Scene Understanding with Discriminative Structured Prediction

Jinhui Yuan Jianmin Li Bo Zhang
State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University
yuan-jh03@mails.tsinghua.edu.cn {lijianmin,dcszb}@tsinghua.edu.cn

Abstract

Spatial priors play crucial roles in many high-level vision tasks, e.g. *scene understanding. Usually, learning spatial priors relies on training a structured output model. In this paper, two special cases of discriminative structured output model,* i.e. *Conditional Random Fields (CRFs) and Max-margin Markov Networks (M^3N), are demonstrated to perform image scene understanding. The two models are empirically compared in a fair manner,* i.e. *using the common feature representation and the same optimization algorithm. Particularly, we adopt online Exponentiated Gradient (EG) algorithm to solve the convex duals of both models. We describe the general procedure of EG algorithm and present a two-stage training procedure to overcome the degeneration of EG when exact inference is intractable. Experiments on a large scale image region annotation task are carried out. The results show that both models yield encouraging results but CRFs slightly outperforms M^3N.*

1. Introduction

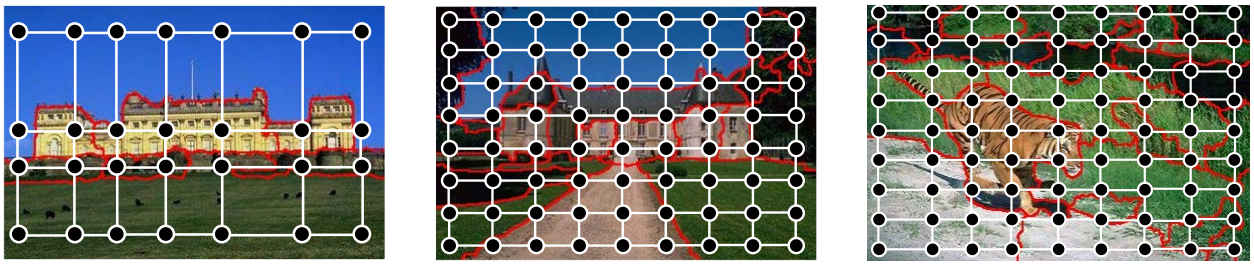
Prior knowledge on the geometrical configuration or spatial dependencies among objects play crucial roles in high-level computer vision tasks, such as object detection [25], object recognition and scene understanding [4, 6, 8, 11, 15, 22, 23, 31]. The basic idea is, to recognize an object, the algorithm should not only consider the local appearance, but also take the spatial context into account. Markov Random Fields (MRFs) has been considered a natural model for exploiting such spatial priors [19]. However, it is trained in generative way. Recent advances in discriminative training technique show prominent advantages over generative ways. For example, Conditional Random Fields (CRFs) [16], relaxing the independence assumption by being conditionally trained, brings significant improvement to generative trained MRFs [12, 15, 21, 20, 22, 31]. Another state-of-the-art method, Max-Margin Markov Networks (M^3N) incorporates the large margin mechanisms into MRFs, making it very appealing [1, 24, 26].

CRFs has been broadly utilized in vision tasks [12, 15, 20, 22, 31], but M^3N has seldom been explored [3]. In particular, little has been done to empirically compare the two discriminative training techniques in computer vision field. We may naturally ask, what about the empirical performance of M^3N on vision problems? Theoretically, CRFs and M^3N differ only in their loss functions [1, 7]. Both methods can be unified in a framework of structured output linear discriminant function [7]. This perspective allows us to present an empirical comparison in a fair manner, *i.e.*, with common feature representation and the same optimization algorithm.

Therefore, the objective of this paper is two-fold. First, we are interested in the empirical comparison of CRFs and M^3N. Second, we also want to demonstrate how the discriminative structured prediction approach can be applied in vision problems, particularly for scene understanding. In the paper, we briefly introduce CRFs and M^3N and present detailed description to other topics which are rare in vision literatures. Particularly, we describe how to map the structured input pattern to feature space and how to learn the parameters within the models. We adopt online Exponentiated Gradient (EG) algorithms to solve the convex duals of both models. Though EG algorithm will converge when exact inference is possible, it will sometimes fail for approximate inference in the graphs with cycles. We design a two-stage EG training strategy to address this problem. Experiments show that discriminative structured prediction are promising approaches for scene understanding.

The paper is structured as follows. In Section 2 we describe the problem setting of scene understanding. In Section 3 we introduce the framework of discriminative structured prediction model, in particular for CRFs and M^3N. Section 4 defines the mapping from the input space to the feature space. Section 5 describes how to train CRFs and M^3N with online EG algorithm. Section 6 discusses the influence of approximate inference on EG algorithm. In Sec-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE



(a) 374078.jpg (b) 215018.jpg (c) 108055.jpg

Figure 1. Region-adaptive label lattice. The state of each filled node indicates the label of a particular grid.

tion 7 we evaluate both models on a scene understanding task. Finally, Section 8 concludes the paper.

2. Problem Description

Our study follows the recent work on automatic image region annotation. Region annotation, also known as region naming or object recognition [6, 8, 20, 22, 23, 31], aims to learn a model which automatically assigns semantic labels to segmented image regions. Frequently, different concepts may display similar appearances, *e.g.* sky and sea often appear in blue regions. Incorporating spatial priors over semantic concepts may reduce such ambiguities [6, 22, 23, 31]. For example, sky often appears above mountains or buildings while sea does not. Nevertheless, it is usually difficult to characterize the spatial layout of regions, due to the irregular shapes and arbitrary sizes. Here we adopt a region-adaptive grid partition approach (see details in [31]). As shown in Figure 1, we apply adaptively partitioned grids to approximating the segmented regions. The nodes in the lattice are firstly annotated, their labels are then propagated to corresponding regions. In this way, region annotation actually becomes grid annotation. The key problem is how to exploit the spatial dependencies of labels of adjacent grids.

