Efficient Inference for Fully-Connected CRFs with Stationarity

Yimeng Zhang Tsuhan Chen
School of Electrical and Computer Engineering, Cornell University
{yz457,tsuhan}@cornell.edu

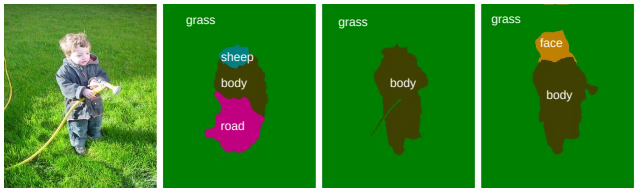
Abstract

The Conditional Random Field (CRF) is a popular tool for object-based image segmentation. CRFs used in practice typically have edges only between adjacent image pixels. To represent object relationship statistics beyond adjacent pixels, prior work either represents only weak spatial information using the segmented regions, or encodes only global object co-occurrences. In this paper, we propose a unified model that augments the pixel-wise CRFs to capture object spatial relationships. To this end, we use a fully connected CRF, which has an edge for each pair of pixels. The edge potentials are defined to capture the spatial information and preserve the object boundaries at the same time. Traditional inference methods, such as belief propagation and graph cuts, are impractical in such a case where billions of edges are defined. Under only one assumption that the spatial relationships among different objects only depend on their relative positions (spatially stationary), we develop an efficient inference algorithm that converges in a few seconds on a standard resolution image, where belief propagation takes more than one hour for a single iteration.

1. Introduction

Object-based image segmentation is one of the most challenging problems in computer vision. Given an image, the goal is to label every pixel with a predefined object category. A good algorithm for this task should be able to incorporate as much context as possible to ensure accurate object categorization, while providing precise segmentation along the object contours. Computational efficiency is also important, especially for high resolution images.

The most popular method is based on the conditional random fields (CRFs) defined over the image pixels, which was originally proposed in the TextonBoost work [27]. Unary potentials in the CRFs capture the low level cues with local texture, color, and location [10, 26, 27]. Edge potentials are typically defined for 4 neighbor pixels to smooth out the prediction. Although this 4-neighbor grid CRF has shown promising results for object segmentation, it fails to capture long-range context information.



(a) Image (b) Grid CRF (c) P^n+Co (d) Ours

Figure 1. Comparison of the object-based image segmentation results using our methods and previous works. (b) 4-neighbor grid CRF [27] (c) Robust P^n model plus object co-occurrence statistics [16] (d) Our method with a fully connected CRF, which captures both long-range color contrasts and the object spatial relationships in addition to their global co-occurrences. Our method preserves the object contours without pre-segmentation and is able to place the *face* at the right place even if its unary probabilities are small. The result with our method is obtained in less than 3 seconds.

Many techniques have been proposed incorporating longer-range information and more contexts by augmenting the traditional CRF with additional potentials, including global image features [31], top-down object detection results [1, 7, 17, 34, 35], label consistency in the same segmented regions [6, 9, 15, 21, 23, 25], and the global co-occurrence statistics among object categories [14, 16]. Efficient inference algorithms have also been associated with the augmented CRFs.

In this paper, we further incorporate the CRFs with pixel-wise spatial relationships among objects, in addition to their co-occurrences in the entire image. We model the object segmentation problem with a fully connected CRF, which allows every pair of pixels in an image to connect with each other. Unlike the grid CRFs, where the edges only serve for the contrast sensitive smoothness, the proposed CRF has edge potentials that encode both color contrasts and spatial arrangements of different object categories.

The context of spatial object interactions has been explored by many previous works on object detection [2, 29], and has been proved of its effectiveness. However, in terms of object segmentation, encoding this information at pixel level into the CRFs remains challenging due to the computational cost. Previous methods [6, 21, 31] first segment the image and then model the object relationships over the

regions. Working on the regions largely reduces the computational complexity and makes it tractable to capture the semantic interactions between every pair of regions. However, the main disadvantage of the segment based methods is the errors introduced by the initial segmented regions. As discussed in [10, 15], this may result in dramatic errors in the final object labeling. Multiple segmentations [25] ease the problem, but increase the computation and may still leave some errors. As shown in [10, 14, 15], directly augmenting the pixel-wise CRFs usually works better. Moreover, the irregular segmented regions usually capture only weak geometry, such as the co-occurrence in the entire image [23, 31], four binary relationship indicators ("inside", "outside", "above", "below") [6, 14], or locations defined over the 3×3 grid of an image [21]. We are able to capture more detailed spatial information with the pixels and preserve object contours at the same time.

A fully connected CRF over image pixels has N^2 number of edges, where N is the number of pixels in an image. For an image of a resolution of 213×320, the number of edges is more than 5×10^9. The current inference methods [11, 12, 24, 33, 36] are impractical in such a case. We propose an efficient inference algorithm for the fully connected CRF, which reduces the complexity at every iteration to $O(N \log N)$. The proposed algorithm relies on one constraint on the edge potentials: Given two object categories, the spatial potentials over two pixels depend only on their relative positions. That is, we assume that the likelihood that two categories co-occur at two particular pixels is only determined with the offset of these pixels. This stationary assumption is a standard assumption, and was made by almost all previous context modeling works [2, 6, 14, 21, 29]. Under this assumption, we show that the gradient of the Quadratic Programming (QP) relaxation of the fully connected CRF can be efficiently computed with the help of convolution. QP relaxation was proposed recently in [24] as another inference method for CRFs and has been proven to be a tight relaxation and solve exactly the MAP problem of the original CRFs. With the proposed algorithm, the QP optimization converges in a few seconds, while max-product belief propagation and the original QP relaxation take more than one hour even for a single iteration on the fully connected CRF.

