Discriminative Random Fields

Sanjiv Kumar (sanjivk@google.com)
Google Research, 1440 Broadway, New York, NY 10018, USA

Martial Hebert (hebert@ri.cmu.edu)
The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract.
In this research we address the problem of classification and labeling of regions given a single static natural image. Natural images exhibit strong spatial dependencies, and modeling these dependencies in a principled manner is crucial to achieve good classification accuracy. In this work, we present Discriminative Random Fields (DRFs) to model spatial interactions in images in a discriminative framework based on the concept of Conditional Random Fields proposed by Lafferty et al (Lafferty et al., 2001). The DRFs classify image regions by incorporating neighborhood spatial interactions in the labels as well as the observed data. The DRF framework offers several advantages over the conventional Markov Random Field (MRF) framework. First, the DRFs allow to relax the strong assumption of conditional independence of the observed data generally used in the MRF framework for tractability. This assumption is too restrictive for a large number of applications in computer vision. Second, the DRFs derive their classification power by exploiting the probabilistic discriminative models instead of the generative models used for modeling observations in the MRF framework. Third, the interaction in labels in DRFs is based on the idea of pairwise discrimination of the observed data making it data-adaptive instead of being fixed a priori as in MRFs. Finally, all the parameters in the DRF model are estimated simultaneously from the training data unlike the MRF framework where the likelihood parameters are usually learned separately from the field parameters. We present preliminary experiments with man-made structure detection and binary image restoration tasks, and compare the DRF results with the MRF results.

Keywords: Image Classification, Spatial Interactions, Markov Random Field, Discriminative Random Field, Discriminative Classifiers, Graphical Models

1. Introduction

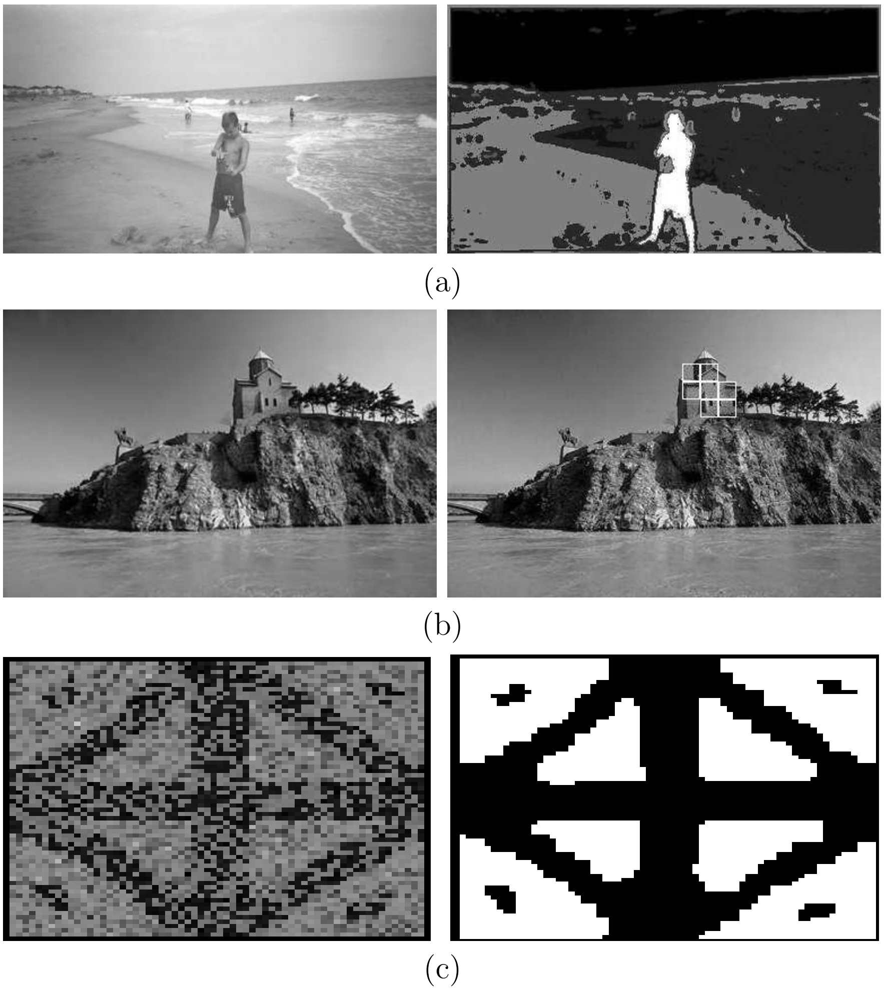
One of the fundamental problems in computer vision is that of *image understanding* or *semantic scene interpretation* i.e., to interpret the scene contained in an image in meaningful entities. For instance, one might be interested in knowing either the general class of a scene e.g., the scene is an *office* or a *beach*, or the event summarizing a scene e.g., the scene is from a *birthday party.* To solve these complex problems, it is important to classify various regions and objects in a scene in meaningful categories. For example, if we can recognize that the scene contains *water* and *sand*, there is high probability that the scene is a *beach.* Similarly, presence of a *birthday cake* is a strong indication of the scene being from a *birthday party.*