3. Discriminative Structured Prediction

Let $(\mathbf{x}, \mathbf{y})$ denote the pair of the grid-based features and labels. The goal of discriminative structure prediction can be thought of learning a $\mathbf{w}$-parameterized linear discriminant function

$$F(\mathbf{w}, \mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}) \rangle, \tag{1}$$

where Φ maps the pattern $(\mathbf{x}, \mathbf{y})$ from input space $\mathcal{X} \times \mathcal{Y}$ to a feature vector $\Phi(\mathbf{x}, \mathbf{y}) \in \mathcal{R}^Q$; $\mathbf{w}$ is a weight vector in $\mathcal{R}^Q$. The definition of feature representation Φ depends on applications. For our task, we will define it in Section 4. With the discriminant function, the prediction rule is determined by

$$\mathbf{y}^* = f(\mathbf{w}, \mathbf{x}) = \arg \max_{\hat{\mathbf{y}} \in \mathbf{G}(\mathbf{x})} F(\mathbf{w}, \mathbf{x}, \hat{\mathbf{y}}), \tag{2}$$

where the function $\mathbf{G}(\mathbf{x})$ enumerates a set of label configuration candidates for input $\mathbf{x}$; the value of $F(\mathbf{w}, \mathbf{x}, \hat{\mathbf{y}})$ can be understood as a score evaluating the compatibility between $\mathbf{x}$ and $\hat{\mathbf{y}}$. This framework unifies many common classification methods. It can not only predict labels of individual objects but also can output meaningful internal structures within $\mathbf{y}$. Both Conditional Random Fields (CRFs, [16]) and Max-Margin Markov Networks (M^{3}N, [24, 26]) are instances of such discriminative structured prediction framework.

CRFs firstly defines a conditional distribution over labels with function $F(\mathbf{w}, \mathbf{x}, \mathbf{y})$

$$p(\mathbf{y}|\mathbf{x}; \mathbf{w}) = \frac{1}{Z(\mathbf{w}, \mathbf{x})} \exp\{F(\mathbf{w}, \mathbf{x}, \mathbf{y})\}, \tag{3}$$

where $Z(\mathbf{w}, \mathbf{x}) = \sum_{\hat{\mathbf{y}} \in \mathbf{G}(\mathbf{x})} \exp\{F(\mathbf{w}, \mathbf{x}, \hat{\mathbf{y}})\}$ is the partition function. Given a training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, the parameter $\mathbf{w}$ can be learned by minimizing the following regularized log-loss [16, 7, 21]

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \sum_{i=1}^{n} \ell_{LL}(i) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \tag{4}$$

where $\ell_{LL}(i) = -\log p(\mathbf{y}_i|\mathbf{x}_i; \mathbf{w})$ and λ is a constant determining the trade-off between empirical risk and model complexity.

M^{3}N is a model of Support Vector Machines (SVMs) with structured output [24, 26]. Learning parameter $\mathbf{w}$ amounts to solving the following constraint quadratic optimization problem

$$\begin{aligned} \arg \min_{\mathbf{w}} \quad & \textstyle\sum_{i=1}^{n} \xi_i + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle \geq e_i(\mathbf{y}) - \xi_i, \ \forall i, \forall \mathbf{y} \in \mathrm{G}(\mathbf{x}_i), \end{aligned} \tag{5}$$

where $\Psi(\mathbf{x}_i, \mathbf{y}) = \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \mathbf{y})$ and $e_i(\mathbf{y})$, defined in Section 5.3, measures the error between the true labels $\mathbf{y}_i$ and the candidate labels $\mathbf{y}$. Assuming $e_i(\mathbf{y}_i) = 0$ for all i, the so-called hinge loss can be written as

$$\ell_{MM}(i) = \max_{\mathbf{y} \in \mathbf{G}(\mathbf{x}_i)} [e_i(\mathbf{y}) - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle]. \tag{6}$$

Hence, the constraint optimization in Equation 5 can be written as

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} \ell_{MM}(i) + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{7}$$

Comparing Equation 4 and 7, we can find that CRFs and M^{3}N differ only in their loss functions. Both models have a regularization term, which is understood as Bayesian parameter estimation with Gaussian priors for CRFs [21] and as large margin criterion for M^{3}N[24, 26].

4. The Definition of Feature Function $\Phi(\mathbf{x}, \mathbf{y})$

Let $(\mathbf{x}^i, y^i)$ denote the feature/label pair for the i-th grid of an image[1], we assume $\mathbf{x}^i \in \mathcal{R}^D$, $y^i \in \Sigma$ and $|\Sigma| = K$. We use $(\mathbf{x}, \mathbf{y})$=$\{(\mathbf{x}^i, y^i), i = 1, \cdots, m\}$ to denote the input pattern for an image with $m = H \times V$ grids where H and V indicate the number of rows and columns respectively. We follow the way described in [1] to define the feature function $\Phi(\mathbf{x}, \mathbf{y})$. In our case, the structure of the input pattern $(\mathbf{x}, \mathbf{y})$ can be characterized by a graphical model similar to that of Figure 2. Each label variable y associates a state node and each low-level feature vector $\mathbf{x}$ associates an observation node. There are two types of cliques in the graph. We denote the set of cliques covering observation-state nodes by $\mathcal{C}^o$ and denote the set of cliques covering state-state nodes by $\mathcal{C}^s$. We define $\Phi(\mathbf{x}, \mathbf{y})$ over the clique set $\mathcal{C} = \mathcal{C}^o \bigcup \mathcal{C}^s$. The components of $\Phi(\mathbf{x}, \mathbf{y})$ can be categorized into two types according to what types of cliques they are defined over. Each component defined over clique in $\mathcal{C}^o$ conjunctively combines an input attribute $x^d \in \mathcal{R}$ (*i.e.* the d-th entry of the low-level feature vector $\mathbf{x}$) with a state $\sigma_l \in \Sigma$

$$\phi^o_{l,d}(\mathbf{x}, \mathbf{y}) = \sum_{(\mathbf{x}^i, y^i) \in \mathcal{C}^o} [\![y^i = \sigma_l]\!] x^{i,d}, \tag{8}$$

where $[\![\cdot]\!]$ denotes the indicator function of the enclosed predicate. There are KD such components in $\Phi(\mathbf{x}, \mathbf{y})$. Each component defined over clique in $\mathcal{C}^s$ deals with a pair of adjacent states $\sigma_l \in \Sigma$ and $\sigma_{\bar{l}} \in \Sigma$

$$\phi^s_{l,\bar{l}}(\mathbf{x}, \mathbf{y}) = \sum_{(y^i, y^j) \in \mathcal{C}^s} [\![y^i = \sigma_l]\!] [\![y^j = \sigma_{\bar{l}}]\!], \tag{9}$$

There are $2K^2$ such components in $\Phi(\mathbf{x}, \mathbf{y})$.