Efficient inference for fully connected CRFs has also been explored by a simultaneous and independent work of Krahenbuhl and Koltun [13]. The main difference is that they make the Gaussian assumption of the *stationary* edge weights and use a Potts model with only label consistency for similar colors. We relax the assumption to any distribution for the stationary weights. With this more general assumption, we can encode more general statistics in the edge weights. For example, we can model the relative spatial relationship (not only "distances) among different categories, such as the geometry relationship between cow and grass. We also provide a different inference algorithm for this more general problem.

The main contributions are as follows: 1) we present a unified model that augments the traditional CRFs to capture the object spatial relationships with fully connected edges. The proposed CRF also captures the long-range color contrast, and therefore preserves the object boundaries without pre-segmenting the image; 2) we propose an efficient inference algorithm for fully connected CRFs with only one stationary constraint and show that convergence can be achieved in only a few seconds for a CRF with 5×10^9 edges. Fig. 1 illustrates the benefit of using our method.

1.1. Related Work

Context: Different contexts have been explored by many previous works in order to improve traditional CRFs. *Global image features* [19, 31] add global image information by transferring whole image labels to a test image from similar training images. *Label consistency in a segment* is enforced by first segmenting the image and then performing labeling on the segments [6, 10, 21, 25]. *Top-down object detection* results are added to improve the recognition performance for structured objects [1, 7, 17, 34, 35]. *Object relationships* have been explored by previous works by modeling weak geometries over segmented regions [6, 14, 21, 23, 31]. Torralba et al. [30] captures pixel-wise relationships for the object detection task, where the goal is to obtain a rough location without precise object contour segmentations. We capture the sematic spatial interactions using the pixels and preserve the object contours at the same time; thus we have a different and harder inference problem.

Long Range CRFs: Long-range interactions in CRFs have been explored before. Sparse CRFs [18] add long-range edges at sparse locations to ensure the computational tractability. The decision tree field [20] and auto context [32] learn the edge potentials and determine the edges to be added based on the learning results. Hierarchical CRFs [14] capture long-range interactions with a higher layer, and therefore may lose detailed information over the pixels. The co-occurrence CRF [16] encodes global object co-occurrence statistics.

Fast Inference with Convolution: Formulating the original problem with convolution to accelerate the computation has a long history in vision. In particular, Felzenszwalb and Huttenlocher [4, 5] formulate the message passing between two nodes in belief propagation as a "min-convolution", which reduces the time from $O(K^2)$ to $O(K)$, where K is the number of categories. The assumption they rely on is that the edge costs depend only on the numerical differences of the labels. Although this assumption works well for tasks like depth estimation, it does not hold for object segmentation where the categories do not

have any numerical meaning. More importantly, we focus on inference efficiency over the space domain, which is defined by the number of pixels N, while they are interested in the number of categories K. For a fully connected graph, the method still needs to pass N^2 number of messages.

2. Approach

We first give a brief description of the traditional 4-neighbor grid CRFs for object-based segmentation, and then describe how we augment it with spatial information and the efficient inference algorithm.

2.1. Grid-CRFs for Object Segmentation

The conditional random fields (CRFs) model the conditional distribution of the class labeling $\mathbf{X}$ given an image $\mathbf{I}$. We use $\mathbf{X} = \{X_1, X_2, ..., X_N\}$ to denote the set of discrete random variables with $X_i \in X$ being associated with every pixel $i \in V$, and takeing a value from the label set $L = \{l_1, ..., l_K\}$, which are the object categories. The labeling X of the image is obtained with a maximum a posteriori (MAP) estimation of the following conditional log-likelihood:

$$\log P(\mathbf{X}|\mathbf{I}) = \sum_{i \in V} \psi_i(x_i) + \sum_{i \in V, j \in N_i} \psi_{ij}(x_i, x_j) - Z(\mathbf{I}), \tag{1}$$

where N_i is the 4-neighbors of pixel i, and Z is a normalization term that does not depend on $\mathbf{X}$. The unary potentials $\psi_i(x_i)$ are defined with the log likelihood of variable X_i taking label x_i, which are usually computed with the texture, color, and location features extracted from a local region around i [10,27]. The edge potential $\psi_{ij}(x_i, x_j)$ typically encodes contrast sensitive smoothness of neighboring pixels [27].

