlessly integrating into a CRF framework that can be optimized using the popular Graph-Cuts method [3] (in contrast to previous attempts by [7, 28]). This integration was made possible by the introduction of robust P^N potentials [13]. Those potentials enable the formulation of grouping constraints on a set of pixels (*e.g*. defined by a superpixel or a detector response) by considering the penalty of mislabeling as a linear function of the number of pixels that disagree with the majority vote.

BMVC 2011 http://dx.doi.org/10.5244/C.25.79

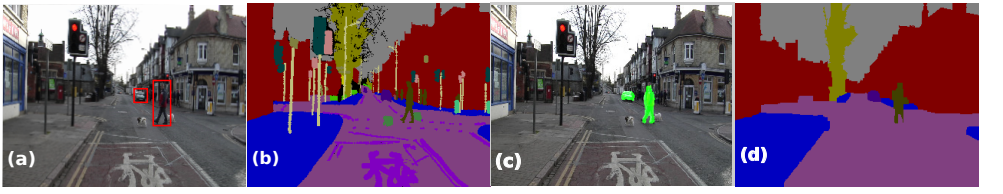


Figure 1: **Top-down segmentations improve multi-class image labeling. (a)** Test image with object detections. **(b)** Ground truth labeled image. Our algorithm uses top-down segmentations **(c)** to produce segmentation results **(d)**. (**Best viewed in color**)

Ladicky *et al*. [17] propose to obtain the support region for a detected object by applying GrabCut [23] 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 a single-stage optimization in a common CRF.

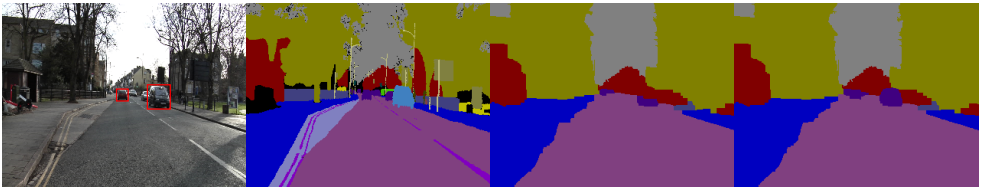
In particular, we propose to integrate class-specific information in the form of *generalized robust higher order potentials* [13]. These potentials make it possible to specify a per-pixel weight which expresses how important a pixel is for preserving object consistency. We show how top-down segmentations from a densely sampled part-based detector [9, 21] can be used to define those weights. As our results demonstrate, the proposed approach yields improved segmentation and image labeling performance.

The paper is structured as follows. The following section discusses related work. Section 2 then introduces the CRF framework for multi-label image segmentation. Section 3 presents our extended approach using generalized robust P^N potentials and Section 4 explains how we obtain the top-down segmentation information. Finally, Section 5 presents experimental results.

Related Work. Graphical models, and CRFs in particular, have proven to be a successful tool for multi-class image labeling. Texton features combined using joint boosting [25] or random forests [24] form strong discriminative classifiers. The output of these classifiers is then fed into a CRF framework in the form of unary potentials to create powerful segmentation systems such as TextonBoost [25].

The introduction of higher order potentials into the CRF framework, as proposed in [13], makes it possible to include context information in a joint optimization scheme. As a result, finer details can still be preserved and contour detail information can successfully be captured. Extensions to this approach have mostly focused on two directions. The first is the use of features which are not based uniquely on color information, such as structure-from-motion (SfM) [9, 26] and stereo depth information [18]. The second direction is the use of multiple hierarchies in the CRF framework. The separate levels of each hierarchy correspond to different image quantizations such as pixels, superpixels, segments, *etc*. The resulting graphical model is then jointly optimized using an associative strategy [16].

The observation that most of the aforementioned methods achieve poor performance on classes containing objects, gave rise to combinations of segmentation algorithms with object detections [12]. The class-specific information that an object detector can provide has been



(a) (b) (c) (d)

Figure 2: **Failure case for GrabCut segmentation. (a)** Test image with two detected cars. GrabCut **(c)** fails to segment a car from the background, resulting in a suboptimal segmentation compared to the ground truth labeling **(b)**. Our method **(d)** obtains better results using information from top-down segmentations in the form of generalized robust P^N potentials

used to segment an object from its background [15, 20, 22] in images which contain at least one instance of the known object class.