5. Learning w with Exponentiated Gradient Algorithm

The EG algorithm is originally proposed by Kivinen and Warmuth to learn linear predictors [13]. Bartlett *et al*. show that EG can solve M^{3}N [5]. Globerson *et al*. show that EG can also solve CRFs [10]. Collins *et al*. present a comprehensive study and provide better theoretical justifications to EG algorithm for solving CRFs and M^{3}N [7]. Previous work [5, 7, 10] show that EG empirically outperforms or are competitive to other state-of-the-art approaches (*e.g.* LBFGS[21], stochastic gradient descent [28] for CRFs and SMO [24], cutting plane algorithm [26] for M^{3}N). EG algorithm solves the dual problems of CRFs and M^{3}N. We firstly introduce the dual problems of both models.

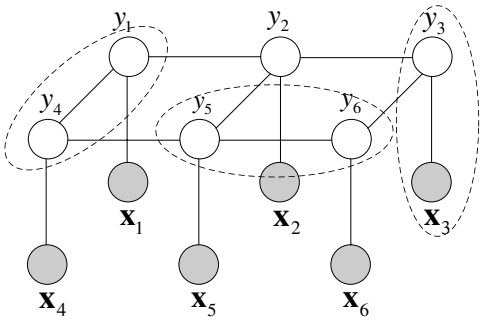


Figure 2. Clique decomposition for graphical model.

5.1. The Dual Problems of CRFs and M^{3}N

Lebanon and Lafferty [7, 18] derive the dual of primal CRFs in Equation 4 as

$$\min_{\boldsymbol{\alpha}} Q_{LL}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{\mathbf{y}} \alpha_{i,\mathbf{y}} \log \alpha_{i,\mathbf{y}} + \frac{1}{2\lambda}\|\mathbf{w}(\boldsymbol{\alpha})\|^2$$
$$\text{s.t.} \sum_{\mathbf{y}} \alpha_{i,\mathbf{y}} = 1, \alpha_{i,\mathbf{y}} \geq 0, \forall i, \forall \mathbf{y} \in \mathbf{G}(\mathbf{x}_i), \tag{10}$$

where $\mathbf{w}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{\mathbf{y}} \alpha_{i,\mathbf{y}} \Psi(\mathbf{x}_i, \mathbf{y})$. The primal solution can be constructed from the dual one according to $\mathbf{w}^* = \frac{1}{\lambda}\mathbf{w}(\boldsymbol{\alpha}^*)$. Tasker *et al*. [5, 7, 24] derive the dual of M^{3}N in Equation 7 as

$$\min_{\boldsymbol{\alpha}} Q_{MM}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{\mathbf{y}} \alpha_{i,\mathbf{y}} e_i(\mathbf{y}) + \frac{1}{2\lambda}\|\mathbf{w}(\boldsymbol{\alpha})\|^2$$
$$\text{s.t.} \sum_{\mathbf{y}} \alpha_{i,\mathbf{y}} = 1, \alpha_{i,\mathbf{y}} \geq 0, \forall i, \forall \mathbf{y} \in \mathbf{G}(\mathbf{x}_i), \tag{11}$$

where $\mathbf{w}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{\mathbf{y}} \alpha_{i,\mathbf{y}} \Psi(\mathbf{x}_i, \mathbf{y})$. We can also get the primal solution from the dual one by $\mathbf{w}^* = \frac{1}{\lambda}\mathbf{w}(\boldsymbol{\alpha}^*)$. In the following, we use $\triangle_i$ to denote the constraints $\{\sum_{\mathbf{y}} \alpha_{i,\mathbf{y}} = 1, \alpha_{i,\mathbf{y}} \geq 0\}$. Constraints $\triangle_i$ imply the feasible dual variables for the i-th example, *i.e.* $\boldsymbol{\alpha}_i = \{\alpha_{i,\mathbf{y}}, \mathbf{y} \in \mathbf{G}(\mathbf{x}_i)\}$, are in a probability simplex.

5.2. Online EG Updates of Dual Variables

Online EG is an iterative algorithm. In each iteration, the dual variables for a specific example is updated. Concretely,

[1] Please distinguish $\mathbf{x}_i$ and $\mathbf{x}^i$. We use subscript for image-level variable while use superscript for grid-level variable

given the current dual variables $\boldsymbol{\alpha}_i$ for the i-th example, the updated dual variables $\boldsymbol{\alpha}_i^{'}$ can be obtained by [7]

$$\alpha_{i,\mathbf{y}}^{'} = \frac{1}{Z_i}\alpha_{i,\mathbf{y}} \exp\{-\eta \nabla_{i,\mathbf{y}}\}, \forall \mathbf{y} \in \mathbf{G}(\mathbf{x}_i), \tag{12}$$

where $\nabla_{i,\mathbf{y}} = \frac{\partial Q(\boldsymbol{\alpha})}{\partial \alpha_{i,\mathbf{y}}}$; $Z_i = \sum_{\hat{\mathbf{y}}} \alpha_{i,\hat{\mathbf{y}}} \exp\{-\eta \nabla_{i,\hat{\mathbf{y}}}\}$ is a normalization constant ensuring the new variables $\boldsymbol{\alpha}_i^{'}$ still constituting a valid probability distribution; the parameter $\eta > 0$ is a learning rate. For the dual problem of CRFs, the gradient is

$$\nabla_{i,\mathbf{y}} = 1 + \log \alpha_{i,y} + \frac{1}{\lambda}\langle \mathbf{w}(\boldsymbol{\alpha}), \Psi(\mathbf{x}_i, \mathbf{y})\rangle \tag{13}$$

and for the dual problem of M^{3}N, the gradient is

$$\nabla_{i,\mathbf{y}} = -e_i(\mathbf{y}) + \frac{1}{\lambda}\langle \mathbf{w}(\boldsymbol{\alpha}), \Psi(\mathbf{x}_i, \mathbf{y})\rangle. \tag{14}$$