In this research we address the problem of classification or labeling of regions in natural images where the term *region* may denote an image pixel, a block of a regular grid, an irregular patch in the image or an object itself. Following the convention, by *natural images* we mean the non-contrived scenes that are encountered commonly in our surroundings i.e., regular indoor and outdoor scenes. These images contain both man-made and other regions or objects occurring in nature. In this research we will deal with the problems where only a single static image of the scene is given, and no 3D geometric or motion information is available. This makes the classification task more challenging.

It is well known that natural images are not a random collection of independent pixels or blocks. It is important to use the contextual information in the form of spatial dependencies in images for the analysis of natural images. In fact, one would like to have total freedom in modeling long range complex data interactions in an image without restricting oneself to small local neighborhoods. This idea forms the core of the research presented in this paper. The spatial dependencies may vary from being local to global and the challenge is how to maintain global spatial consistency using models that only need to consider relatively local dependencies.

1.1. The Nature of Spatial Interactions

There are typically two types of spatial interactions one would like to model for the purpose of classification and labeling. First is the notion of spatial smoothness of labels in natural images. According to this, neighboring sites tend to have similar labels (except at the discontinuities). For example if a pixel in Fig. 1(a) has label *sky*, there is a high probability that the neighboring pixels also have the same label except at the boundaries. In fact, this underlying smoothness of image labels is the reason that one can hope to recover the true image from its corrupted version in image denoising tasks (Fig. 1(c)), which is otherwise an ill-posed problem. In addition to spatial interaction in labels, there are also complex interactions in the observed data that might be required for the task of classification. Consider a task of detecting structured textures (e.g., man-made structures such as buildings) in a given image. The data belonging to this type of textures is highly dependent on its neighbors. This is because, in man-made structures, the lines or edges at spatially adjoining sites follow some underlying organization rules rather than being random (See Fig.1(b)). The task of labeling image regions encompasses a wide range of applications in computer vision including (semantic) segmentation, image denoising, texture recognition etc. (Fig. 1). Ideally one would like to find a computational model that can learn relevant dependencies of different types automatically in a single consistent framework using the training data. In this paper we address the challenge of how to model arbitrarily complex dependencies in the observed image data as well as the labels in a principled manner.



(a)

(b)

(c)

Figure 1. Various tasks in computer vision that require explicit consideration of spatial dependencies for the purpose of region labeling. (a) Segmentation and labeling of input image in meaningful regions. (b) Detection of structured textures such as buildings. (c) Image denoising to restore images corrupted by noise.

1.2. Modeling Spatial Interactions

While modeling spatial interactions in images, it is important to take into consideration statistical variations in data within each class and other uncertainties due to image noise etc. This naturally leads toward probabilistic modeling of classification problems including the spatial dependencies. In probabilistic models, the final classification task can be seen as inference over these models with respect to some cost function. As discussed before, natural images exhibit long range dependencies and manipulating these global interactions is of fundamental interest in classification. However, direct modeling of global interactions becomes computationally intractable even for a small image. On the contrary, usually one can encode the structure of local dependencies in an image easily from which one would like to make globally consistent predictions. This paradox can be resolved to a large extent by *graphical models*. Graphical models combine two areas viz. *graph theory* and *probability theory*, and provide a powerful yet flexible framework for representing and manipulating *global* probability distributions defined by relatively *local* constraints.

This paper is organized as follows. In Section 2 we discuss the background research in modeling spatial interactions in computer vision. Specifically we mention commonly used causal and noncausal models, and highlight their limitations when applied to vision tasks. In Section 3 we present a noncausal Discriminative Random Field (DRF) model along with parameter learning and inference over these models. Experiments with man-made structure detection task are also presented. Section 4 describes modifications in the original DRF framework and further experiments with binary image denoising task. Finally, we highlight the contributions of this paper in Section 5, and also discuss further extensions of the proposed DRF framework as the future work.