However, very recently multi-class image labeling algorithms have been also combined with object detectors for the goal of building systems for scene analysis and scene understanding. This has first been done by Wojek *et al*. [28] and Ess *et al*. [7], who take advantage of the temporal consistency of a video sequence and combine image labeling with object detection for analyzing an urban environment scene. However, the resulting graphical models become quite complex and therefore difficult to optimize with standard Graph-Cuts methods. Recently, other works integrate also geometry [10] and physics information [11] to improve performance on the scene understanding task.

The work that is most closely related to our approach is that of Ladicky *et al*. [17]. In this work, object detections are incorporated into a CRF framework and inference is performed using Graph-Cuts. The integration of object detections at the pixel-level is achieved using hard segmentations provided by the GrabCut algorithm [23]. Classcut [1] and Grabcut in combination with the sum-product algorithm for doing the probabilistic inference could be also used as alternatives to provide soft segmentations in case the objects of interest are of an unknown class. However, there are object detection systems available which provide more detailed, object-specific, soft, top-down segmentations [21]. In this paper, we use a dense extension of the ISM detector [21] that is based on the idea of Hough Forests [9].

2 CRFs for Multi-Class Image Labeling

Markov Random Fields (MRFs) and Conditional Random Fields (CRFs) have been extensively used for modeling multi-class image labeling problems. In this section we provide a brief description of the CRF framework and the relevant notation. We closely follow the notation convention used in [14].

Consider a set of random variables $\mathbf{Y} = \{Y_1, Y_2, ..., Y_N\}$. The random field is defined on a graph $\mathscr{G} = (V, E)$ consisting of N pixels (*i.e*., $|V| = N$), where each node of the graph is associated with a random variable Y_i. Each variable $Y_i \in \mathbf{Y}$ has a domain $\mathscr{L} = \{l_1, l_2, ..., l_k\}$, where the value of Y_i represents a region assignment for pixel i (*e.g*., building, car, road, sky). A label assignment $\mathbf{y} = \{Y_i | i \in V\}$ with components lying in $\mathscr{L}$ represents a sample from the configuration space $\mathbf{Y}^N$.

We denote the probability assigned to vector $\mathbf{y}$ of the configuration space by $p(\mathbf{y})$. The random variables are said to form an MRF if $p(\mathbf{y})$ is strictly positive for all $\mathbf{y}$ and $p(y_i | y_{\mathscr{N}_i}) = p(y_i | y_{V \setminus i})$, where $\mathscr{N}_i$ represents all neighboring nodes of i in $\mathscr{G}$, while $V \setminus i$ denotes all nodes of the graph $\mathscr{G}$ except for i.

A CRF is an MRF globally conditioned on the set $\mathbf{X}$ of observed variables [19]. The

distribution $p(\mathbf{y}|\mathbf{X})$ over the discrete random vector $\mathbf{y}$ of the CRF is a Gibbs distribution and has the following form: $p(\mathbf{y}|\mathbf{X}) = \frac{1}{Z}\exp(-\sum_{\mathscr{C}\in\mathbf{C}} V_{\mathscr{C}}(\mathbf{y}))$, where Z is a normalizing constant and $\mathbf{C}$ denotes the set of all *maximal cliques* in the graph $\mathscr{G}$. A clique is defined to be any complete subgraph of the nodes V of $\mathscr{G}$. The symbol $V_{\mathscr{C}}(\mathbf{y})$ represents the clique potentials which constitute real functions depending only on the random variables contained in the clique $\mathscr{C}$. According to Maximum A Posteriori (MAP) estimation, the most probable label assignment $\hat{\mathbf{y}}$ of the CRF is obtained as $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathbf{Y}^N} p(\mathbf{y}|\mathbf{X}) = \arg\min_{\mathbf{y}\in\mathbf{Y}^N} E(\mathbf{y})$, where $E(\mathbf{y})$ is the corresponding Gibbs energy.