2.2. Fully Connected CRFs with Stationary Edges

We formulate the labeling problem with a fully connected CRF defined over the image pixels. The conditional log-likelihood is the same as the grid CRFs (Equ. 1), except that the edges are defined over all pairs of pixels. In other words, the neighborhood of a pixel i is defined with all other pixels $N_i = V$. The unary potential is the same as grid CRFs. The main difference is that the edge potential $\psi_{ij}(x_i, x_j)$ is a combination of the color contrast $\varphi_{ij}(x_i, x_j)$ and the spatial relationships between two categories, $\phi_{ij}(x_i, x_j)$.

$$\psi_{ij}(x_i, x_j) = \underbrace{\phi_{ij}(x_i, x_j)}_{\text{spatial relation}} \underbrace{\varphi_{ij}(x_i, x_j)}_{\text{color contrast}}. \tag{2}$$

Color term: The color contrast term $\varphi_{ij}(x_i, x_j)$ encourages the same label when the colors I_i, I_j are similar, and different labels otherwise. Let $g(I_i - I_j)$ denote the gaussian function of the color difference: $g(I_i - I_j) = \exp(-\theta_c ||I_i - I_j||^2)$. The color contrast term is defined as follows.

$$\varphi_{ij}(x_i, x_j) = \begin{cases} g(I_i - I_j) & \text{if } x_i = x_j \\ 1 - g(I_i - I_j) & \text{otherwise .} \end{cases} \tag{3}$$

Spatial term: The spatial term $\phi_{ij}(x_i, x_j)$ is the log-likelihood of the spatial distribution $f(x_i, x_j, p_i, p_j)$ of these two categories, i.e. the probability that two categories x_i, x_j co-occur at positions p_i, p_j. We make the assumption that this probability only depends on the relative positions of the two pixels $p_i - p_j$.

$$\phi_{ij}(x_i, x_j) = \log(\varepsilon_s + f(x_i, x_j, p_i - p_j)) \tag{4}$$

$$f(x_i, x_j, p_i - p_j) = \frac{1}{Z} P(x_i, x_j) P(p_i - p_j | x_i, x_j), \tag{5}$$

where Z is a normalization term, and $f(x_i, x_j, p_i - p_j)$ is computed as a combination of the global co-occurrence probability $P(x_i, x_j)$ and the relative position distribution given the two categories $P(p_i - p_j | x_i, x_j)$. Note that when $x_i = x_j$, the spatial term computes the likelihood that the same category occurs at a relative position. This likelihood can capture the shape and size information of the objects.

2.3. Quadratic Programming Relaxation

To obtain a MAP estimation of a CRF model, many inference methods have been proposed. In this paper, we adopt the quadratic programming (QP) relaxation method [24], which has been shown to perform comparable to other inference methods, such as LP relaxation [12] and tree-reweighted message passing [11].

The labeling problem that maximizes the conditional probability can be viewed as an integer program by adding a binary variable $\mu_i(x_i)$, which indicates whether a pixel i is labeled with x_i.

$$\mu_i(x_i) = \begin{cases} 1 & \text{if } X_i = x_i \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

With $\mu_i(x_i)$, the MAP estimation for the conditional probability $P(X|I)$ (Equ. 1) can be formulated as a quadratic integer program, which is further relaxed to the following quadratic program.

$$\begin{aligned} \max_{\mu} \; & \sum_{i; x_i} \psi_i(x_i) \mu_i(x_i) + \sum_{i,j; x_i, x_j} \psi_{ij}(x_i, x_j) \mu_i(x_i) \mu_j(x_j) \\ \text{s.t.} \; & \sum_i \mu_i(x_i) = 1 \\ & 0 \le \mu_i(x_i) \le 1 \end{aligned} \tag{7}$$

This QP relaxation has been proven to be a tight relaxation, and solves exactly the original MAP problem [24].

Although the optimization problem can also be formulated as a minimization problem of the Gibbs energy, which

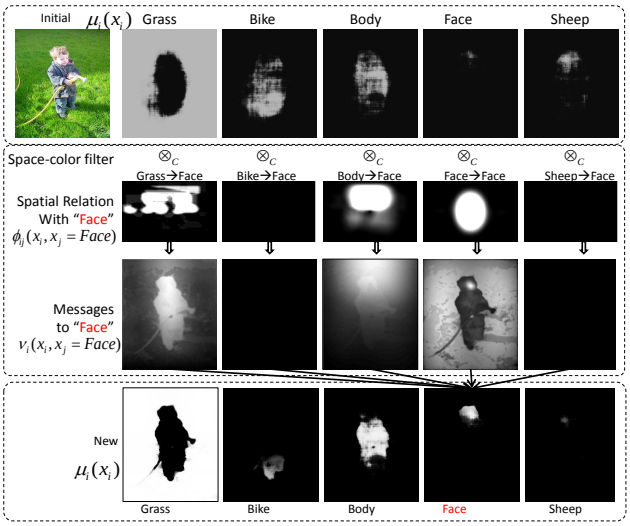


Figure 2. The illustration of the inference algorithm. The top row shows the initial $\mu_i(x_i)$ obtained from the unary terms. The second row shows the spatial relationships of different categories with "Face" (to save space, we show them in smaller images). The third row illustrates the messages sent to "Face" $\nu_i(x_i, x_j)$. The bottom row shows the updated $\mu_i(x_i)$ after normalization.

is the negative of the above objective function, the QP relaxation is usually formulated with maximization [24]. For convenience, we make the coefficients in the objective function positive by adding a constant to the log likelihood of the unary probability $\psi_i(x_i)$ and the spatial distribution $f(x_i, x_j, p_i - p_j)$ (Equ. 5).