For online EG algorithm, Collins *et al*. [7] prove that, to get an approximate optimal solution with accuracy ϵ, CRFs require $O(\log(\frac{1}{\epsilon}))$ EG updates and M^{3}N requires $O(\frac{1}{\epsilon})$ EG updates. A notable problem is that the size of $\boldsymbol{\alpha_i}$ equals to that of $\mathbf{G}(\mathbf{x}_i)$, which means the number of dual variables may be exponential in size, *e.g.* $|\boldsymbol{\alpha_i}| = K^m$ for an image with m grids. Directly operating $\boldsymbol{\alpha}$ is infeasible. Next, we introduce how to overcome this challenge.

5.3. Part Factorization Trick for Dual Variables

As aforementioned in Section 5.1, for the i-th example, the feasible dual variables $\boldsymbol{\alpha}_i$ constitute a probability distribution. Taskar *et al*. [24] originally show that the distribution $\boldsymbol{\alpha}_i$ can be represented by polynomial number of marginal terms. Thus, we do not need to directly manipulate the exponential number of dual variables. Specifically for EG algorithm, Bartlett *et al*. [5] and Collins *et al*. [7] show that, if we constrain the probability distribution $\boldsymbol{\alpha}_i$ to Gibbs distribution, an efficient EG update algorithm can be designed, meanwhile it does not affect the theoretical convergence properties. Here we apply their results and introduce how to implement it in our scenario.

Given a clique $c \in \mathcal{C}$, we define $\mathcal{Y}(c)$ to be the set of possible label configurations for that clique and define $y(c)$ to be the value of $\mathbf{y}$ on that clique. For example, if $c \in \mathcal{C}^o$, $\mathcal{Y}(c)$ equals to Σ, while if $c \in \mathcal{C}^s$, $\mathcal{Y}(c)$ equals to $\Sigma \times \Sigma$. We decompose each $\mathbf{y} \in \mathbf{G}(\mathbf{x})$ into set of parts based on the clique decomposition $\mathcal{C}$, with a part for each clique. Concretely, the set of parts for input pattern $(\mathbf{x}_i, \mathbf{y}_i)$ is defined as

$$R(\mathbf{x}_i, \mathbf{y}_i) = \{(c, y(c)) | c \in \mathcal{C}\}, \tag{15}$$

and the set of parts for all the possible patterns with observation $\mathbf{x}_i$ is defined as

$$R(\mathbf{x}_i) = \bigcup_{\mathbf{y} \in \mathbf{G}(\mathbf{x}_i)} R(\mathbf{x}_i, \mathbf{y}) = \{(c, a) | c \in \mathcal{C}, a \in \mathcal{Y}(c)\}. \tag{16}$$

It is straightforward that $R(\mathbf{x}_i, \mathbf{y}_i)$ and $R(\mathbf{x}_i)$ have $3m–H–V$ and $mK+(2m–H–V)K^2$ elements respectively. If we define a variable $\theta_{i,r} \in \mathcal{R}$ for each part $r \in R(\mathbf{x}_i)$ and constrain $\boldsymbol{\alpha}_i$ in exponential families, any $\boldsymbol{\alpha}_i$ can be determined by $\boldsymbol{\theta}_i$[5, 7]

$$\alpha_{i,\mathbf{y}} = \boldsymbol{\sigma}(\boldsymbol{\theta}_{i,\mathbf{y}}) = \frac{\exp\{\sum_{r \in R(\mathbf{x}_i, \mathbf{y})} \theta_{i,r}\}}{\sum_{\mathbf{y}' \in \mathbf{G}(\mathbf{x}_i)} \exp\{\sum_{r \in R(\mathbf{x}_i, \mathbf{y}')} \theta_{i,r}\}}, \tag{17}$$

which means, K^m components of $\boldsymbol{\alpha}_i$ can be represented by $\boldsymbol{\theta}_i$ with much fewer components (*i.e.* $mK+(2m–H–V)K^2$). Moreover, the following lemma shows that, multiplicatively updating $\boldsymbol{\alpha}_i$ with Equation 12 can be accomplished by additively updating $\boldsymbol{\theta}_i$.

Lemma 1 *(Collins et al. [7]) For a given $\boldsymbol{\alpha} \in \triangle^n$, and for a given $i \in [1, \cdots, n]$, take $\boldsymbol{\alpha}_i^{'}$ to be the updated value for $\boldsymbol{\alpha}_i$ derived using an EG step in Equation 12. Suppose that for some G_i and $g_{i,r}$, we can write $\nabla_{i,\mathbf{y}} = G_i + \sum_{r \in R(\mathbf{x}_i, \mathbf{y})} g_{i,r}$ for all $\mathbf{y} \in \mathbf{G}(\mathbf{x}_i)$. Then if $\boldsymbol{\alpha}_i$ can be parameterized in an exponential form according to Equation 17, that is, $\boldsymbol{\alpha}_i = \boldsymbol{\sigma}(\boldsymbol{\theta}_i)$ with some $\boldsymbol{\theta}_i \in \mathcal{R}^{|R(\mathbf{x}_i)|}$, we define $\theta_{i,r}^{'} = \theta_{i,r} - \eta g_{i,r}$ for all $r \in R(\mathbf{x}_i)$, it follows that $\boldsymbol{\alpha}_i^{'} = \boldsymbol{\sigma}(\boldsymbol{\theta}_i^{'})$.*

