Are Spatial and Global Constraints Really Necessary for Segmentation?

Aurélien Lucchi[1] Yunpeng Li[1] Xavier Boix[2] Kevin Smith[1] Pascal Fua[1]
[1] Computer Vision Laboratory, EPFL, Lausanne
[2] BIWI, ETH Zurich

Abstract

Many state-of-the-art segmentation algorithms rely on Markov or Conditional Random Field models designed to enforce spatial and global consistency constraints. This is often accomplished by introducing additional latent variables to the model, which can greatly increase its complexity. As a result, estimating the model parameters or computing the best maximum a posteriori *(MAP) assignment becomes a computationally expensive task.*

In a series of experiments on the PASCAL and the MSRC datasets, we were unable to find evidence of a significant performance increase attributed to the introduction of such constraints. On the contrary, we found that similar levels of performance can be achieved using a much simpler design that essentially ignores these constraints. This more simple approach makes use of the same local and global features to leverage evidence from the image, but instead directly biases the preferences of individual pixels. While our investigation does not prove that spatial and consistency constraints are not useful in principle, it points to the conclusion that they should be validated in a larger context.

1. Introduction

Segmenting natural images into semantically consistent regions is of fundamental importance to computer vision and image understanding, and good performance on the PASCAL [4] and MSRC [26] datasets has become a *de facto* standard for success. A number of recent approaches have pushed the state of the art on these data sets by introducing sophisticated graphical models that include constraints on both local spatial smoothness and global consistency.

Markov Random Fields (MRF) and Conditional Random Fields (CRF) are at the heart of many modern segmentation approaches. A recent trend has been to model global constraints as latent variables, which interact with local variables either directly or through intermediate hierarchies.

However, these global variables are usually coupled with *global features* which are computed at a global or very large scale. Thus, it unclear whether it is the constraints or the features that give these models their power. We therefore ask the question, what happens when we collapse the superstructures of global variables and incorporate their features directly into the local variables with which they interact?

In this paper, we show that in doing so, we obtain very similar performance levels by using only global features that already appear in the literature [8] to leverage evidence from the image and bias the preferences of individual pixels or superpixels. This results in a much simpler model than those found in other recent approaches, which rely on complex hierarchies and global context models. Moreover, when using these global features, removing the local spatial smoothness term only results in minimal performance degradation.

To demonstrate our point, we compare increasingly complex versions of a state-of-the-art CRF-based segmentation algorithm patterned after the model proposed by Gonfaus *et al*. [8] that has been shown to yield excellent performance on the PASCAL and the MSRC datasets. As will be shown, for the MSRC dataset, the simplest model which relies exclusively on image features outperforms state-of-the-art models which enforce spatial smoothness and global consistency constraints. On the PASCAL dataset, the same simple model shows only a slight decline in performance.

This is not to say that spatial constraints or other kinds of global constraints are not useful in principle. However, it does suggest that recently reported results do not conclusively demonstrate their usefulness, at least in the context of the MSRC and PASCAL datasets.

2. Motivation and Related Work

Markov Random Fields were originally introduced as generative models [2], where each variable is exclusively associated with its own observation. In the context of image segmentation, this is to say that the data term, or "unary potential" as it is sometimes referred to, of a given superpixel [1] can only draw evidence from within itself. This requirement has proved to be too stringent for most vision tasks includ-

[1] For simplicity, we will treat superpixels as atomic image regions for the remainder of the text. However, for most algorithms, pixels can be used interchangeably.

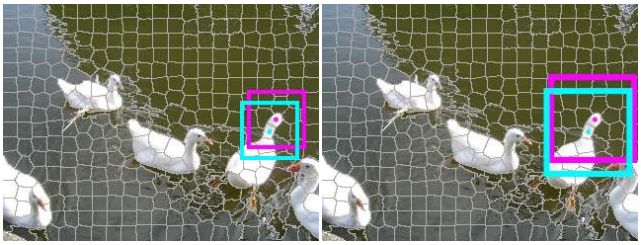
ing segmentation, and recent approaches have opted for the more flexible Conditional Random Field (CRF) [15], which allows the label of a superpixel to depend on features collected from itself and its neighbors.

The Potts model is commonly used to enforce local spatial smoothness constraints for image segmentation (Fig. 2). Despite its continued popularity, there has been a considerable amount of work searching for more sophisticated spatial smoothness terms. Gould *et al*. [9] use relative location features to model class-specific spatial dependencies between pixels. These are used to supplement appearance based features when producing the final segmentation. In their model, the smoothness term is pre-computed at a separate stage, not modeled jointly with the data term during inference. Batra *et al*. [1] also learn class-specific affinities in a CRF framework, where these affinities model the relationship between visual words instead of between pixels. Galleguillos *et al*. [7] model the full joint transition likelihood between neighboring pixels, where the parameters of their spatial model are set via simple counting of co-occurrences in the labeled images. Though this is arguably sub-optimal, the method was shown to perform well in practice.

It is well known that leveraging data from larger support gives features more discriminative power. However, this has a subtle side effect: *it reduces the significance of the spatial term*, namely the smoothness or consistency constraints that are encoded as edges of the CRF. As shown in Figure 1, if the features of a superpixel are computed from a neighborhood much greater than the superpixel itself, the smoothness constraint between adjacent superpixels becomes almost redundant since they are associated with highly correlated, largely overlapping observations.