2.4. Iterative Update Procedure

We initialize μ with the unary probabilities, and iteratively update it with the gradient of the objective function. The gradient of the objective function Q of Equ. 7 is as follows.

$$q_i(x_i) = \frac{\partial Q}{\partial \mu_i(x_i)} \tag{8}$$

$$= \psi_i(x_i) + 2 \sum_{x_j} \sum_{j} \psi_{ij}(x_i, x_j) \mu_j(x_j) \tag{9}$$

$$= \psi_i(x_i) + 2 \sum_{x_j} \nu_i(x_i, x_j), \tag{10}$$

with the symmetric edge potentials $\psi_{ij}(x_i, x_j)$.

Given the gradient $q_i(x_i)$, we can adopt the fixed point iteration to perform gradient ascent and maintain the constraints in (7) at the same time.

$$\mu_i^{t+1}(x_i) = \frac{\mu_i^t(x_i) q_i(x_i)}{\sum_{x_i} \mu_i^t(x_i) q_i(x_i)}, \tag{11}$$

where $\mu_i^t(x_i)$ is the value of $\mu_i(x_i)$ at the t^{th} iteration.

When the edge potentials do not define a negative definite matrix, gradient ascent may converge to a local maximum, as other iterative update methods, such as max-product belief propagation or mean field. While we can follow [24] to change the values on the diagonal of the matrix to make a convex approximation, we found the original edge potentials produce a reasonable result in practice.

2.5. Gradient as Image Filtering

The main bottleneck for computing the gradient (Equ. 10) of a fully connected CRF is the computation for $\nu_i(x_i, x_j)$, which is the weighted summation of the messages from all other pixels to i when the categories are x_i, x_j. A naïve implementation would have computational complexity $O(N^2)$ for computing this term for all pixels, where N is the number of pixels. Combining with the edge potential definitions (Equ. 2, 3, and 5), for two categories (x_i, x_j), $\nu_i(x_i, x_j)$ is calculated as follows. For simplicity, we omit x_i, x_j in the variables.

$$\begin{aligned} \nu_i &= \sum_j \underbrace{\phi_{ij}(x_i, x_j)}_{\text{spatial}} \underbrace{\varphi_{ij}(x_i, x_j)}_{\text{color}} \mu_j \\ &= \begin{cases} \sum_j \phi(p_i - p_j) g(I_i - I_j) \mu_j & \text{if } x_i = x_j \\ \sum_j \phi(p_i - p_j) \left(1 - g(I_i - I_j)\right) \mu_j & \text{otherwise .} \end{cases} \end{aligned} \tag{12}$$

The intuition of this formulation is that for the same category, we prefer propagating the messages to similar color pixels, while for different categories, we propagate the messages to pixels of different colors, which are more likely to come from different objects (illustrated Fig. 2).

Since $\phi(p_i - p_j)$ only depends on the relative position $p_i - p_j$, if we do not have the color term, the above equation can be solved for all pixels i by treating $\phi(p_i - p_j)$ as a filter for an image valued with $\mu_i(x_i)$. Image filtering can be greatly accelerated with Fast Fourier Transform (FFT), which reduces the complexity to $O(N \log N)$.

However, if we have the color term, the equation does not take the form of convolution. Instead, it is a filtering on the space and color dimension at the same, where the color filter is a Gaussian. Image filtering on both space and color dimensions has been studied before, as the edge preserving image smoothing problem, which is also called Bilateral Filtering [3, 22, 28]. In bilateral filtering, the spatial filter is also a Gaussian. Although we have a different space term, we can borrow the idea from the algorithms proposed for this problem.

Borrowing the idea from [3], if we fix the color value I_i for the pixel i, Equ. 12 will become a convolution of the function $H_j^c = g(I_c - I_j)\mu_j$. Therefore, we discretize the color values into C clusters $\{I^c\}$, and compute a linear filtering for each I_c.

$$\nu_i^c = \textstyle\sum_j \phi(p_i - p_j) g(I_c - I_j) \mu_j \tag{13}$$

$$= \textstyle\sum_j \phi(p_i - p_j) H_j^c, \tag{14}$$

when $x_i = x_j$. The final output ν_i is a linear interpolation between the output ν_i^c of two closest values I_c of I_i.

The resulting computation complexity is $O(CN \log N)$, where C is the number of clusters we create for an image. In practice, we found $10 - 15$ clusters produce good results on most images. We also adopt the down-sampling scheme as described in [3] to further speed up the computation (detailed in the next section). On a 213×320 resolution image, computing the gradient (Equ. 10) for all pixels takes less than 0.1 seconds. Although further acceleration can be expected with more recent algorithms on the bilateral filtering problem [22], we leave deep exploration on this line for future work.

The illustration of the iterative update is shown in Fig. 2

2.6. Learning

We perform piecewise training as previous works [10,16, 27], which learn each potentials with ground truth instead of a joint learning of all potentials at the same time. For the unary term, we employ the same parameters as [10].