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

$$E(\mathbf{y}) = \sum_{i\in V}\psi_i(y_i) + \sum_{(i,j)\in E}\psi_{ij}(y_i,y_j) + \sum_{c\in\mathscr{S}}\psi_c(\mathbf{y}_c), \tag{1}$$

which has been shown to achieve state-of-the art performance for the multi-class image labeling problem [13]. 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 [13], those segments are created by an unsupervised multi-level mean-shift segmentation [6]. 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. More specifically, the robust P^N potentials are defined as:

$$\psi_c(\mathbf{y}_c) = \begin{cases} N_i(\mathbf{y}_c)\frac{1}{Q}\gamma_{max} & \text{if } N_i(\mathbf{y}_c) \le Q, \\ \gamma_{max} & \text{otherwise}, \end{cases} \tag{2}$$

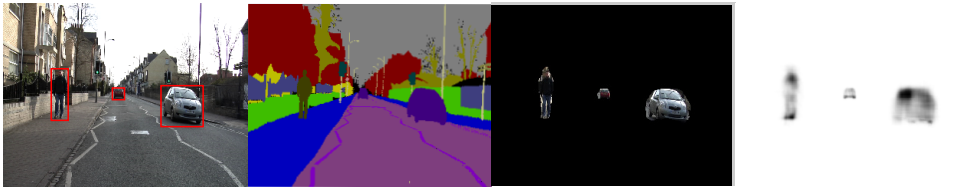
where $N_i(\mathbf{y}_c)$ represents the number of pixels in the segment c that are inconsistent with the majority label; Q is the truncation parameter; and $\gamma_{max} = |c|^{\theta_a}(\theta_p^h + \theta_v^h G(c))$, as defined in [13].

The particular strength of this formulation lies in the fact that higher order potentials capture the fine texture and contour details better than the pairwise potentials, thus providing impressive results particularly on label classes that represent *stuff* [12] (*e.g.* road, sky, grass). However, lacking any object specific information, such methods underperform on classes representing *things* (*e.g.* cars, pedestrians). Recently, Ladicky *et al.* [17] proposed to incorporate the output of an object detector into the CRF framework as a higher order potential. In particular, GrabCut segmentation [23] is used to obtain the foreground and background pixels inside the detection bounding box provided by a sliding window object detector. Each detection hypothesis h_d consists of a detection score H_d and a set of pixels belonging to the specific object. An auxiliary binary variable x_d is introduced for each object detection hypothesis indicating whether this hypothesis should be adopted or not. Hypotheses having a low score are filtered using a threshold H_t.

The detector potential then takes the following form:

$$\psi_d(\mathbf{y}_d, H_d, l_d) = \min_{x_d\in\{0,1\}} (-f(\mathbf{y}_d, H_d)x_d + g(N_d, H_d)x_d) \tag{3}$$

where $f(\mathbf{y}_d, H_d) = w_d|\mathbf{y}_d|\max(0, H_d - H_t)$ represents the strength of the hypothesis and $g(N_d, H_d) = \frac{w_d}{p_d}\max(0, H_d - H_t)N_d$ represents the cost for partial inconsistency as a linear function of the number of inconsistent pixels. As a result, this method manages to integrate object detections into a CRF framework, which can be optimized using standard Graph-Cuts algorithms like α-expansion [13].



(a) (b) (c) (d)

Figure 3: **Top-down segmentations. (a)** Results of the part-based detector of [8] applied to a test image, producing two detections for cars and one for pedestrians. The detection bounding boxes are then fed into the Hough Forests framework described in Section 3, which provides top-down segmentations **(c)** together with foreground probabilities **(d)** in the form of per-pixel weighting factors. As can be seen, the top-down segmentations are of comparable quality as the ground truth annotation **(b)**.

However, the integration of the object detections is not optimal in several respects. Firstly, the introduction of the GrabCut segmentation makes a hard decision for the foreground/background pixels inside each detection bounding box without considering object-specific nor context information. Secondly, the later CRF optimization may not be able to compensate for errors made during the assignment of foreground and background labels to the pixels inside the object, since only the number of consistent pixels counts, but not their importance for the detection. We argue that only one segmentation stage should take place in the processing pipeline, enabling joint optimization for the object existence and the scene segmentation. Figure 2 illustrates an example where the GrabCut segmentation fails to extract the pixels corresponding to the object inside the bounding box. Using more detailed information about the object from a detector-specific top-down segmentation, these problems can be overcome.