Indeed, several other methods have found success using simpler models that do not enforce smoothness, raising the question about the necessity of more elaborate CRFs. It was shown in [24] that much greater performance gains could be attained by finding powerful image features rather than enforcing spatial constraints. This finding was echoed by Verbeek and Triggs [29] who concluded that complex smoothness priors help only in the absence of global contextual information provided by image-wide aggregate features, and by the success of the Semantic Texton Forests [25] which do not involve any spatial model.

Consequently, recent CRF models for segmentation have shifted their focus from regularizing local smoothness to encouraging global consistency [13, 14, 8]. A further motivation for this is the belief that information stored in the local nodes is incapable of capturing the "big picture" or overview of the scene. To do this, Shotton *et al*. [25] train an image-level classifier and use its prediction to modify the probabilities given by the local pixel-level classifiers. Similar to many other approaches [10, 22, 16, 21], the image-level cues are inferred separately from the labels of the pixels. Despite the lack of any spatial constraints, this method achieved impressive performance that surpassed the state-of-the-art. Joint inference of pixel labels and image-level preference was studied by Kohli *et al*. [13], where robust P^N potentials are used to encourage consistency between the labels of local and global variables. Ladicky *et al*. [14] further propose an associative hierarchical CRF model, which has a more elaborate structure that includes intermediate layers in between. In these approaches, the label sets for the global nodes are limited to the set of semantic classes, possibly augmented by an additional "background" label. This restricts image-level preference to a single class at most, ignoring scenarios where there is more than one dominant object type in the image.



(a) (b)

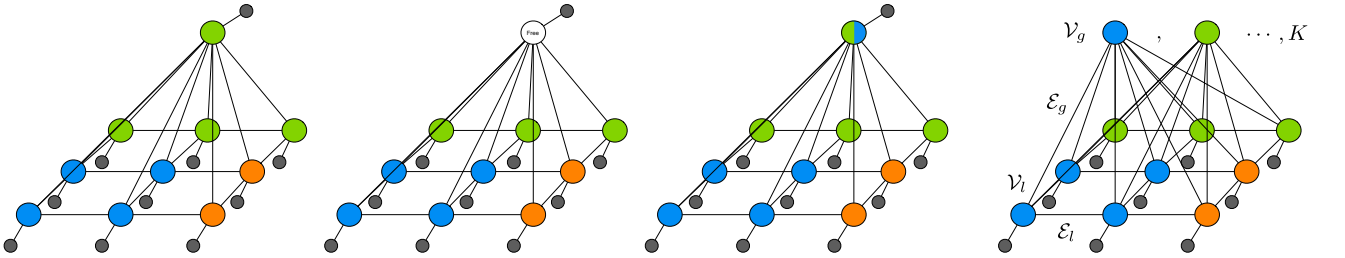
Figure 1. *Spatial support of local image features for neighboring superpixels computed at different scales.* The bounding boxes show the extent of the spatial support of the *local features* described in Section 4.1, which are similar to the features used in [8] and other works. In *(a)*, one side of the bounding box is 4 times the mean superpixel width. In *(b)*, it is 6 times the mean superpixel width. Because the regions significantly overlap, their class predictions will be highly correlated. This has the effect of reducing the impact in performance attributed to the spatial term.

To address this issue Gonfaus *et al*. [8] use more expressive constraints called "harmony potentials", which model global preferences using the power set over all semantic classes. Although this makes it possible to have multiple preferences at the global level, the exponential sized power set is prohibitively expensive to search and has to be heuristically truncated to make it computationally affordable. Hence in practice, only a small subset is used.

An important but often overlooked detail of these hierarchical models is the use of specialized global features designed to enforce global consistency. As we will show, directly embedding these features into a much simpler graphical model results in similar performance.

3. Segmentation using CRF Models

In order to investigate the effect of imposing spatial and global constraints, we designed a CRF model inspired by the one proposed by Gonfaus *et al*. [8], shown in Fig. 2(d).



(a) Potts model (b) Robust P^N model [12] (c) Harmony model [8] (d) Class independent model (DPG)

Figure 2. *Increasingly more sophisticated CRFs for modeling global preference*. Standard CRFs include local nodes $\mathcal{V}_l$ connected by edges $\mathcal{E}_l$. Observations are denoted by gray nodes. Class labels are indicated by color. High level preferences can be encoded by adding global nodes $\mathcal{V}_g$ connected by edges $\mathcal{E}_g$. **(a)** The Potts model penalizes all the local nodes with a label different from the global node. **(b)** The robust P^N potential is similar to the Potts model but adds an extra "free label" label that does not penalize local nodes. **(c)** The harmony potential allows different labels to coexist in a power set. However, the size of the power set makes optimization difficult. **(d)** The class independent model (DPG) used in this work models each of the K classes with its own global node to make the inference more tractable.

In Section 4, we will explore the contributions of the data term, spatial constraints, and global preferences by comparing increasingly complex versions of this model.

3.1. CRFs with Global Preferences

The standard CRF model can be extended to incorporate global consistency constraints as seen in Fig. 2(d). Such models typically contain a set of nodes $\mathcal{V} = (\mathcal{V}_l, \mathcal{V}_g)$ where the global nodes $\mathcal{V}_g$ encode high-level preferences in addition to local nodes $\mathcal{V}_l$ which represent superpixels.