For the edge potential, we need to learn the co-occurrence distribution $P(l, l')$ of two categories l, l' and their relative spatial distribution $P(dp_x, dp_y|l, l')$; we use (dp_x, dp_y) to denote the position offset on x, y axes. We perform a maximum likelihood estimation for these distributions using the training images. The co-occurrence distribution can be easily obtained by counting the number of times categories l, l' appear in the same image. To compute $P(dp_x, dp_y|l, l')$, for each image which has objects l, l', we count the frequency l and l' occur at relative positions (dp_x, dp_y). This can be efficiently computed with a cross-correlation over the pixel-wise ground truth for categories l and l' (Fig. 3). Note that the order of l and l' matters. Learning the distributions on 276 images for each pair of 21 categories takes less than 25 seconds.

To avoid over-fitting, we compute a quantized spatial distribution. Specifically, rather than computing pixel-wise $P(dp_x, dp_y|l, l')$, we compute a binned relative spatial distribution with $M_x \times M_y$ bins. In practice, we use step size 5 for both x and y axes. The quantization also enables us to perform image filtering (previous subsection) using the down-sampling scheme as [3] with little quality loss. Specifically, we do filtering on the down-sampled image space, but perform the final interpolation with full-scale images. Note that this is quite different from performing inference on a resized input image, which will lose detailed information. We refer the readers to [3] for the detailed algorithm. With the down-sampling scheme, the computational complexity for computing the gradients reduces to $O(CM_xM_yl \log(M_xM_y))$, where C is the number of color clusters created for an image.

Other model parameters, such as different weights given to unary and edge potentials, are manually set to minimize the segmentation errors on a validation set.

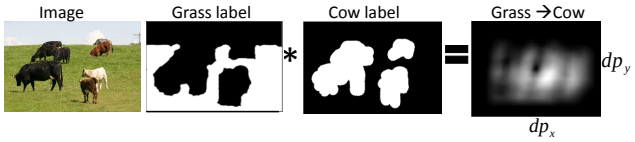


Figure 3. The illustration for computing the frequency that the pixels labeled with "grass" and "cow" appear at different relative positions (dp_x, dp_y). The rightmost image shows the resulting (dp_x, dp_y) space, which can be obtained with a cross-correlation from the ground truth of the two categories.

2.7. Practical Issues

At each iteration, we need to do the image filtering (section 2.5) for every pair of categories. We can first compute the Fourier transform for the scores of all categories to avoid K^2 times of FFT computation, where K is the number of categories. Moreover, when the co-occurrence probability of two categories is very low, we do not make propagation between them. Finally, we pre-compute the Fourier transforms for the edge potentials to avoid recomputing them at every iteration.

3. Experiments

We perform the experiments on the Sowerby-7 [8] and MSER-23 [27] datasets. The Sowerby dataset contains 7 object classes of 104 images, and the MSRC dataset has 23 object classes of 591 images. We compare favorably with the state-of-the-art object segmentation approaches, including 4-neighbor grid CRFs [27], robust P^n CRFs [10], and the robust P^n plus object co-occurrence CRFs [16]. For implementation of these methods, we use the publicly available code provided by the authors. We consider our baseline as different inference methods where the same low level features are used, and treat other works that encode additional features, such as features from the whole image [19, 31] or image segments [15], as complementary to ours.

3.1. Synthetic Data

We first perform an experiment on the synthetic data to evaluate the proposed inference algorithm. In this experiment, we would like to know how well the proposed fully connected CRF work when we do not have reliable unary potentials but know the exact spatial object relationships. For a selected 213x320 resolution image from the MSRC dataset, we learn relative spatial distribution between different objects using the ground truth of the same image. For the unary potentials, we randomly generate the class probabilities for each pixel from a uniform distribution over all categories included in the image.

Analysis: The inference results of the synthetic data are shown in Fig. 4. No surprisingly, using the 4-neighbor grid CRF, where no unary and spatial cues are available, we can-

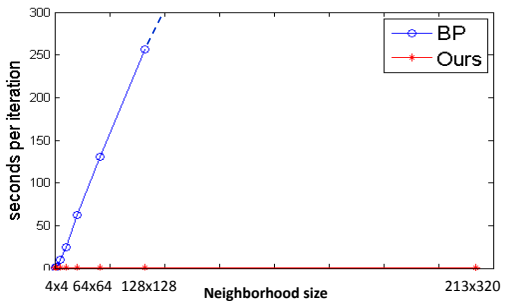


Figure 5. Average running time (seconds) per iteration for max-product belief propagation and the proposed method for a CRF defined over 213x320 pixels with edges over different neighborhood size.

not locate the objects correctly. With relative spatial distribution, we can predict a reasonable layout of the objects, but do not preserve good object boundaries when color contrasts are not included. With color contrast in the edge potentials, we can provide reliable object segments. The QP optimization converges after around 20-30 iterations in both cases. In all the following experiments, we always set the maximum iteration to 30.

Running time: We compare the running of our method with max-product loopy belief propagation (BP) in Fig. 5. We implemented the proposed method using Matlab with some C++ help on a server with Quad-Core 2.66G Intel Xeon CPU and 12G memory. We vary the number of neighborhood pixels each node connects to in the CRF. The running time of BP for each iteration is linear to neighborhood size, while the proposed inference algorithm is almost constant. The reason is that the main computation of the proposed algorithm is a FFT over the image, which is $O(N \log N)$ regardless of the neighborhood size (N is the total number of pixels in the image). When all pixel pairs are connected, BP runs out of memory on our computer, while the estimated running time would be more than one hour even if enough memory is provided. In comparison, our algorithm is less than 0.1 second.