3 Image Labeling with Generalized Higher Order Potentials

As discussed in the previous section, we argue that object detections should be integrated into the CRF in a fashion that enables joint optimization in a single segmentation stage. For this, we build upon the idea of generalized robust P^N potentials [13].

Definition. *Generalized* robust P^N potentials provide a structured framework for incorporating the class-specific information provided by an object detector. As introduced in [13], they take the following form:

$$\psi_c(\mathbf{y}_c) = \min\left\{\min_{k\in\mathscr{L}}((P - f_k(\mathbf{y}_c))\theta_k + \gamma_k), \gamma_{max}\right\}. \tag{4}$$

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

$$P = \sum_{i\in c} w_i^k, \quad \forall k \in \mathscr{L} \tag{5}$$

$$f_k(\mathbf{y}_c) = \sum_{i\in c} w_i^k \delta_k(y_i) \tag{6}$$

$$\delta_k(y_i) = \begin{cases} 1 & \text{if } y_i = k, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where the weights w_i^k should have the following properties: $w_i^k \geq 0, i \in c, k \in \mathscr{L}$. These 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.

Generalized Top-Down Segmentation Potentials. Top-down figure-ground 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 (Fig. 3). These soft decisions can also be interpreted as weights indicating the importance of each pixel in the preservation of the object's label consistency. In particular, if a pixel has a high foreground probability p_{fig} then a higher cost should be assigned in the case that this pixel takes a label different from the object's. On the other hand, in the case that a pixel has a low probability being part of the foreground a lower cost is assigned for this pixel not taking the object's label.

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. Thus, the new energy function of our proposed model takes the form:

$$E(\mathbf{y}) = E_{pix}(\mathbf{y}) + \sum_{c \in \mathscr{S}'} \psi_c(\mathbf{y_c}), \tag{8}$$

where $E_{pix}(\mathbf{y})$ summarizes the pixel-based energy terms from eq. (1). $\mathscr{S}' = \mathscr{S} \cup \mathscr{D}$ is the union of the segments $\mathscr{S}$ produced by unsupervised mean-shift segmentations (similar to [13]) and the detections $\mathscr{D}$, together with their corresponding top-down segmentations given by an object detector.

The generalized higher order potentials used in our model can be divided into two categories. The first category contains the higher order potentials capturing the consistency of the image segments coming from the unsupervised segmentations. In this case the pixels within the segment are weighted uniformly and the definition of γ_{max} and γ_k follows the approach of [13]. The second category contains the higher order potentials coming from object detection. The weights in this category come from the p_{fig} probability of the top-down object segmentations. Following [17], we define γ_{max} to be proportional to the detection hypothesis' strength and γ_k inversely proportional to the cost taken for partial inconsistency.

Implementation. The unary potentials are a combination of densely computed local features such as location, color, texton and HOG features. We follow the procedure of [25] to extract a bag of textons by convolving a 17-dimensional filter bank with the training images. The filter responses are then clustered and a texton map is formed by assigning each pixel to the closest cluster center. Similar maps are obtained by clustering other types of features. A set of shape filters are formed as triplets of the feature type, the cluster number and a rectangular region. The responses to these filters are collected and corresponding weak classifiers are formed by comparing the shape filter responses to a set of predefined thresholds. The weak classifiers are then fed into a joint boosting framework which learns a strong classifier out of the weak ones. For pairwise potentials between the pixels of the image a standard contrast sensitive Potts model is used [23]. Inference is performed in the CRF using standard α-expansion [2, 3, 13]. The values of the variables x_d are then computed similar to [17].

4 Obtaining the Top-Down Segmentation

An important question is how to obtain the top-down segmentations, such that they fulfill the properties described in the previous section. In the following, we show how this can be realized using the ISM top-down segmentation formalism [21] on top of a dense Hough Forest classifier [9].