The edges $\mathcal{E} = (\mathcal{E}_l, \mathcal{E}_g)$ represent interactions at two different levels. $\mathcal{E}_l$ model the relationship between neighboring local nodes. $\mathcal{E}_g$ link local nodes to global nodes, which serve to bias local labels to be consistent with global preferences. Thus, the extended energy function takes the form:

$$E_w(Y|X) = \sum_{i\in\mathcal{V}_l} \underbrace{D_i(y_i)}_{\text{data term}} + \sum_{(i,j)\in\mathcal{E}_l} \underbrace{P_{ij}(y_i, y_j)}_{\text{pairwise term}} + \sum_{(i,g)\in\mathcal{E}_g} \underbrace{G_{ig}(y_i, y_g)}_{\text{global term}} \quad (1)$$

where $y_i \in \{1, \cdots, K\}$ are class labels for superpixels and $y_g \in \{0, 1\}$ represent the states of global preferences.

Minimizing this energy function is equivalent to performing MAP inference on the CRF. Though this is in general NP-hard on graphs with loopy structures, good approximate solutions can be found using efficient energy minimization techniques (we use belief propagation [19]).

3.2. Energy Function

The energy function in Eq. 1 consists of three terms: the data term, the spatial term, and the global term, described below.

Data Term The data term $D_i(y_i)$ encourages agreement between a node's label y_i and the local image evidence x_i. We model it as a linear combination of the output scores given by S classifiers c_s, such as support vector machines (SVM) trained to predict the label of a superpixel. It is written as

$$D_i(y_i) = \sum_{s=1}^{S} w^D_{y_i,s} c_s(x_i, y_i). \quad (2)$$

Spatial Pairwise Term The pairwise term $P_{ij}(y_i, y_j)$ represents the cost of transition from class y_i to y_j, and is expressed in a non-parametric form as

$$P_{ij}(y_i, y_j) = w^P_{y_i,y_j}. \quad (3)$$

Like the standard contrast-dependent Potts model [26], this term encodes valid configurations according to the labels of neighboring nodes (y_i, y_j). It also considers the difference in color $||x_i - x_j||^2$ between pairs of superpixels, as well as their position relative to one another, allowing the model to capture geometric relationships such as "sky should appear above grass" (as illustrated in Fig. 3).

Global Term To enforce global consistency, a set of global nodes are introduced, resulting in a global term

$$G_{ig}(y_i, y_g) = w^G_{y_i,y_g}. \quad (4)$$

As shown in Fig. 2(d), this term expresses the dependency between the local nodes and K global nodes whose labels are inferred jointly with the local nodes. Our approach is similar to the harmony potential approach [8], except that it does not need to model the full power set of all class labels. This makes it more computationally tractable, but sacrifices the capability of modeling category co-occurrence. In contrast, other approaches like [12] can only bias the local nodes towards a single class label.

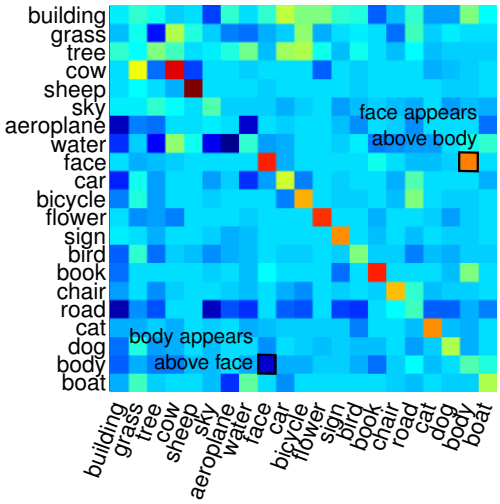


Figure 3. *Learned spatial relationships in the pairwise term.* The pairwise term $w^P(x_i, x_j, y_i, y_j)$ in matrix form where columns indicate classes y_i belonging to superpixel i and rows indicate classes y_j belonging to neighboring j for the MSRC-21 dataset. Red colors indicate that y_i is likely to appear above y_j. Left-of and right-of relationships are also learned but not shown here. Gradient information is also ignored for this illustration.

3.3. Parameter Learning

As in [27, 20], we express the energy function of the CRF (Eq. 1) in a linear form $E_w(X, Y) = w\Psi(X, Y)$, where $\Psi(X, Y)$ is a vector defined as $\begin{bmatrix}\Psi^D & \Psi^P & \Psi^G\end{bmatrix}$. Here, Ψ^D is a vector with S entries with $\Psi^D_s = c_s(x_i, y_i)$. For the pairwise interaction $\Psi^P(X, Y)$, assuming that the image X is ignored results in a K-by-K table of values where $\Psi^P(X, Y) = \sum_{(i,j)\in\mathcal{E}_l} \Psi^P_{i,j,y_i,y_j}$ is a sum of indicator variables in which the (a,b)-th entry is defined as

$$\Psi^P_{i,j,y_i,y_j}(a, b) = I(y_i = a, y_j = b). \tag{5}$$

Gradient information and spatial relationships are also included in $\Psi^P(X, Y)$. The global term Ψ^G is formulated in a similar manner.