3.2. Segmentation without Unary Cues

Now we ask ourselves the question: in real cases, if we know what objects are present in an image, how well can we perform object localization and segmentation with only the relative positions of different categories? That is, we do not use any classifiers of the low level cues, such as texture or color, which can be computationally expensive for both training and testing. Instead, we only learn the relative spatial distributions between two categories. We experiment on the Sowerby dataset [8], where we can safely assume all images have the same categories: *sky, road, vegetable*, and *building*. We ignore the *car* and *sign* categories which are only present in a few images. We randomly select 50% images to train the spatial distribution, and use the rest 50% for testing.

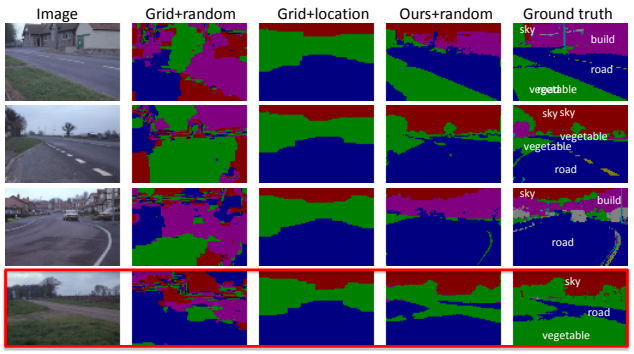


Figure 6. Qualitative results on the Sowerby dataset when only absolute or relative locations are learned during training. From left to right, we show the test image, grid CRF with random unary potential, grid CRF with learned location potentials, proposed method with random unary potential, and the ground truth.

	Grid+random	Grid+loc	P^n+loc	Ours+random	Ours+loc
Global	22.2	68.2	62.1	76.4	**78.9**
Avg. Recall	24.0	54.5	63.0	67.8	**70.8**
Avg. Precision	21.8	49.4	61.7	67.7	**70.7**

Table 1. Quantitative results on the Sowerby dataset when only absolute or relative locations are learned during training. We show the pixel-wise global accuracy (%) over all images, and recall (%) and precision (%) averaged over different categories.

Qualitative analysis: Fig. 6 illustrates the predicted segmentation results of example testing images using different methods. We first compare our method with 4-neighbor grid CRF when randomly generated unary potentials are used. Since the grid CRF only has neighboring color contrast information, the predicted labeling is quite random. To make fair comparison, we learn absolute location distribution for different classes and encode it as the unary potential. With the location information, the grid CRF performs better, but provides similar prediction for all images. On the contrary, the proposed fully connected CRF can make reasonable object arrangements and preserve object boundaries for different test images. We can explain our method as follows: the spatial relationship term in the edge potentials predicts a rough object layout, and the color contrast term among all pixel pairs enforces a better object segmentation. The bottom row of Fig. 6 shows a typical error made by our algorithm, when the input image has an usual spatial layout among objects.

Quantitative comparison: Table 1 presents the pixel-wise accuracy using different methods. In addition to the grid CRF, we also compare with the robust P^n CRF [10], which encourages the label consistency within each segmented region.

3.3. Segmentation with Unary Cues

Finally, we show the experiment results on the MSRC-23 dataset when more unary information is used. We use the same experiment settings as [27]: 45% of the images

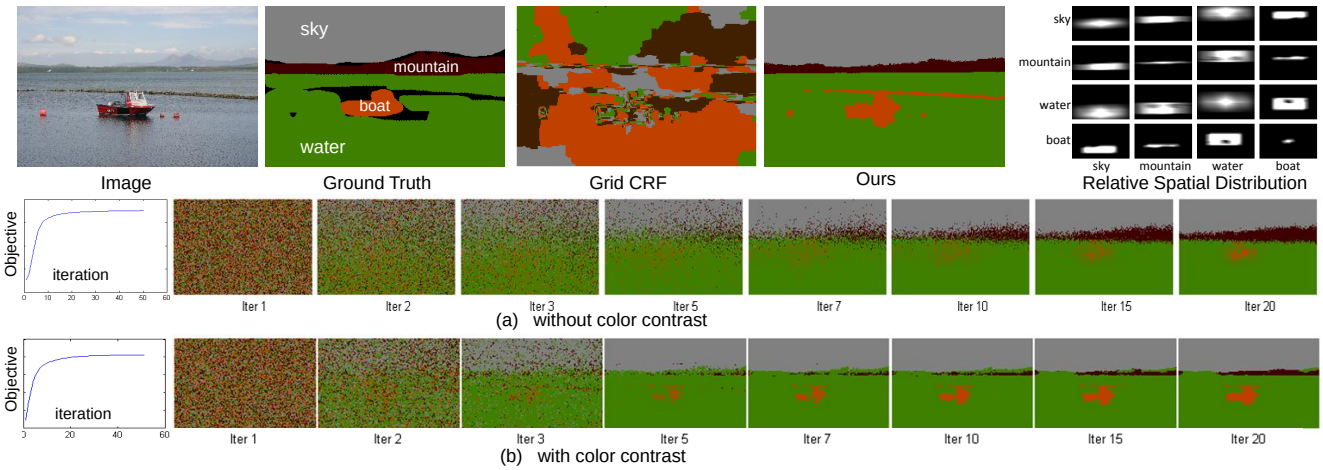


Figure 4. Analysis on the synthetic data where no unary cues are used. The rightmost image on the top row shows the relative spatial distribution of each pair of categories. The middle and bottom rows show the performance of our method with and without color contrast term in the edge potentials respectively. The leftmost images of these two rows plot the changes of the objective values in the QP relaxation at each iteration. The rest images show the changes of the predicted labels as we perform iterative updates.