This approach builds up a random forest classifier that processes densely sampled image patches. Each internal node of each tree performs a simple binary test, comparing the values of two pixel locations in one of 32 feature channels (we use the same features as [9]). As patches are passed down the tree, they are sorted into visually similar groups. The leaf nodes thus correspond to the entries of a discriminatively trained visual codebook. The idea behind Hough Forests is to now 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, similar to [21].

As shown in [21], 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 back to the patches they were originating from. This process can be used to infer local figure-ground labels [21], object part annotations, or even depth maps or surface orientations from a single image [27]. In our approach, we extend the Hough Forest classifier with this top-down segmentation formalism, using figure-ground labels learned from annotated training examples. The resulting procedure is very intuitive to implement. 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 patch $Seg(v_j)$. 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}.

$$\begin{aligned} p(\mathbf{p}=fig|h) &= \frac{1}{\sum_{v_j\in h} w_{v_j}} \sum_{\mathbf{P}_i(\mathbf{p})} \frac{1}{|\mathbf{P}_i|} \sum_{v_j\in votes(\mathbf{P}_i)} w_{v_j} Seg(v_j) \\ p(\mathbf{p}=gnd|h) &= \frac{1}{\sum_{v_j\in h} w_{v_j}} \sum_{\mathbf{P}_i(\mathbf{p})} \frac{1}{|\mathbf{P}_i|} \sum_{v_j\in votes(\mathbf{P}_i)} w_{v_j} (1-Seg(v_j)) \end{aligned} \tag{9}$$

As shown in [21], this results in the correct probabilities. In practice, the top-down procedure thus boils down to a simple weighted summation of figure-ground patches, which can be implemented very efficiently on the GPU. Example results are shown in Fig. 3.

In our experiments, we build up Hough Forest detectors for pedestrians and cars using the procedure described above. In principle, the detector outputs, together with their top-down segmentations, can directly be used to define the generalized P^N potentials. However, we later on want to compare to a baseline using the part-based detector of [8]. In order to provide a fair comparison, we therefore use the detection bounding boxes returned by [8] in order to *query* the corresponding location in the Hough voting space and compute the resulting top-down segmentations for those locations.

5 Experimental Results

Datasets. We evaluate our method on the CamVid database [5]. The CamVid database contains over 10 minutes of high quality 30 Hz footage at 960×720 pixel resolution. The videos were captured by a driving car equipped with a camera mounted inside the vehicle. The database is divided in four video sequences, three captured at daylight (`0006R0`, `0016E5`, `Seq05VD`) and one captured at dusk (`0001TP`). Ground truth labeled frames are provided for these four sequences at a rate of 1 fps. The pixels in the annotated frames take one of 32 label classes. We closely follow the experimental setup of [4, 17, 26] by using the same subset of 11 class categories, dividing the data into 367 training and 233 test images, and downscaling all images by a factor of 3. Object detections are provided for two classes, namely cars and pedestrians, which are the main traffic participants.