To train the model, two learning steps are required. For the data term Ψ^D, S multi-class SVM classifiers corresponding to S local and global feature scales are trained. We also learn the parameters of the energy function, $w = \begin{bmatrix}w^D & w^P & w^G\end{bmatrix}$, which encode prior knowledge about relationships between the various object classes including pairwise relationships. We learn these parameters using the standard margin-rescaling Structured SVM (SSVM) [28].

The SSVM framework can be viewed as minimizing an upper bound on the average training loss (up to a constant factor C), as long as a labeling with cost no higher than that of the ground truth can be found for every training example. While this condition is not guaranteed due to the intractability of exact energy minimization on loopy graphs, an approximation can be found using efficient energy minimization techniques [27, 5].

The SSVM framework finds parameters w balancing model complexity and empirical loss. The loss function measures how incorrect a labeling H is compared to the ground truth Y. A natural choice of loss function is the 0-1 loss function that penalizes any error with the same weight without considering the labels: $\delta(h_i, y_i) = I(h_i \neq y_i)$. However, some classes occur more often, so in order to ensure a better balance, we weight errors inversely proportional to the frequency of a class

$$\delta(h_i, y_i) = \begin{cases} \frac{1}{frequency(y_i)} & \text{, if } h_i \neq y_i \\ 0 & \text{, otherwise.} \end{cases} \tag{6}$$

We also consider an alternative method to estimate w, which was proposed in [8]. In this approach, a Gibbs-like sampling algorithm changes a single parameter at each iteration. A new value is drawn from a Gaussian distribution with $\mu = 0$ and $\sigma = 1$. If the new parameter improves the score, it is kept. The drawback of this method is that it does not scale well with the number of parameters and therefore cannot be applied to high order CRFs. However, we found this simple method to be very competitive when only considering the data term, as we will see in the next section.

4. Experiments

We conducted experiments on two popular datasets for multi-class segmentation, the MSRC-21 [26] and PASCAL VOC 2010 [4] datasets. We compare five increasingly complex versions of the CRF model described in Sec. 3, as well as previously published models. Below, we provide details related to our implementation, experimental setup, and evaluation procedure.

4.1. Implementation

Prior to image segmentation, we extract features from the image. To do so, we over-segment the image into superpixels using the SLIC algorithm [23]. These superpixels correspond to local nodes in the CRF models. For each superpixel, we then extract *local features* at multiple scales, as well as *global features* over the entire image. These features are fed to classifiers in the data term, and a final solution is inferred using belief propagation.

Local features It has been shown that using a combination of features computed from both the superpixels and its surrounding area is more effective than using features just from the superpixel itself [6, 8]. Therefore, for each superpixel we extract a set of quantized visual words [30] over five different neighborhood scales, which provides a histogram-like descriptor. However, unlike [6, 8] who concatenated the features to create a single feature vector, we build a bag-of-words descriptor at each scale.

To build the bag-of-words representation, we extract patches over a grid with 50% overlap at several scales (12, 24, 36 and 48 pixels). These patches are described by shape (SIFT) and color (RGB histogram) features. We then use k-means to build a dictionary containing $1,000$ words for the shape features and 400 words for the color features. The assignment of a query patch to a dictionary term is done using a nearest neighbor search. The feature vectors are created for 5 different scales by extracting patches inside the superpixel alone, then extending the neighborhood size by factors of 1, 2, 4 and 6 respectively.

Global features Global features are similar to the local features, but extracted over the entire image. They are fed to a classifier who returns a single response for the whole image. For the VOC 2010 data set, we used a bag-of-words representation of the whole image, based on shape SIFT, color SIFT [3], together with spatial pyramids [17]. For MSCR-21, we used a simpler bag-of-words representation based on SIFT and RGB histograms.

Learning As described in Sec. 3.3, extracted features are fed to an SVM classifier trained such that its response $c(x_i, y_i)$ represents its perceived cost of assigning superpixel i to the class label y_i in the data term. We also train a structured SVM to learn the parameters w of the energy function $E_w(X, Y)$ in Eq. 1, which represents a linear combination of low-level classifier outputs on local (and global) features. The parameters cover all possible category combinations between two superpixels, two types of spatial relations (left-right and top-bottom), and 10 discretized gradient values. In total, the *DP model* incorporates 8862 parameter values, and the *DPG model* incorporates 9744 values. During training, we sample a total of 8000 superpixels for MSRC-21, and $20,000$ for VOC 2010 with equal numbers of positive and negative examples for each class.

4.2. Experimental Methodology

Data Sets The MSRC-21 dataset contains 591 images, with objects from 21 categories. To compare results with those of other methods, we use the standard split of the dataset [26]. The VOC 2010 dataset contains 20 object classes plus a background class. The images are divided into 3 subsets: training, validation, and testing.

Models Tested In order to better understand how spatial and global constraints affect performance, we tested five increasingly more complex versions of the CRF model described in Sec. 3. Descriptions of the models appear below.

- *D model* – Includes only the data term, consisting of the SVM classifiers scores. Equivalent to an energy function $E_w(Y|X) = \sum_{i \in \mathcal{V}_l} D_i(y_i)$.
- *DP model* – Considers both the data and pairwise terms of the energy function, i.e. $E_w(Y|X) = \sum_{i \in \mathcal{V}} D_i(y_i) + \sum_{(i,j) \in \mathcal{E}_l} P_{ij}(y_i, y_j)$.
- *DG model* – Considers the data term and the global term without the pairwise term, i.e. $E_w(Y|X) = \sum_{i \in \mathcal{V}_l} D_i(y_i) + \sum_{(i,g) \in \mathcal{E}_g} G_{ig}(y_i, y_g)$.
- *DPG model* – The full model described in Eq. 1, including the data, pairwise and global terms.
- *D-sampling* – Like the *D model*, only considers the data term. Instead of learning parameters using the SSVM, uses the sampling method of Sec. 3.3.