for training, 10% for validation, and the rest 45% for testing. Following [27], we ignore the *mountain* and *horse* categories which have too few training images. For the unary potentials, we use the same texture, color, and location classifiers as [10] by applying their published code.

Qualitative analysis: Fig. 7 illustrates the labeling results of example test images using different CRFs. Grid CRF does not capture class relationships, and prefers smoother object boundaries. Incorporating object co-occurrence statistics [16] in the Robust P^n CRF [10] provides better object boundaries and remove the objects that rarely co-occur in the same image, eg. *cat* and *book* (first row). The proposed CRF incorporates both object co-occurrence information and their spatial arrangements, and produce better class predictions. For example, *road* is more likely to appear below *grass* than *water* (first row); *boat* has higher possibility to appear right above *water* than *building* (third row); and *car* rarely appears right above *grass* (bottom row). Our approach benefits from modeling everything in a unified CRF, and thus is able to make the object layout adjustment only when the unary potentials are more confused. The second row in the figure is an example where the color contrast and spatial relation together leads to a better prediction. Since *sky* is very likely to appear above *grass*, and the colors are quite different for a particular region, we can make the correct changes. Moreover, with the long-range color contrast, we can also provide more precise object boundaries, eg. *bird* in the first row, and *tree* in the bottom row.

Quantitative comparison: Table 2 compares different CRFs with the recognition rates of different categories. Our method improves previous works on most of the categories by 1% to 14%. Note that the ground truth provided by the MSRC dataset is quite rough, and therefore more precise object boundaries may not reflect better accuracies. Table 3 presents the global accuracies over all images, and precision and recall averaged over different categories. Robust P^n+co-occurrence [16] performs better than Robust P^n by incorporating global co-occurrence statistics. By further encoding spatial relationships, we improve the Robust P^n+co-occurrence by 1.9% in global accuracy. Since the global accuracy is biased for categories that have more pixels, such as *grass* or *tree*, more improvement is observed for average precision (3.6%) and recall (2.7%). The proposed method also runs faster than Robust P^n+co-occurrence. Excluding the time for obtaining the unary potentials, the proposed method process an image in 2 to 8 seconds, while robust P^n+co-occurrence [16] requires 8 to 30 seconds per image. Gaussian CRF [13] also captures long-range color contrast. By incorporating spatial interactions among different object categories, we improve the method by 1.0% in global accuracy and 2.4% in average recall.

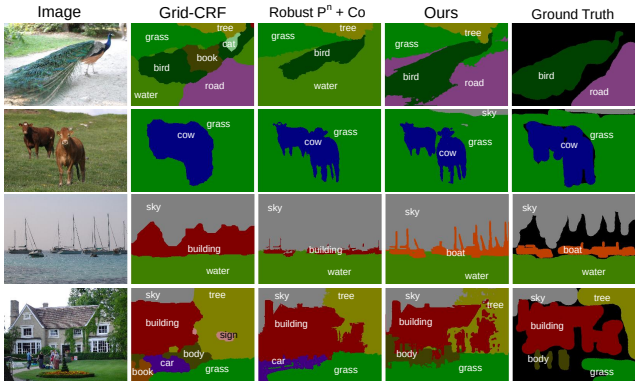


Figure 7. Results on the MSRC dataset with different CRFs

Errors: Fig. 8 shows example errors made by our method. When we do not have a good support from the unary potential, we predict the labels with more common spatial arrangements, which can be wrong in some cases.

	building	grass	tree	cow	sheep	sky	plane	water	face	car	bike	flower	sign	bird	book	chair	road	cat	dog	body	boat
Unary	68.8	96.5	**90.1**	82.9	85.0	94.2	**87.9**	80.4	88.7	77.0	92.5	86.1	62.1	41.5	93.8	66.5	86.3	81.0	53.9	75.0	30.6
Grid-CRF	69.0	96.8	88.6	82.9	85.9	94.6	87.4	81.8	88.2	78.0	92.9	87.0	63.5	41.6	94.1	66.3	86.0	81.3	54.2	75.0	30.2
Robust P^n [10]	70.0	96.8	89.7	83.9	86.9	94.6	87.4	81.8	89.1	78.0	92.7	88.0	62.5	41.2	94.1	66.3	87.0	82.3	**54.6**	75.7	29.2
Robust P^n + Co [16]	71.5	97.0	89.8	84.7	**87.7**	94.8	87.4	81.9	88.9	78.9	93.6	89.0	62.6	42.1	94.6	66.8	87.5	82.5	51.9	76.5	28.8
Ours	**75.4**	**98.1**	88.7	**86.1**	86.2	**96.3**	87.5	**85.0**	**93.9**	**79.1**	**93.7**	**90.0**	**64.2**	**55.9**	**94.8**	**81.6**	**88.2**	**85.4**	53.6	**79.1**	**32.1**

Table 2. Recognition rates (%) of different categories in the MSRC dataset. Recognition rate is computed as the recall of each category.