The above lemma requires that the gradients in Equation 13 and 14 can be factorized into the sum of a global value G_i and some part-based values $g_{i,r}$ for any $r \in R(\mathbf{x}_i)$. Next, we show how to accomplish it. To do this, we firstly show $\Psi(\mathbf{x}_i, \mathbf{y})$ and $e_i(\mathbf{y})$ for $\mathbf{y} \in \mathbf{G}(\mathbf{x}_i)$ can be factorized into parts. According to the definition of feature functions in Equation 8 and 9, it is easy to get

$$\Psi(\mathbf{x}_i, \mathbf{y}) = \sum_{r \in R(\mathbf{x}_i, \mathbf{y}_i)} \Phi(\mathbf{x}_i, r) - \sum_{r \in R(\mathbf{x}_i, \mathbf{y})} \Phi(\mathbf{x}_i, r), \tag{18}$$

where the components of $\Phi(\mathbf{x}, r)$ are defined as

$$\phi_{l,d}^o(\mathbf{x}, r) = [\![(\mathbf{x}^i, y^i) \in r]\!][\![y^i = \sigma_l]\!] x^{i,d}, \tag{19}$$

$$\phi_{l,l}^s(\mathbf{x}, r) = [\![(y^i, y^j) \in r]\!][\![y^i = \sigma_l]\!][\![y^j = \sigma_{\bar{l}}]\!]. \tag{20}$$

Also $e_i(\mathbf{y})$ can be defined in factorization style

$$e_i(\mathbf{y}) = \sum_{r \in R(\mathbf{x}_i, \mathbf{y})} e_i(r) = \sum_{r \in R(\mathbf{x}_i, \mathbf{y})} [\![r \notin R(\mathbf{x}_i, \mathbf{y}_i)]\!]. \tag{21}$$

With the above factorization results of $\alpha_{i,\mathbf{y}}$ (in Equation 17), $\Psi(\mathbf{x}_i, \mathbf{y})$ (in Equation 18) and $e_i(\mathbf{y})$ (in Equation 21), the gradient $\nabla_{i,\mathbf{y}} = \frac{\partial Q_{LL}(\boldsymbol{\alpha})}{\partial \alpha_{i,\mathbf{y}}}$ can be factorized as

$$G_i = 1 - \log Z(\mathbf{y}) + \frac{1}{\lambda}\langle \mathbf{w}(\boldsymbol{\alpha}), \Phi(\mathbf{x}_i, \mathbf{y}_i)\rangle, \tag{22}$$

$$g_{i,r} = \theta_{i,r} - \frac{1}{\lambda}\langle \mathbf{w}(\boldsymbol{\alpha}), \Phi(\mathbf{x}_i, r)\rangle, \tag{23}$$

where $Z(\mathbf{y}) = \sum_{\mathbf{y}' \in \mathbf{G}(\mathbf{x}_i)} \exp\{\sum_{r \in R(\mathbf{x}_i, \mathbf{y}')} \theta_{i,r}\}$ is a normalization constant. And the gradient $\nabla_{i,\mathbf{y}} = \frac{\partial Q_{MM}(\boldsymbol{\alpha})}{\partial \alpha_{i,\mathbf{y}}}$ can be factorized as

$$G_i = \frac{1}{\lambda} \langle \mathbf{w}(\boldsymbol{\alpha}), \Phi(\mathbf{x}_i, \mathbf{y}_i) \rangle, \tag{24}$$

$$g_{i,r} = -e_i(r) - \frac{1}{\lambda} \langle \mathbf{w}(\boldsymbol{\alpha}), \Phi(\mathbf{x}_i, r) \rangle. \tag{25}$$

Therefore, suitable updates over part-based variables $\boldsymbol{\theta}_i$ for Q_{LL} and Q_{MM} objectives respectively are [7]

$$\theta'_{i,r} = \theta_{i,r} - \eta \Big(\theta_{i,r} - \frac{1}{\lambda} \langle \mathbf{w}(\boldsymbol{\alpha}), \Phi(\mathbf{x}_i, r) \rangle \Big), \tag{26}$$

$$\theta'_{i,r} = \theta_{i,r} - \eta \Big(- e_i(r) - \frac{1}{\lambda} \langle \mathbf{w}(\boldsymbol{\alpha}), \Phi(\mathbf{x}_i, r) \rangle \Big). \tag{27}$$

6. Learning with Approximate Inference

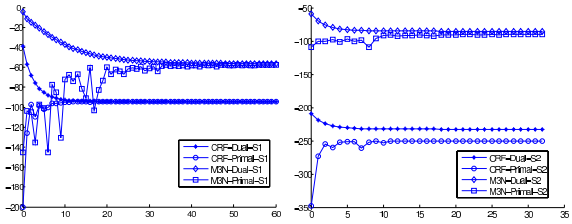
We need to solve three types of inference problems either for training or for decoding the structured output models. They are, (i) computing marginals, (ii) calculating partition functions (iii) finding maximum a posterior (MAP) label configuration. Since there exist cycles in the graph structure of our problem, exact inference is intractable. We resort to approximate inference. In the following, we will identify the cases when we need to solve these inference problems and discuss what are the influences of approximate inference on EG.