We also compared against four state-of-the-art CRF approaches including [25, 11, 14, 8] on the MSRC data set, and six reported methods for the VOC 2010 data set which are shown in Tables 1 and 2.

Local vs. Global+Local To test the effect of directly introducing global features to the data term on the various CRF models, we repeated each experiment twice. First, we provided only local features to the classifiers in the data term. In the second round, we included the global features as well.

Evaluation Metrics For MSRC-21, we measure performance for a given category by computing its pixel-wise classification accuracy. Overall performance is measured by averaging per-category classification accuracy across all categories. A global pixel-wise accuracy is also reported. For the VOC 2010 dataset, a similar procedure is used, but performance is measured by the *Jaccard index* instead of pixel-wise accuracy. The Jaccard index is the ratio of the areas of the intersection between what has been segmented and the ground truth, and of their union. It is written as

$$\text{VOC} = \frac{\text{True Pos}}{\text{True Pos} + \text{False Pos} + \text{False Neg}} . \quad (7)$$

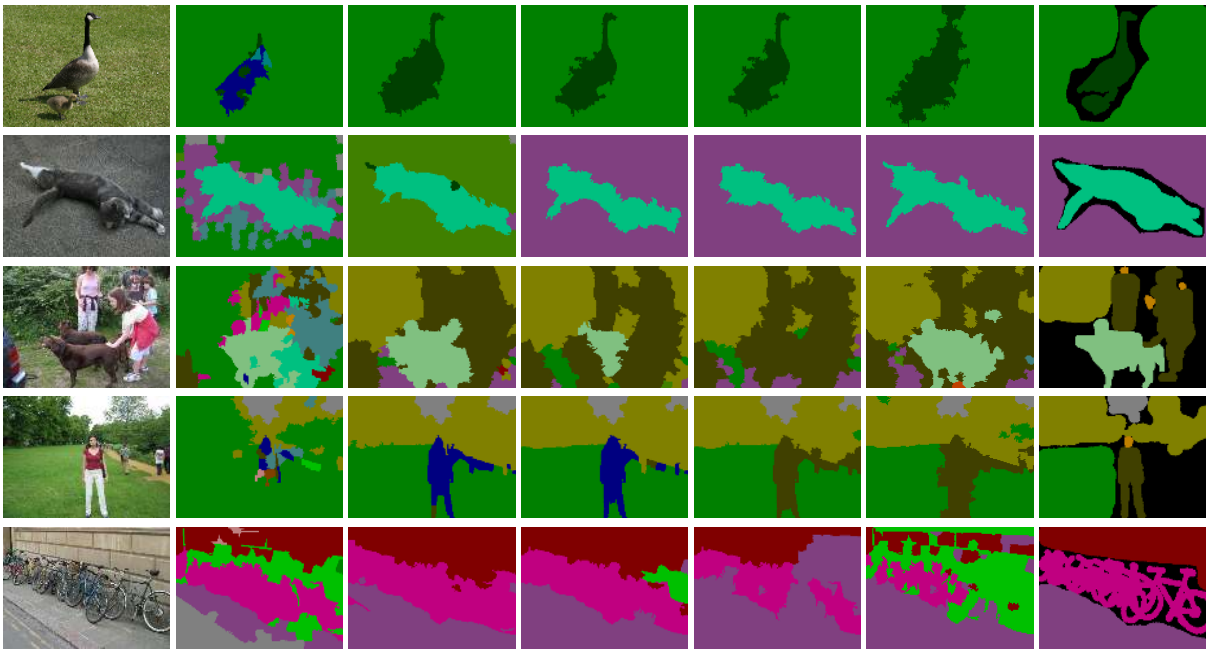
4.3. Results

MSRC-21 The results for MSRC-21 appear in Table 1. When only local features are considered, there is a clear advantage to adding spatial and global constraints (as indicated in red). The pairwise term alone leads to an increase by 6%, and adding the global term results in another 5% increase. The average per-category accuracy of the various CRF models ranged from 58% to 69%.

The second set of experiments introduces global features into the data term. Under these conditions, previous gains from adding the spatial and global constraints disappear, while the overall performance of all methods increased. In fact, the simple *D model* and *D-sampling* models now outperform the higher order CRFs. These results demonstrate