	Unary	Grid	P^n [10]	P^n + Co [16]	Gaussian [13]	Ours
Global	84.1	84.6	84.7	85.1	86.0	**87.0**
Avg. Precision	79.3	80.5	80.6	81.9	n/a	**85.5**
Avg. Recall	77.1	77.4	77.7	78.0	78.3	**80.7**

Table 3. Performance comparison on the MSRC dataset with pixel-wise global accuracies (%) over all images, precision (%) and recall (%) averaged over different categories. Except the Gaussian CRF [13], we show the performance using the code by [16]. Therefore, exactly the same unary potentials are used. For [13], we cite the performance in their paper.

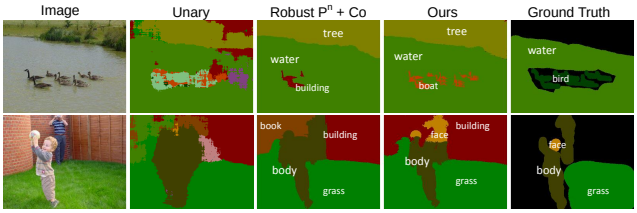


Figure 8. Example errors made by our approach on the MSRC dataset.

4. Conclusion

We proposed a fully connected CRF which encodes spatial relationships among different objects and preserves object contours at the same time. We also developed an efficient algorithm to inference the fully connected CRF. The experiment results demonstrated the efficiency and effectiveness of the proposed CRF. For future work, we would like to investigate more sophisticated learning of the model parameters.

Acknowledgement: The authors thank Dr. Dhruv Batra for the suggestions and discussions on the idea and writing of the paper.

References

[1] L. Bertelli, T. Yu, D. Vu, and B. Gokturk. Kernelized structural SVM learning for supervised object segmentation. In *CVPR*, 2011. 1, 2

[2] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 95(1):1–12, 2011. 1, 2

[3] F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *SIGGRAPH*, 2002. 4, 5

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 2

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 70(1):41–54, 2006. 2

[6] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008. 1, 2

[7] J. M. Gonfaus, X. B. Bosch, J. van de Weijer, A. D. Bagdanov, J. S. Gual, and J. G. Sabaté. Harmony potentials for joint classification and segmentation. In *CVPR*, 2010. 1, 2

[8] X. M. He, R. S. Zemel, and D. Ray. Learning and incorporating top-down cues in image segmentation. In *ECCV*, pages I: 338–351, 2006. 5, 6

[9] A. Ion, J. Carreira, and C. Sminchisescu. Probabilistic joint image segmentation and labeling. In *NIPS*, 2011. 1

[10] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008. 1, 2, 3, 5, 6, 7, 8

[11] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 28(10):1568–1583, 2006. 2, 3

[12] N. Komodakis and N. Paragios. Beyond loose lp-relaxations: Optimizing mrfs by repairing cycles. In *ECCV*, 2008. 2, 3

[13] P. Krahenbuhl and V. Koltun. Efficent inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 2, 7, 8

[14] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005. 1, 2

[15] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009. 1, 2, 5

[16] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010. 1, 2, 5, 7, 8

[17] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, 2010. 1, 2

[18] Y. Li and D. P. Huttenlocher. Sparse long-range random field and its application to image denoising. In *ECCV*, 2008. 2

[19] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009. 2, 5

[20] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli. Decision tree fields. In *ICCV*, 2011. 2

[21] D. Parikh, C. L. Zitnick, and T. Chen. From appearance to context-based recognition: Dense labeling in small images. In *CVPR*, 2008. 1, 2

[22] S. Paris and F. Durand. A fast approximation of the bilateral filter using a signal processing approach. *IJCV*, 81(1):24–52, 2009. 4, 5

[23] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 1, 2

[24] P. Ravikumar and J. Lafferty. Quadratic programming relaxations for metric labeling and markov random field map estimation. In *ICML*. ACM Press, 2006. 2, 3, 4

[25] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 1, 2

[26] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 1

[27] J. Shotton, J. M. 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. 1, 3, 5, 6, 7

[28] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, 1998. 4

[29] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):169–191, 2003. 1, 2

[30] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*, 2004. 2

[31] T. Toyoda and O. Hasegawa. Random field model for integration of local information and global information. *PAMI*, 30(8):1483–1489, Aug. 2008. 1, 2, 5

[32] Z. W. Tu. Auto-context and its application to high-level vision tasks. In *CVPR*, 2008. 2

[33] Y. Weiss and W. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *Information Theory*, 47(2):736 –744, 2001. 2

[34] Y. Weiss and A. Levin. Learning to combine bottom-up and top-down segmentation. In *ECCV*, 2006. 1, 2

[35] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object detection for multi-class segmentation. In *CVPR*, 2010. 1, 2

[36] R. Zabih, O. Veksler, and Y. Y. Boykov. Fast approximate energy minimization via graph cuts. In *ICCV*, 1999. 2