2. Background

The issue of incorporating spatial dependencies in various image analysis tasks has been of continuous interest in vision community. In the vision literature, broadly two different categories of approaches have been used to address this issue: *non-probabilistic* and *probabilistic*. We categorize a framework as non-probabilistic if the overall labeling objective is not given by a consistent probabilistic formulation even if the framework utilizes probabilistic methods to address parts of it.

Among the non-probabilistic approaches, other than the weak measures of capturing spatial smoothness of natural images using filters with local neighborhood supports (Guo et al., 2003)(Kumar et al., 2003), perhaps the most popular one is (*Relaxation Labeling RL*) proposed by Rosenfeld et al. (Rosenfeld et al., 1976). This work was inspired by the work of Waltz (Waltz, 1975) concerned with discrete relaxation on how to impose a global consistency on the labelings of idealized line drawings where the object and object primitives were assumed to be given. Since the introduction of RL, several probabilistic relaxation approaches have been suggested to provide a better explanation of the original heuristic updates of the label responsibilities (Kittler and Pairman, 1985)(Kittler and Hancock, 1989)(Christmas et al., 1995). In spite of successes of probabilistic RL in several applications, there are many ad-hoc assumptions in various RL frameworks (Kittler, 1997). For example, either the labels are assumed to be independent given the relational measurements at two or more sites (Christmas et al., 1995) or conditionally independent in local neighborhood of a site given its label (Kittler and Hancock, 1989). Probably the most important problem with RL is that the model parameters, i.e., the compatibility coefficients are chosen on a heuristic basis. In fact, it is not even clear how to interpret these coefficients (Kittler and Illingworth, 1985). The probabilistic versions of RL do allow viewing compatibilities as conditional probabilities. However, even these interpretations are valid only for the first iteration. The meaning of the probabilities yielded by subsequent iterations is increasingly speculative (Kittler and Illingworth, 1985).

In the probabilistic schemes, two types of graphical models, i.e. causal and noncausal models have been used to incorporate spatial contextual constraints in vision problems.

2.1. Causal Models

Causal models are global probability distributions defined on directed graphs using local transition probabilities. If a causal graph is acyclic[1]. If we denote the set of labels on image sites by $\boldsymbol{x}$,

$$P(\boldsymbol{x}) = \prod_i P(x_i|pa_i),$$

where pa_i is the parent of node i. The causal models assume that the observed image is produced by a latent causal hierarchical process. A particular form of such models is a causal tree in which each node has only one parent. These models have been used with some success in various segmentation and labeling problems (Bouman and Shapiro, 1994)(Cheng and Bouman, 2001)(Feng et al., 2002)(Wilson and Li, 2003)(Kumar and Hebert, 2003c).

There are several advantages of causal trees. These models can encode long range interactions in images explicitly through the latent hierarchical process. Also, the causal trees contain no cycles and hence allow the use of very efficient techniques for exact parameter learning and inference. In spite of these advantages, there are several problems associated with these models. The main problem with the tree-structured models is that they suffer from the nonstationarity of the induced random field, leading to 'blocky' smoothing of the image labels (Feng et al., 2002). According to this, there exists an imposed difference in the behavior of interactions between neighboring nodes at different locations in the image, purely dictated by the tree structure even though there is no a-priori reason for such a difference. This problem exists in all the tree-structured models whether causal or noncausal. One way to solve this problem is to dynamically adapt the tree-structure to a given input image. This idea was explored in *dynamic trees* (Williams and Adams, 1999) but the inference over tree-structure still remains an intractable problem. Other possibility is to use more complex causal structures instead of simple trees. But this makes the problem of parameter learning and inference hard.

The second important drawback is that, when trained discriminatively, the causal models sometimes suffer from the *label bias* problem which unfairly favors

[1] The directed acyclic graphs are popularly known as *Bayesian Networks.*

labels with fewer successors due to the need of normalizing each link to be a proper transition probability (Bottou, 1991)(Lafferty et al., 2001). On the other hand, in noncausal models this problem does not arise as one needs to define potential functions for each clique (which are not required to sum to one) and there is a universal normalizing constant for the whole distribution known as the partition function. Finally, since causal models are developed in a generative framework, some crude approximations are required to make data generative model computationally tractable while retaining some expressive power of the model. This problem exists with conventional noncausal models also and we will discuss this in detail in Section 2.2. To avoid the problems associated with the causal models, in this paper we will focus on noncausal or undirected graphical models.