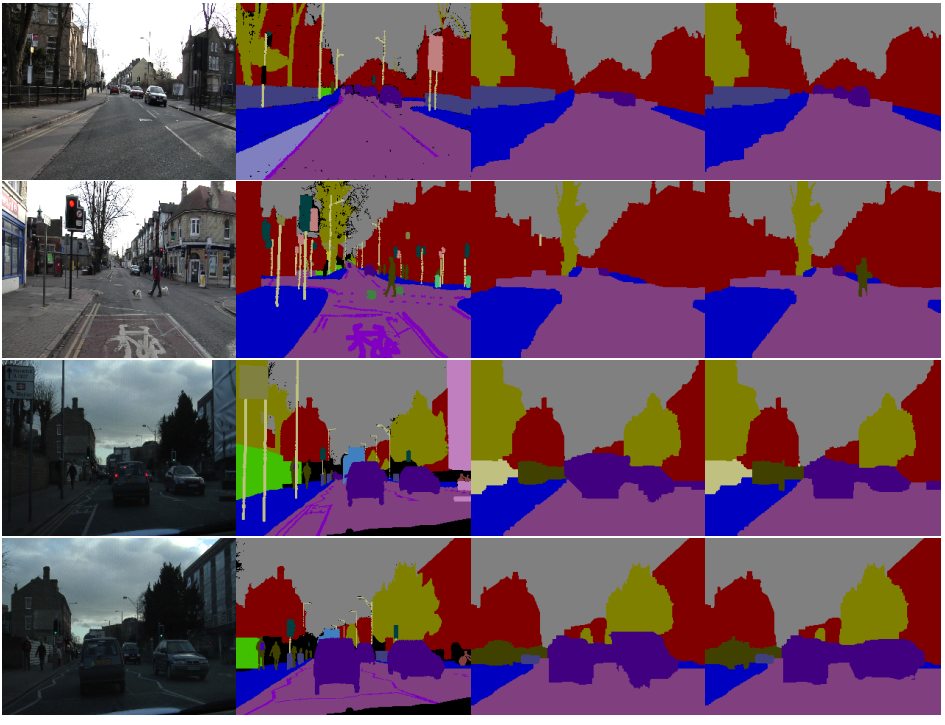
	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Sidewalk	Bicyclist	**Global**	**Average**
Recall													
[9] (SfM)	*46.2*	*61.9*	*89.7*	*68.6*	*42.9*	*89.5*	***53.6***	*46.6*	*0.7*	*60.5*	*22.5*	*69.1*	*53.0*
[26] (SfM)	***84.5***	*72.6*	***97.5***	*72.7*	*34.1*	***95.3***	*34.2*	*45.7*	*8.1*	*77.6*	*28.5*	***83.8***	*59.2*
[17] (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
[17] (with 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**
Our reimpl. [17]	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
Our approach	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
Inter. *vs* Union													
[26] (SfM)	***71.6***	*60.4*	***89.5***	*58.3*	*19.4*	*86.6*	***26.1***	*35.0*	*7.2*	*63.8*	*22.6*	-	*49.2*
[17] (no det.)	70.0	63.7	**89.5**	58.9	17.1	86.3	20.0	35.8	9.2	**64.6**	23.1	-	48.9
[17] (with det.)	**71.5**	**63.7**	89.4	64.8	**19.8**	86.8	23.7	35.6	**9.3**	**64.6**	**26.5**	-	**50.5**
Our reimpl. [17]	69.7	62.7	89.0	58.7	13.1	86.2	20.0	35.3	6.6	62.3	15.3	-	47.2
Our approach	70.8	63.2	89.3	**71.2**	13.2	**86.8**	**25.7**	**35.8**	6.7	63.8	16.1	-	49.3

Table 1: Quantitative results on the CamVid sequence. 'Global' performance refers to the overall percentage of pixels correctly classified, and 'Average' is the average of all the class measures. 'no det' indicates that no detections are availabale and 'with det' indicates that detections are available. Percentages in *italics* indicate results obtained from methods using SfM features. These approaches are not directly comparable to our method, because we do not include this cue in our approach. It can be clearly seen that our method performs better than the baseline approach on all categories. The inclusion of the top-down segmentations in the form of generalized P^N potentials is shown to be beneficial, as we outperform [17] on the categories for which object detections are available (cars, pedestrians). It is important to mention that results from [17] have been computed using detector responses for all the 5 *things* classes. This fact explains their superior performance on bicyclists, sign-symbols and column-poles.

Experimental setup. In order to make a fair comparison to the approach of [17], we use the following experimental setup. We employ the state-of-the-art part-based object detector from [8] to obtain the same detection bounding boxes for cars and pedestrians. We then tried to faithfully reimplement the algorithm proposed in [17] as a baseline. For that we perform GrabCut segmentation inside each detection bounding box to assign foreground or background labels to the pixels inside it. The higher-order potentials are finally formed exactly as in [17]. The difference in the implementation lie in the fact that we do not use a hierarchical CRF. Since there will invariably be small differences between Ladicky's original parameters and our reimplementation, we report comparisons to both baselines wherever possible.

For our method we take as input the same detections bounding boxes [8], from which we query top-down segmentations, as described in Section 4. The higher-order potentials are then formed as described in Section 2. The unary and pairwise potentials are the same for both approaches. Both the baseline method and the proposed one have been implemented in C++. The current implementation takes around 6.5 hours to train on the CamVid dataset and 20-25 seconds for solving the graph-cuts optimization problem to infer the image labeling of a test image on a Intel Core 2 Duo, 2.0GHz, 3GB RAM laptop.