First, either for getting the primal solution according to $\mathbf{w}^* = \frac{1}{\lambda} \mathbf{w}(\boldsymbol{\alpha}^*)$ or for performing the updates in Equation 26 and 27, we need to calculate $\mathbf{w}(\boldsymbol{\alpha})$ by

$$\mathbf{w}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \Phi(\mathbf{x}_i, \mathbf{y}_i) - \sum_{i=1}^{n} \sum_{r \in R(\mathbf{x}_i)} \mu_{i,r}(\boldsymbol{\theta}_i) \Phi(\mathbf{x}_i, r), \tag{28}$$

where $\mu_{i,r}(\boldsymbol{\theta}_i) = \sum_{\mathbf{y}: r \in R(\mathbf{x}_i, \mathbf{y})} \alpha_{i,\mathbf{y}}$ [7]. Note that $\boldsymbol{\alpha}_i$ follows gibbs distribution as defined by Equation 17, implying $\mu_{i,r}(\boldsymbol{\theta}_i)$ can be thought of marginal probability for part r. For this case, we use loopy sum-product algorithm. Second, we need to calculate the partition functions to obtain the objective values for both the primal and dual models of CRFs. For this case, we use the Bethe free energy approximation approach [30]. Third, we need to solve Equation 2 for decoding both CRFs and M^{3}N, and we also need to calculate the hinge loss for each example to get the objective value of primal M^{3}N. All these cases resort to the third type of inference problem. We adopt Tree Reweighted Message Passing approach for MAP inference [29].

Note that approximate inference may affect the convergence properties of EG algorithm. With approximate inference, we can only get approximate gradient $g_{i,r}$ for update. However, we do not know whether EG will converge



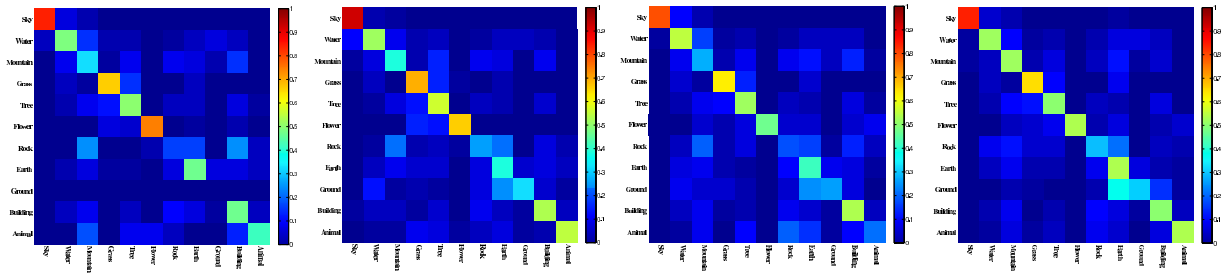
(a) Stage 1 (b) Stage 2

Figure 3. Primal and dual objective values for CRFs and M^{3}N in two-stage training procedure. The coordinates of x-axis indicate the number of iterations.

with approximate inference, since the theoretical guarantees for convergence of EG are obtained by assuming exact gradients can be computed [5, 7, 10]. In our experiments, both CRFs and M^{3}N trained by EG with approximate inference yield very poor performances, even worse than those of multi-class Logistic Regression (MLR) and multi-class Support Vector Machines (MSVMs). Note that MLR and MSVMs are special cases of CRFs and M^{3}N respectively, which only use the feature functions in Equation 8 but do not incorporate the label interaction features in Equation 9. Kulesza *et al*. [14] observe similar phenomena that learning may fail with approximate inference. They argue that approximate inference can reduce the expressivity of models and may lead the learning algorithm astray. In our case, the reason may be that, the errors in $\mathcal{C}^s$ caused by approximate inference are propagated to cliques $\mathcal{C}^o$ and make the EG algorithm fail. To overcome this problem, we design a two-stage training approach to prevent the error propagations. In the first stage, only the part variables for the cliques in $\mathcal{C}^o$ are updated, which amounts to training models of MLR and MSVMs. In this stage, exact inference can be performed, convergence is theoretically guaranteed. As the example shown in Figure 3(a), the dual gap in the first stage converges to zero. In the second stage, the part variables for cliques in $\mathcal{C}^o$ are kept unchanged and only the variables for cliques in $\mathcal{C}^s$ are updated. In this stage, approximate inference is performed. Also see the example in Figure 3(b), the dual gap gradually shrinks but does not converge to zero. With two-stage setting, the inaccuracies caused by approximate inference will not affect the weights with respect to the cliques in $\mathcal{C}^o$.

7. Experiments

In this section, we evaluate both CRFs and M^{3}N on the task of scene understanding. In particularly, we are interested in two problems, *i.e*. the effectiveness of discriminative structured prediction and the empirical comparison between CRFs and M^{3}N.



(a) MLR (b) CRFs (c) MSVMs (d) M^{3}N

Figure 4. Confusion matrix of average categorization accuracy for the four implemented algorithms. The brightness of intersected block indicates the probability of classifying the concept in y-axis as the concept in x-axis.

Table 1. Categorization accuracies of the four approaches.