2.2. Noncausal Models

Noncausal models are global probability distributions defined on undirected graphs using local clique potentials, i.e.,

$$P(\boldsymbol{x}) \propto \prod_{c \in \mathcal{C}} \psi_c(x_c),$$

where $\mathcal{C}$ is the set of all the cliques[2] in the graph, and $\psi_c(x_c)$ are clique potentials, i.e. positive functions of clique variables x_c. Noncausal graphs are more suited to handle interactions over image lattices since usually there exists no natural causal relationships among image components. Even though computationally tractable, the tree-structured noncausal models suffer from similar problems as the causal trees described in Section 2.1 except the label-bias problem. So, in the following discussion we will explore arbitrary undirected graphs with cycles. Undirected graphical models are sometimes popularly referred to as *random fields* in computer vision, statistical physics and several other areas.

Markov Random Fields (MRFs) are the most popular undirected graphical models in vision, which allow one to incorporate local contextual constraints in labeling problems in a principled manner (Li, 2001). MRFs were made popular in vision by early work of Geman and Geman (Geman and Geman, 1984), and Besag (Besag, 1986). MRFs are generally used in a probabilistic generative framework that models the joint probability of the observed data and the corresponding labels. In other words, let $\boldsymbol{y}$ be the observed data from an input image, where $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i \in S}$, $\boldsymbol{y}_i$ is the data from the i^{th} site, and S is the set of sites. Let the corresponding labels at the image sites be given by $\boldsymbol{x} = \{x_i\}_{i \in S}$. In the MRF framework, the posterior over the labels given the data is expressed using the Bayes' rule as,

$$P(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{x}, \boldsymbol{y}) = P(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$$

[2] A clique is a fully connected subgraph of the original graph

where the prior over labels, $P(\boldsymbol{x})$ is modeled as a MRF. For computational tractability, the observation or likelihood model, $p(\boldsymbol{y}|\boldsymbol{x})$ is assumed to have a factorized form, i.e. $p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i \in S} p(\boldsymbol{y}_i|x_i)$ (Besag, 1986)(Li, 2001)(Feng et al., 2002)(Xiao et al., 2002). However, as noted by several researchers (Bouman and Shapiro, 1994)(Pieczynski and Tebbache, 2000)(Wilson and Li, 2003)(Kumar and Hebert, 2003c), this assumption is too restrictive for several natural image analysis applications. For example, consider a class that contains man-made structures (e.g. buildings). The data belonging to such a class is highly dependent on its neighbors. This is because, in man-made structures, the lines or edges at spatially adjoining sites follow some underlying organization rules rather than being random (See Figure 1(b)). This is also true for a large number of texture classes that are made of structured patterns and other object detection applications where geometric (and possibly appearance) relationships between different parts of the object are crucial for its detection in cluttered scenes (Weber et al., 2000)(Fergus et al., 2003)(Felzenszwalb and Huttenlocher, 2000).

Some efforts have been made in the past to model the dependencies in the observed image data. In (Kittler and Pairman, 1985), a technique was presented that assumes the noise in the data at neighboring sites to be correlated, which is modeled using an auto-normal model. However, the authors do not specify a field over the labels, and classify a site by maximizing the local posterior over labels given the data and the neighborhood labels. In the context of hierarchical texture segmentation, Won and Derin (Won and Derin, 1992) model the local joint distribution of the data contained in the neighborhood of a site assuming all the neighbors from the same class. They further approximate the overall likelihood to be factored over the local joint distributions. Wilson and Li (Wilson and Li, 2003) assume the difference between observations from the neighboring sites to be conditionally independent given the label field. In the context of multiscale random field, Cheng and Bouman (Cheng and Bouman, 2001) make a more general assumption. They assume the difference between the data at a given site and the linear combination of the data from that site's parents to be conditionally independent given the label at the current scale.