		building	grass	tree	cow	sheep	sky	airplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**Global**	**Average**
Local features S=5	*D model*	54	94	59	72	67	95	70	66	86	45	93	72	68	27	52	27	43	70	1	19	37	67	58
	DP model	53	86	72	79	75	95	93	49	85	38	81	83	64	39	63	49	68	68	38	63	9	70	64
	DG model	34	91	70	83	70	97	83	65	82	93	91	69	67	13	86	64	65	83	23	31	20	72	66
	DPG model	54	88	83	79	82	95	87	70	85	81	**97**	69	72	27	88	46	60	74	27	49	28	75	69
	D-sampling	52	83	74	50	72	89	86	68	73	69	83	67	69	22	68	25	67	54	14	46	50	69	61
Local+Global features S=6	*D model*	64	94	91	72	87	97	90	76	72	83	86	88	93	**62**	90	**89**	85	**97**	0	**83**	0	85	77
	DP model	58	87	83	73	78	94	95	78	85	68	96	89	71	41	96	83	85	87	49	52	38	80	76
	DG model	54	86	**93**	80	94	90	87	**88**	74	80	85	86	**96**	35	96	80	65	96	0	77	26	81	76
	DPG model	65	87	87	84	75	93	94	78	83	72	93	86	70	50	93	80	86	78	28	58	27	80	76
	D-sampling	50	83	87	81	84	90	**97**	72	75	79	90	95	79	52	**97**	81	80	89	51	64	**60**	79	**78**
	[25]	49	88	79	**97**	**97**	78	82	54	**87**	74	72	74	36	24	93	51	78	75	35	66	18	72	67
	[11]	53	**97**	83	70	71	98	75	64	74	64	88	67	46	32	92	61	89	59	66	64	13	78	68
	[14]	**80**	96	86	74	87	**99**	74	87	86	**87**	82	**97**	95	30	86	31	**95**	51	**69**	66	09	**86**	75
	[8]	60	78	77	91	68	88	87	76	73	77	93	**97**	73	57	95	81	76	81	46	56	46	77	75

Table 1. *MSRC-21 segmentation results.* For each category, the pixel-wise classification rate is provided. Bold entries indicate best performance. *Global* reports the pixel-wise classification rate over the entire data set. *Average* reports the mean of all category classification rates. The first five rows show results for the data term only (*D model*), the data and pairwise terms (*DP model*), data and global terms (*DG model*), and the full model (*DPG model*). *D-sampling* includes only the data term, but learns parameter values using Gibbs sampling instead of the SSVM. The second five rows show results when the global features are added. For reference, scores reported for other methods are reported in the last four rows. Scores in red indicate that when only local features are considered, there is an advantage to adding spatial and global constraints. Scores highlighted in yellow show that introducing global features eliminates the previous gains attributed to adding the spatial and global constraints, while increasing the overall performance of all methods.



(a) (b) (c) (d) (e) (f) (g)

Figure 4. *Example segmentations from the MSCR-21 dataset.* (a) Original images (b) *D model* with local features, (c) *DPG model* with local features, (d) *D model* with local+global features, (e) *DPG model* with local+global features, (f) *D-sampling* with local+global features, (g) Ground-truth.

that the presence of the global classifier in the data term can boost performance to levels similar to or even better than the more complex models that include spatial and global constraints. To the best of our knowledge, the *D model* trained with the sampling method achieves the highest average score ever reported on MSRC-21.

PASCAL VOC 2010 Results for the PASCAL VOC 2010 dataset appear in Table 2. The results mirror our findings on the MSRC-21 dataset. In red, we can see that when only local features are considered, there is a clear advantage to using spatial and global constraints. But the results highlighted in yellow show that providing global features to the

		Background	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dinning Table	Dog	Horse	Motorbike	Person	Potted Plant	Sheep	Sofa	Train	TV/Monitor	**Average**
VALIDATION SET																							
Local features S=5	*D model*	31.7	9.7	1.1	9.1	0.9	3.0	27.4	12.9	14.4	0.0	21.2	0.0	0.0	3.7	26.8	22.9	0.0	30.0	9.6	23.1	0.0	11.8
	DP model	29.3	13.0	2.2	2.8	7.7	12.8	36.6	24.3	19.7	3.6	19.2	0.2	14.6	1.9	25.3	17.7	1.3	12.6	6.1	19.1	5.5	13.1
	DG model	76.2	30.8	6.8	1.6	0.0	4.4	34.1	0.7	23.5	0.4	28.3	0.3	2.6	6.5	46.2	11.4	0.0	0.0	2.2	39.9	0.1	15.0
	DPG model	67.3	22.4	15.5	18.0	15.4	0.0	28.4	24.5	18.7	0.0	30.5	4.6	0.5	6.0	28.9	23.0	3.6	0.0	12.9	33.1	7.3	17.2
	D-sampling	76.2	23.3	11.7	6.4	6.8	9.0	24.4	24.0	13.6	3.2	11.4	0.0	14.7	10.3	24.1	24.5	4.0	20.3	3.8	14.6	11.7	16.1
Local+Global features S=6	*D model*	73.0	38.9	13.3	21.1	25.7	12.2	37.7	32.7	29.0	0.4	35.5	11.5	9.5	18.8	30.1	1.8	6.5	36.1	8.1	43.7	19.9	24.1
	DP model	76.7	29.8	16.8	6.0	21.6	13.2	39.8	33.0	16.1	0.0	25.5	0.5	21.8	13.8	45.1	34.1	3.0	35.0	2.9	47.2	26.3	24.2
	DG model	74.1	27.8	0.2	22.9	20.8	16.7	32.5	29.1	25.4	7.1	30.8	14.9	15.7	16.0	45.5	29.3	5.7	32.3	17.2	42.0	0.0	24.1
	DPG model	64.3	27.1	18.2	23.1	21.0	15.0	35.3	29.2	23.7	7.8	16.9	21.5	17.3	18.8	30.7	31.5	6.7	27.8	12.2	39.5	26.7	24.5
	D-sampling	78.8	44.1	21.0	16.9	28.7	24.8	59.3	40.0	30.3	7.0	26.8	6.8	18.2	17.0	35.2	34.3	31.2	18.7	11.5	**47.3**	18.1	29.3
TEST SET																							
	BONN SVR	**84.2**	52.5	27.4	**32.3**	**34.5**	**47.4**	60.6	**54.8**	**42.6**	**9.0**	32.9	**25.2**	27.1	32.4	47.1	38.3	**36.8**	**50.3**	**21.9**	35.2	40.9	**39.7**
	BROOKES	70.1	31.0	18.8	19.5	23.9	31.3	53.5	45.3	24.4	8.2	31.0	16.4	15.8	27.3	48.1	31.1	31.0	27.5	19.8	34.8	26.4	30.3
	STANFORD	80.0	38.8	21.5	13.6	9.2	31.1	51.8	44.4	25.7	6.7	26.0	12.5	12.8	31.0	41.9	**44.4**	5.7	37.5	10.0	33.2	32.3	29.1
	UC3M	73.4	45.9	12.3	14.5	22.3	9.3	46.8	38.3	41.7	0.0	35.9	20.7	**34.1**	**34.8**	33.5	24.6	4.7	25.6	13.0	26.8	26.1	27.8
	UOCTTI	80.0	36.7	23.9	20.9	18.8	41.0	**62.7**	49.0	21.5	8.3	21.1	7.0	16.4	28.2	42.5	40.5	19.6	33.6	13.3	34.1	**48.5**	31.8
	Harmony FG-BG	80.2	**57.0**	**28.7**	29.3	31.7	27.0	57.6	48.5	35.2	8.3	29.9	22.6	25.2	33.0	52.6	35.9	25.2	39.7	16.9	43.4	24.7	35.8
	DPG model	64.8	33.4	16.6	17.8	23.4	17.2	45.7	35.0	30.3	6.0	21.5	21.0	21.9	29.6	32.6	29.6	23.3	24.9	15.7	26.4	21.1	26.6
	D-sampling	77.9	49.4	23.1	19.2	24.8	26.1	52.4	44.9	32.9	6.5	**35.8**	22.3	25.5	21.9	**58.1**	34.6	26.8	39.9	17.5	38.0	25.3	33.5