Discussion of Results. An overview of the experimental results on the CamVid dataset is shown in Table 1. For comparison, we list the results of [9, 17, 26] reported on the same dataset. Note that [9, 26] use structure-from-motion (SfM) features, which are not available for our approach. It can be seen that our baseline implementation achieves slightly lower performance than [17] on some *stuff* categories. This can be explained by two reasons.



(a) (b) (c) (d)

Figure 4: **Qualitative results on CamVid. (a)** Test images. **(b)** Ground truth labeled images. **(c)** Segmentations obtained using the baseline approach of [7]. **(d)** Segmentations obtained using our method. As can be seen, our approach achieves better segmentations for the cars and pedestrians classes. (**Best viewed in color**)

Firstly, we use a simple one-layer CRF, instead of a two-level hierarchical CRF as in [7]. Second, we do not use dedicated detectors for `column poles`, `sign-symbols`, and `bicyclists`, since those classes would require additional training data that we do not have. The global performance level is however very similar (82.3% *vs*. 83.8%), which is why we are confident our results will transfer also to the full system of [7].

When replacing the GrabCut segmentations by our proposed top-down segmentations, our approach consistently improves upon the baseline, indicating that the proposed integration of top-down segmentations indeed improves the overall results. Also compared to [7], our approach achieves significant improvements for the `car`, `pedestrian`, and `road` categories, where our improved segmentations show most effect. Even though the baseline we started from was slightly below the one from [7], this improvement is strong enough such that our approach reaches a similar global performance level (83.2% *vs*. 83.8%). The improvement is also seen visually in the form of more accurate segmentations, as illustrated in Fig. 4 and in the supplementary material. Together, those results demonstrate the advantage of our proposed use of *generalized* robust P^N potentials.

6 Conclusions

We have presented a novel framework for combining object detections with a CRF framework for the multi-class image labeling problem. Top-down segmentations in the form of

generalized higher order potentials deliver soft object segmentations on the pixel-level and therefore enable joint optimization in a single energy minimization formulation. We have evaluated our approach on the CamVid dataset achieving state-of-the-art performance, as well as good generalization capabilities. We achieve better performance on the classes for which object detections are available, which indicates the potential of the proposed approach.

As future work, we will explore extensions of our current system into the time domain, as we believe that temporal consistency conveys important information from which our system could benefit.

Acknowledgements

This project has been funded, in parts, by the EU project EUROPA (ICT-2008-231888) and the cluster of excellence UMIC (DFG EXC 89).

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Classcut for unsupervised class segmentation. In *ECCV*, 2010.

[2] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9):1124–1137, 2004.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2002.

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

[5] 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.

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

[7] A. Ess, T. Mueller, H. Grabner, and L. van Gool. Segmentation-Based Urban Traffic Scene Understanding. In *BMVC*, 2009.

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. In *CVPR*, 2008.

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

[10] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009.

[11] A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010.

[12] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008.

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

[14] D. Koller and N. Friedman. *Probabilistic graphical models*. MIT Press, 2009.

[15] M.P. Kumar, P.H.S. Torr, and A. Zisserman. OBJ CUT. In *CVPR*, 2005.

[16] L. Ladickỳ, C. Russell, P. Kohli, and PHS Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

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

[18] L. Ladickỳ, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and PHS Torr. Joint optimization for object class segmentation and dense stereo reconstruction. In *BMVC*, 2010.

[19] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[20] D. Larlus, J. Verbeek, and F. Jurie. Category Level Object Segmentation by Combining Bag-of-Words Models and Markov Random Fields. In *CVPR*, 2008.

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

[22] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. In *ECCV*, 2006.

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

[24] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[25] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1):2–23, 2009.

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

[27] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, and L. Van Gool. Shape-from-Recognition: Recognition Enables Meta-Data Transfer. *CVIU*, 113(12):1222–1234, 2009.

[28] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *ECCV*, 2008.