All the above techniques make simplifying assumptions to get some sort of factored approximation of the likelihood for tractability. This precludes capturing stronger relationships in the observations in the form of arbitrarily complex features that might be desired to discriminate between different classes. A novel pairwise MRF model is suggested in (Pieczynski and Tebbache, 2000) to avoid the problem of explicit modeling of the likelihood, $p(\boldsymbol{y}|\boldsymbol{x})$. They model the joint $p(\boldsymbol{x}, \boldsymbol{y})$ as a MRF in which the label field $P(\boldsymbol{x})$ is not necessarily a MRF. But this shifts the problem to the modeling of pairs $(\boldsymbol{x}, \boldsymbol{y})$. The authors model the pair by assuming the observations to be the true underlying binary field corrupted by correlated noise. However, for most of the real-world applications, this assumption is too simplistic. In our previous work (Kumar and Hebert, 2003c), we modeled the data dependencies using a pseudolikelihood approximation of a conditional MRF for computational tractability. In this paper, we explore alternative ways of modeling data dependencies which allow elimination of these approximations in a principled manner.

Another thing to note is that the interaction among labels in MRFs is modeled by the term $P(\boldsymbol{x})$, which is seen as a prior in the Bayesian view. The main drawback of this view is that the label interactions do not depend on the observed data $\boldsymbol{y}$. This prohibits one from modeling data-dependent interactions in labels that are necessary for a variety of tasks. For example, while implementing local smoothness of labels in image segmentation, it may be desirable to use observed data to modulate the smoothness according to the image intensity gradients (Boykov and Jolly, 2001)(Blake et al., 2004). Further, in parts based object detection, to model interactions among object parts, we need observed data to enforce geometric (and possibly photometric) constraints. This is also the case for modeling higher level interactions between objects or regions in an image. In this paper, we present models which allow interactions among labels based on unrestricted use of observations as necessary. This step is crucial to develop models that can incorporate interactions of different types within the same framework.

In related work, taking the non-probabilistic view of energy-based graphical model, Boykov and Jolly (Boykov and Jolly, 2001) have proposed an energy form that uses observed data to model pairwise interaction between labels. In this work, the smoothness parameter of the Ising model was modulated by a Gaussian over the intensity difference between a pair of pixels. Using such contrast-sensitive interactions, they have shown interesting results in the area of interactive image segmentation. However, their approach has two main drawbacks. Firstly, there is no direct probabilistic interpretation of their energy. As the authors themselves state, the choice of modulating the smoothing parameter by using image observations is rather 'ad-hoc'. Being non-probabilistic, parameter learning is hard in these models. The authors tune the smoothing and the modulation parameters by hand. Secondly, the non-probabilistic energy-based view eliminates the possibility of computing labels that are optimal for minimizing sitewise errors (i.e., sitewise zero-one loss function). This is due to the fact that there is no concept of marginals in energy-based view, which are required for minimizing the sitewise errors.

Recently, Blake et al. (Blake et al., 2004) have given a probabilistic interpretation of the contrast-sensitive image segmentation formulation suggested by Boykov and Jolly (Boykov and Jolly, 2001). They have proposed to learn the observation model parameters along with the modulation parameters using pseudo-likelihood. This alleviates one of the main problems with the original non-probabilistic formulation of (Boykov and Jolly, 2001). However, the parameters of the foreground and background models are learned separately. This forces one to use 'post-hoc' averaging schemes to estimate the modulation and the interaction parameters. Another limitation of this approach is that the interactions among observed data are restricted to site pairs. On the contrary, the model presented in this paper allows arbitrary interactions among data from multiple sites, potentially from the whole image, without any added computational complexity.

In MRF formulations of binary classification problems, the label interaction field $P(\boldsymbol{x})$ is commonly assumed to be a homogeneous and isotropic Ising model (or Potts model for multiclass labeling problems) with only pairwise nonzero potentials. If the data likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ is approximated by assuming that the observed data is conditionally independent given the labels, the posterior distribution[3] over labels can be written as,

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z_m} \exp\left(\sum_{i \in S} \log p(\boldsymbol{s}_i(\boldsymbol{y_i})|x_i) + \sum_{i \in S} \sum_{j \in \mathcal{N}_i} \beta_m x_i x_j \right), \tag{1}$$

where β_m is the interaction parameter of the MRF, and $\boldsymbol{s}_i(\boldsymbol{y_i})$ is a *single-site* feature vector, which uses data only from a single site i, i.e., $\boldsymbol{y}_i$. Note that even though only the label prior, $P(x)$ was assumed to be a MRF, the assumption of conditional independence of data implies that the posterior given in (1) is also a MRF. This allows one to reap the benefits of readily available tools of inference over a MRF. If the conditional independence assumption is not used, the posterior will usually not be a MRF making the inference difficult.