Table 2. *PASCAL VOC 2010 segmentation results.* For each category, the Jaccard index is provided. Bold entries indicate the best performance. *Average* indicates the mean of the scores across all categories. The first ten rows show results reported on the validation set, for increasing CRF complexity, as in Table 1. Tests are made for local features only, as well as local+global features. Results highlighted in red indicate that local and global constraints improve performance when only local features are provided. But the results highlighted in yellow show that providing global features to the model eliminates the need for complex models. The last 8 rows compare the performance of our approach to other reported methods on the official VOC 2010 test set. Our models tend to underperform relative to the other methods because our learning procedure does not optimize for the Jaccard index.

model eliminates the need for complex models.

The lower section of the table compares our models to some of the best reported results on the PASCAL VOC 2010 test set. BONN SVR [18] obtained the best results, but their method tries to produce globally consistent segmentations by exploiting characteristic shapes of objects. However, they do not report results on the MSRC-21 dataset, which may prove to be more difficult for their model as it contains classes such as grass, building, water, and sky, which are difficult to characterize shapes. Our *D-sampling* model achieves a similar score to the second highest competitor, the Harmony potential model [8]. Note that our models trained with the structured SVM tend to underperform as they do not optimize for the VOC score.

5. Summary and Discussion

While we believe that spatial and global constraints are useful in principle, their relative weakness when compared to simpler models that consider global image features on the PASCAL and MSRC-21 datasets is worth reflecting upon. This suggests two possibilities for further consideration. First, we should reconsider the effectiveness of current approaches to modeling global and spatial constrains in CRF frameworks. Second, we should consider the possibility that these datasets, while useful, have shortcomings that need to be addressed if they are to be used to validate segmentation approaches that employ sophisticated constraints. For instance, in many images of the MSRC-21 dataset the ground truth is imprecise. This has the effect of arbitrarily penalizing correct labels near boundaries. While annotation quality of the PASCAL dataset is more precise, the current state-of-the-art on this dataset is such that state-of-the-art methods struggle to correctly label even half of the pixels. Performance differences attributed to the complexity of the CRF model may be overshadowed by other error sources that lead to such poor overall performance.

Resources used in this paper are publicly available at http://cvlab.epfl.ch/data/dpg/index.php.

This work was supported in part by the Swiss National Science Foundation and the EU ERC grant MicroNano. Xavier Boix aknowledges the financial support of IURO (FP7-ICT-24314).