	MLR	CRFs	MSVMs	M^{3}N
Accuracy	0.597	0.623	0.591	0.612

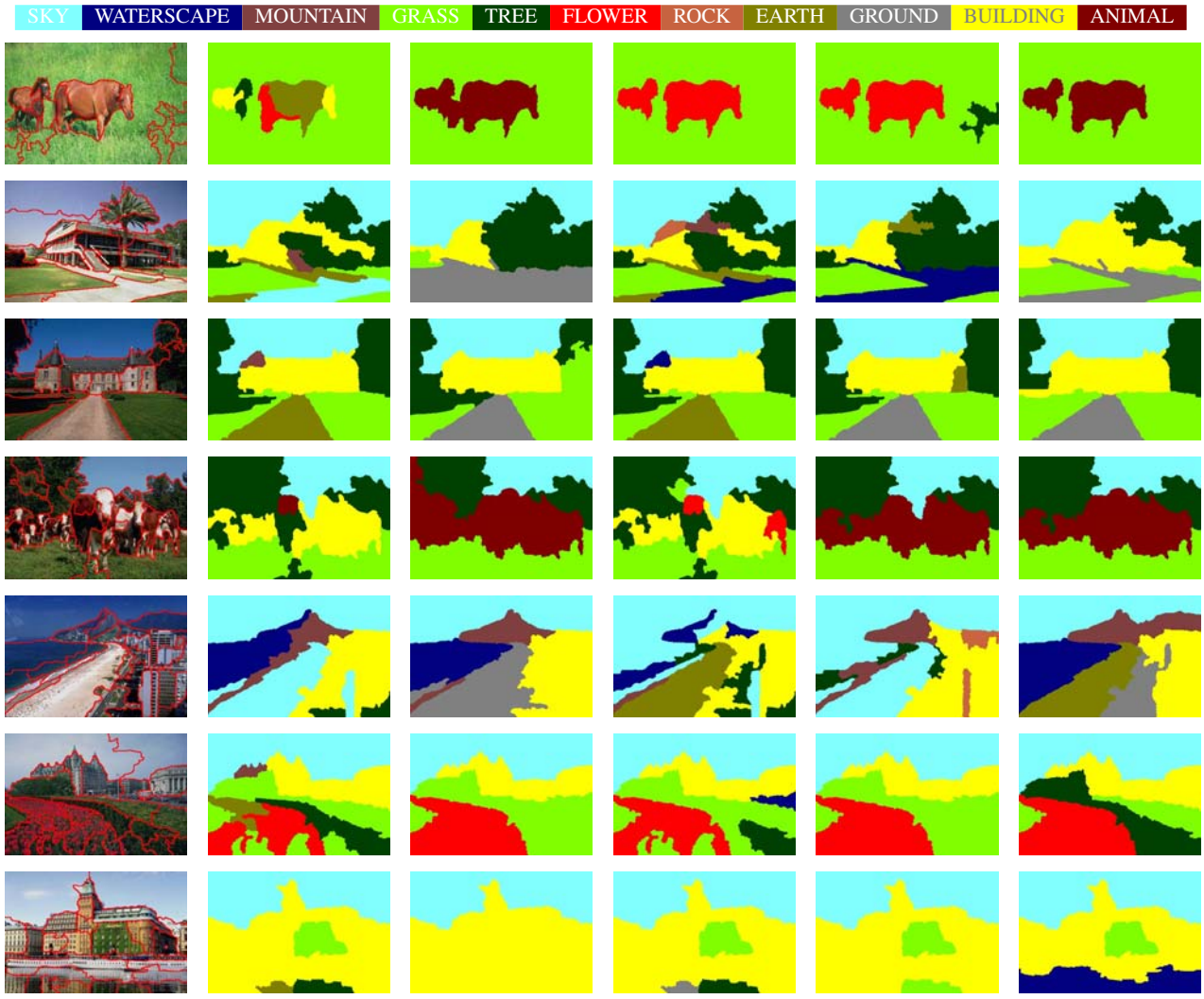
7.1. Experimental Setup

Discriminative training for structured output model requires image set with region-level groundtruth. Unfortunately, in most available image set, descriptive keywords are associated with entire images rather than individual regions. Some image sets with region-level groundtruth contain only several hundred images (*e.g.* MSRC [22], Sowerby [12]). Alternatively, we use a much larger scale data set [31]. 4002 outdoor images are chosen from Corel Stock Photo CDs. All the images are segmented into regions by JSEG algorithm [9]. Totally, 104,626 regions are obtained. For each region, 9-dimensional color moment in HSV color space and 20-dimensional Pyramid-structured wavelet texture are extracted to describe its appearance. One of 11 semantic concepts are manually annotated to each region, including sky, water, mountain, grass, tree, flower, rock, earth, ground, building and animal. The data set is randomly split into two sub-sets in equal size for training and testing respectively. For every algorithm, we use the same region-adaptive approach to construct the grid-structure graphical model. More detailed information on data set can be found in [31].

7.2. Results

We implement four approaches for comparison, including multi-class Logistic Regression (MLR), multi-class SVMs (MSVMs), CRFs and M^{3}N. The first two approaches are special cases for CRFs and M^{3}N respectively. They only use feature functions defined in Equation 8, which amounts to learning the mapping between appearance features and semantic labels without considering the spatial dependencies among labels. The last two approaches use both features in Equation 8 and 9, taking the spatial dependencies among adjacent labels into account. For all the algorithms, we use the fixed learning rate $\eta = 0.5$ and $\lambda = 0.005$. 408 images from training set are held out for cross validation purpose. We found that $\lambda = 2$ is a good choice for all the models. In the following, we report the performance of each algorithm under the best chosen parameters. In Table 1 we summarize the grid-based categorization accuracies for all the approaches. Note that the criterion adopted here is different from that used in [31], which uses averaged recall and precision to measure the performances. We have several observations from the results. First, the performances yielded by structured output models outperform those of non-structured output models. For example, CRFs obtains a relative 4.4% increasement over MLR and M^{3}N gains a relative 3.6% performance increasement compared to MSVMs. Second, we have not found the superiors of hinge-loss models over log-loss models, because MLR and CRFs have outperformed MSVMs and M^{3}N respectively. This is inconsistent with the conclusions of previous work [24, 26] which find that M^{3}N slightly outperforms CRFs. We guess this may be due to inaccuracies caused by approximate inference, because the previous results favoring M^{3}N are obtained in tasks where exact inference is tractable (*e.g.* graphical models without cycles). Nevertheless, our result is consistent with the observation of Vapnik [27], who has not found the superiors of SVMs over logistic regression. Therefore, it is difficult to say which is better. Both hinge loss (*e.g.* MSVMs, M^{3}N) and log loss (*e.g.* MLR, CRFs) are state-of-the-art methods. The final observation is that EG for M^{3}N converges slower than that for CRFs, which is consistent with the theoretical results and empirical observations in [7].