Now, if we turn our attention again toward the original aim of this work, we are interested in classification of image sites. For classification purposes, we want to estimate the posterior over labels given the observations, i.e., $P(\boldsymbol{x}|\boldsymbol{y})$. In a generative framework, one expends efforts to model the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$, which involves implicit modeling of the observations. In a discriminative framework, one models the distribution $P(\boldsymbol{x}|\boldsymbol{y})$ directly. This saves one from making simplistic assumptions about the data. This view forms the core theme of the model we present in this paper as discussed in the following sections.

As noted in (Feng et al., 2002), a potential advantage of using the discriminative approach is that the true underlying generative model may be quite complex even though the class posterior is simple. This means that the generative approach may spend a lot of resources on modeling the generative models which are not particularly relevant to the task of inferring the class labels. Moreover, learning the class density models may become even harder when the training data is limited (Rubinstein and Hastie, 1997). A more complete comparison between the discriminative and the generative models for the linear family of classifiers has been presented in (Rubinstein and Hastie, 1997)(Ng and Jordan, 2002).

[3] With a slight abuse of notation, we will use the term 'MRF model' to indicate this posterior in the rest of the paper.

3. Discriminative Random Field (DRF)

In this paper, we present a noncausal model called Discriminative Random Field [4] based on the concept of Conditional Random Field (CRF) proposed by Lafferty et al. (Lafferty et al., 2001) in the context of segmentation and labeling of 1-D text sequences. The CRFs directly model the posterior distribution $P(\boldsymbol{x}|\boldsymbol{y})$ as a Gibbs field. This approach allows one to capture arbitrary dependencies among the observations without resorting to any model approximations. CRFs have been shown to outperform the traditional Hidden Markov Model based labeling of text sequences (Lafferty et al., 2001). Our model further enhances the 1-D CRFs by proposing the use of local discriminative models to capture the class associations at individual sites as well as the interactions on the neighboring sites on 2-D regular as well as irregular lattices. The proposed DRF model permits interactions in both the observed data and the labels in a principled manner.

We first restate in our notations the definition of CRFs as given by Lafferty et al. (Lafferty et al., 2001). Let the observed data from an input image be given by $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i \in S}$ where $\boldsymbol{y}_i$ is the data from i^{th} site and $\boldsymbol{y}_i \in \Re^c$. The corresponding labels at the image sites are given by $\boldsymbol{x} = \{x_i\}_{i \in S}$. In this work we will be concerned with binary classification, i.e. $x_i \in \{-1, 1\}$. The random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly distributed, but in a discriminative framework, a conditional model $P(\boldsymbol{x}|\boldsymbol{y})$ is constructed from the observations and labels, and the marginal $p(\boldsymbol{y})$ is not modeled explicitly.

CRF Definition: *Let $G = (S, E)$ be a graph such that $\boldsymbol{x}$ is indexed by the vertices of G. Then $(\boldsymbol{x}, \boldsymbol{y})$ is said to be a conditional random field if, when conditioned on $\boldsymbol{y}$, the random variables x_i obey the Markov property with respect to the graph: $P(x_i|\boldsymbol{y}, \boldsymbol{x}_{S-\{i\}}) = P(x_i|\boldsymbol{y}, \boldsymbol{x}_{\mathcal{N}_i})$, where $S - \{i\}$ is the set of all nodes in the graph except the node i, $\mathcal{N}_i$ is the set of neighbors of the node i in G, and $\boldsymbol{x}_\Omega$ represents the set of labels at the nodes in set Ω.*

Thus, a CRF is a random field globally conditioned on the observations $\boldsymbol{y}$. The condition of positivity requiring $P(\boldsymbol{x}|\boldsymbol{y}) > 0 \ \forall \ \boldsymbol{x}$ has been assumed implicitly. Now, using the Hammersley-Clifford theorem (Hammersley and Clifford,) and assuming only up to pairwise clique potentials to be nonzero, the joint distribution over the labels $\boldsymbol{x}$ given the observations $\boldsymbol{y}$ can be written as,

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z} \exp\left(\sum_{i \in S} A_i(x_i, \boldsymbol{y}) + \sum_{i \in S} \sum_{j \in \mathcal{N}_i} I_{ij}(x_i, x_j, \boldsymbol{y}) \right) \qquad (2)$$

where Z is a normalizing constant known as the partition function, and -A_i and -I_{ij} are the unary and pairwise potentials respectively. With a slight abuse of