References

[1] D. Batra, R. Sukthankar, and T. Chen. Learning Class-Specific Affinities for Image Labelling. In *CVPR*, 2008. 2

[2] J. E. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Royal Stat. Soc., B*, 36(2), 1974. 1

[3] K. E. A. V. de Sande, T. Gevers, and C. G. M. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *PAMI*, 32(10):1582–1596, 2010. 5

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes

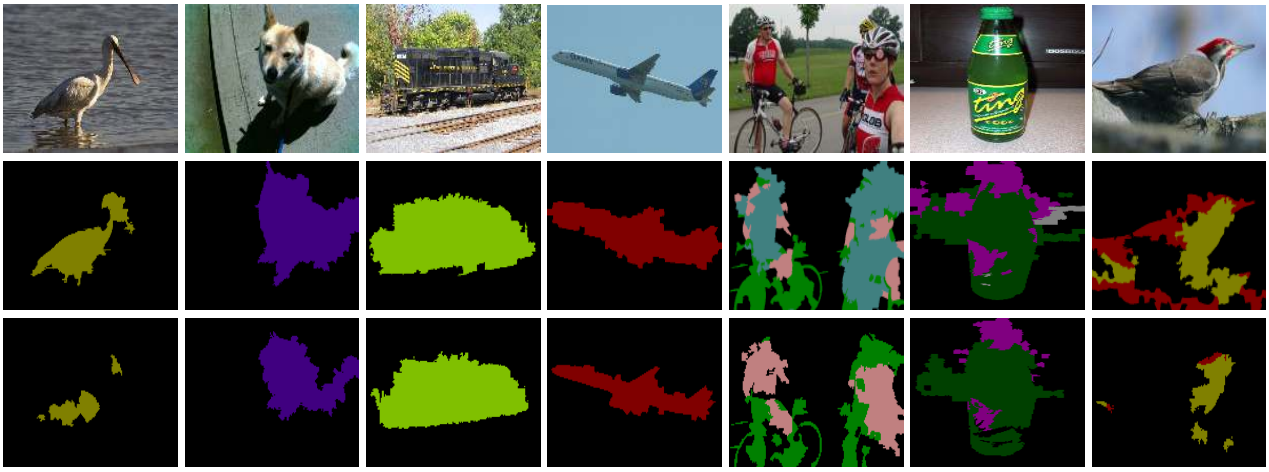


Figure 5. *Example segmentations from the PASCAL VOC 2010 test set.* (top) Original image. (middle) *DPG model* (bottom) *D-sampling*.

Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 1, 4

[5] T. Finley and T. Joachims. Training structural SVMs when exact inference is intractable. In *ICML*, 2008. 4

[6] B. Fulkerson, A. Vedaldi, and S. Soatto. Class Segmentation and Object Localization With Superpixel Neighborhoods. In *ICCV*, 2009. 4

[7] C. Galleguillos, A. Rabinovich, and S. Belongie. Object Categorization Using Co-Occurrence, Location and Appearance. In *CVPR*, 2008. 2

[8] J. Gonfaus, X. Boix, J. Weijer, A. Bagdanov, J. Serrat, and J. Gonzalez. Harmony Potentials for Joint Classification and Segmentation. In *CVPR*, pages 3280–87, 2010. 1, 2, 3, 4, 5, 6, 7

[9] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-Class Segmentation With Relative Location Prior. *IJCV*, 80(3):300–316, 2008. 2

[10] X. He, R. Zemel, and D. Ray. Learning and Incorporating Top-Down Cues in Image Segmentation. In *ECCV*, 2006. 2

[11] J. Jiang and Z. Tu. Efficient Scale Space Auto-Context for Image Segmentation and Labeling. In *CVPR*, 2009. 5, 6

[12] P. Kohli, M. Kumar, and P. Torr. P^3 and Beyond: Solving Energies With Higher Order Cliques. In *CVPR*, 2007. 3

[13] P. Kohli, L. Ladicky, and P. Torr. Robust Higher Order Potentials for Enforcing Label Consistency. In *CVPR*, 2008. 2

[14] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative Hierarchical CRFs for Object Class Image Segmentation. In *ICCV*, 2009. 2, 5, 6

[15] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001. 2

[16] D. Larlus and F. Jurie. Combining Appearance Models and Markov Random Fields for Category Level Object Segmentation. In *CVPR*, 2008. 2

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006. 5

[18] F. Li, J. Carreira, and C. Sminchisescu. Object Recognition as Ranking Holistic Figure-Ground Hypotheses. In *CVPR*, 2010. 7

[19] K. Murphy. Bayesian Map Learning in Dynamic Environments. In *NIPS*, pages 1015–1021, 1999. 3

[20] S. Nowozin, P. Gehler, and C. Lampert. On Parameter Learning in Crf-Based Approaches to Object Class Image Segmentation. In *ECCV*, 2010. 4

[21] N. Plath, M. Toussaint, and S. Nakajima. Multi-Class Image Segmentation Using Conditional Random Fields and Global Classification. In *ICML*, 2009. 2

[22] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in Context. In *ICCV*, 2007. 2

[23] A. Radhakrishna, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic Superpixels. Technical report, EPFL, June 2010. 4

[24] D. Ramanan. Learning to Parse Images of Articulated Bodies. In *NIPS*, 2006. 2

[25] J. Shotton, M. Johnson, and P. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation. In *CVPR*, 2008. 2, 5, 6

[26] 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), January 2009. 1, 3, 4, 5

[27] M. Szummer, P. Kohli, and D. Hoiem. Learning CRFs Using Graph Cuts. In *ECCV*, 2008. 4

[28] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004. 4

[29] J. Verbeek and B. Triggs. Scene Segmentation With Conditional Random Fields Learned from Partially Labeled Images. In *NIPS*, 2007. 2

[30] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local Features and Kernels for Classification of Texture and Object Categories: a Comprehensive Study. *IJCV*, 73(2):213–238, 2007. 4