For details on the classification accuracies of each category, we show the confusion matrices of the four models in Figure 4. As shown in Figure 4(a) and 4(c), MLR and MSVMs tend to classifying mountain and building as rock. When combined spatial context, CRFs and M^{3}N reduces such ambiguities and boost the accuracies of rock and animal. We also present several examples of region annotation in Figure 5. For most images, we can find that models with spatial priors (*i.e.* CRFs and M^{3}N) improve the recogni-



(a) JSEG (b) MLR (c) CRFs (d) MSVMs (e) M^3N (f) Groundtruth

Figure 5. Some example results. The first row shows the color-concept correspondence relationship. The first column show the segmented regions by JSEG. From the second to the fifth columns, we show the annotation results by MLR, CRFs, MSVMs and M^3N respectively. The last column shows the groundtruth.

tion accuracies compared to those without spatial priors(*i.e.* MLR and MSVMs). However, note that if the clues from spatial priors dominate the clues from local appearances, the approaches may yield over-smoothed results, *e.g.* the image in the last row.

8. Conclusions

In this paper, CRFs and M^3N are demonstrated to solve scene understanding task. More specifically, we describe how to unifiedly represent local appearances and spatial priors with structured output model. We also show how to solve both models with online EG algorithm. In particularly, we discuss the influences of approximate inference on EG approach. Both CRFs and M^3N yield encouraging results and their performances are comparable. Theoretically, the two implemented models differ only in their loss function, *i.e.* with log-loss and hinge loss respectively. Altun *et al.* [1] proved that both log-loss and hinge loss upper bound the desired zero-one loss. Our work can be thought as an empirical study to the effect of loss functions on scene understanding. Finally, we would like to point out that, though discriminative structured prediction models is unified with linear discriminant function, that does not mean the method can only model linear dependencies. Note that both duals for CRFs and M^3N are determined only by the inner product matrix of feature representation, which means the mod-

els can be conveniently kernelized just as what SVMs does [2, 17, 24]. With nonlinear kernels, more complex dependencies among $\mathbf{x}$ and $\mathbf{y}$ can be modeled. It is promising to evaluate whether kernelized model will bring benefits to the categorization accuracies to tasks such as scene understanding.

9. Acknowledgements

This work was supported by the National Natural Science Foundation of China under the grant No. 60621062 and 60605003, the National Key Foundation R&D Projects under the grant No. 2003CB317007, 2004CB318108 and 2007CB311003, and the Basic Research Foundation of Tsinghua National Laboratory for Information Science and Technology (TNList). Finally, special thanks go to Prof. Michael Collins, Terry Koo and Dr. Xavier Carreras for providing the package *egstra* and having extensive discussions on EG algorithms.

References

[1] Y. Altun and T. Hofmann. Large margin methos for label sequence learning. In *Proc. of EuroSpeech*, 2003. 1, 3, 7

[2] Y. Altun, A. J. Smola, and T. Hofmann. Exponential families for conditional random fields. In *Proc. of UAI*, pages 2–9, 2004. 8

[3] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A. Ng. Discriminative learning of markov random fields for segmentation of 3D scan data. In *Proc. of CVPR*, pages 169–176, 2005. 1

[4] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5:617–629, August 2004. 1

[5] P. L. Bartlett, M. Collins, B. Taskar, and D. McAllester. Exponentiated gradient algorithms for large-margin structured classification. In *Proc. of NIPS*, 2005. 3, 4, 5

[6] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proc. of ECCV*, pages 350–362, 2004. 1, 2

[7] M. Collins, A. Globerson, T. Koo, X. Carreras, and P. Bartlett. Exponentiated gradient algorithms for conditional random fields and max-margin markov networks. In *to appear in JMLR*, 2008. 1, 2, 3, 4, 5, 6

[8] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Proc. of CVPR*, pages 10–17, 2005. 1, 2

[9] Y. Deng and B.S.Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(8):800–810, 2001. 6

[10] A. Globerson, T. Y. Koo, X. Carreras, and M. Collins. Exponentiated gradient algorithms for log-linear structured prediction. In *Proc. of ICML*, pages 305–312, 2007. 3, 5

[11] L. Gu, E. P. Xing, and T. Kanade. Learning GMRF structures for spatial priors. In *Proc. of CVPR*, pages 1–6, 2007. 1

[12] X. He, R. S. Zemel, and M. Á. C.-P. nán. Multiscale conditional random fields for image labeling. In *Proc. of CVPR*, pages 695–702, 2004. 1, 6

[13] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput.*, 132(1):1–63, 1997. 3

[14] A. Kulesza and F. Pereira. Structured learning with approximate inference. In *Proc. of NIPS 20*, pages 785–792, 2008. 5

[15] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proc. of ICCV*, pages 1150–1159, 2003. 1

[16] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, pages 282–289, 2001. 1, 2

[17] J. Lafferty, X. Zhu, and Y. Liu. Kernel conditional random fields: representation and clique selection. In *Proc. of ICML*, pages 64–71, 2004. 8

[18] G. Lebanon and J. Lafferty. Boosting and maximum likelihood for exponential models. In *Proc. of NIPS*, 2002. 3

[19] S. Z. Li. *Markov random field modeling in image analysis*. Springer-Verlag New York, Inc., 2001. 1

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

[21] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proc. of NAACL*, 2003. 1, 2, 3

[22] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proc. of ECCV*, pages 1–15, 2006. 1, 2, 6

[23] A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. In *Proc. of CVPR*. 1, 2

[24] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *Proc. of NIPS*, 2004. 1, 2, 3, 4, 6, 8

[25] A. Torralba. Contextual priming for object detection. *Int. J. Comput. Vision*, 53(2):169–191, 2003. 1

[26] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, 2005. 1, 2, 3, 6

[27] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. 6

[28] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy. Accelerated training of conditional random fields with stochastic gradient methods. In *Proc. of ICML*, pages 969–976, 2006. 3

[29] M. Wainwright, T. Jaakkola, and A. Willsky. MAP estimation via agreement on trees: message-passing and linear programming. *IEEE Trans. Infomation Theory*, 51(11):3697–3717, 2005. 5

[30] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. Technical report, IJCAI 2001 Distinguished Lecture track, 2001. 5

[31] J. Yuan, J. Li, and B. Zhang. Exploiting spatial context constraints for automatic image region annotation. In *Proc. of ACM Multimedia*, pages 595–604, 2007. 1, 2, 6