[4] An earlier version of this work appeared in International Conference on Computer Vision (ICCV 03)(Kumar and Hebert, 2003b).

notations, in the rest of the paper we will call A_i the *association potential* and I_{ij} the *interaction potential.*

There are two main differences between the conditional model given in (2) and the original MRF framework given in (1). First, in the conditional fields, association potential at any site is a function of all the observations $\boldsymbol{y}$ while in MRFs (with the assumption of conditional independence of the data), the association potential is a function of data only at that site, $\boldsymbol{y}_i$. Second, the interaction potential for each pair of nodes in MRFs is a function of only labels, while in the conditional models it is a function of labels as well as all the observations $\boldsymbol{y}$. As will be shown later, these differences play crucial roles in modeling arbitrary interactions in natural images in a principled manner.

The DRF model we present in this paper is a specific type of CRF defined in (2), and thus inherits all its advantages over the traditional MRF as described above. In the DRF model, we extend the specific 1-D sequential CRF form proposed in (Lafferty et al., 2001). There are two main extensions: First, the unary and pairwise potentials in DRFs are designed using arbitrary local discriminative classifiers. This allows one to use domain-specific discriminative classifiers for structured data rather than restricting the potentials to a specific form. Taking a similar view, several researchers have recently demonstrated good results using different classifiers such as probit (Qi et al., 2005), boosting (Torralba et al., 2005) and even neural network (He et al., 2004). This view is consistent with one of the key motivations behind this work in which we wanted to develop models that allow one to leverage the power of discriminative classifiers in problems where data has interactions rather than being independent.

Second, instead of being 1-D sequential models, the DRFs are defined over 2-D image lattices which generally induce graphs with loops. This makes the problem of parameter learning and inference significantly harder. To the best of our knowledge, ours is the first work that introduced CRF-based models in computer vision for image analysis. Recently, a number of researchers have demonstrated the utility of such models in various computer vision applications (Murphy et al., 2003)(He et al., 2004)(Quattoni et al., 2004)(Weinman et al., 2004)(Szummer and Qi, 2004)(Torralba et al., 2005)(Qi et al., 2005)(Wang and Ji, 2005). The discriminative fields make it possible to tie different computer vision applications in a single framework in a seamless fashion. In the rest of this paper we assume the random field given in (2) to be homogeneous and isotropic, i.e. the functional forms of A_i and I_{ij} are independent of the locations i and j. Henceforth we will leave the subscripts and simply use the notations A and I. Note that the assumption of isotropy can be easily relaxed at the cost of a few additional parameters.

3.1. Association Potential

In the DRF framework, the association potential, $A(x_i, \boldsymbol{y})$, can be seen as a measure of how likely a site i will take label x_i given image $\boldsymbol{y}$, ignoring the effects of other sites in the image. Suppose, $\boldsymbol{f}(.)$ is a function that maps an arbitrary patch in an image to a feature vector such that $\boldsymbol{f} : \mathcal{Y}_p \rightarrow \Re^l$. Here $\mathcal{Y}_p$ is the set of all possible patches in all possible images. Let $\omega_i(\boldsymbol{y})$ be an arbitrary patch in the neighborhood of site i in image $\boldsymbol{y}$ from which we want to extract a feature vector $\boldsymbol{f}(\omega_i(\boldsymbol{y}))$. Note that the neighborhood used for the patch $\omega_i(\boldsymbol{y})$ need not be the same as the label neighborhood $\mathcal{N}_i$. Indeed, $\omega_i(\boldsymbol{y})$ can potentially be the whole image itself. For clarity, with slight abuse of notation, we will denote the feature vector $\boldsymbol{f}(\omega_i(\boldsymbol{y}))$ at each site i by $\boldsymbol{f}_i(\boldsymbol{y})$. The subscript i indicates the difference just in the feature vectors at different sites, *not* in the functional form of $\boldsymbol{f}(.)$. Then, $A(x_i, \boldsymbol{y})$ is modeled using a local discriminative model that outputs the association of the site i with class x_i as,

$$A(x_i, \boldsymbol{y}) = \log P'(x_i | \boldsymbol{f}_i(\boldsymbol{y})), \tag{3}$$

where $P'(x_i | \boldsymbol{f}_i(\boldsymbol{y}))$ is the local class conditional at site i. This form allows one to use an arbitrary domain-specific probabilistic discriminative classifier for a given task. This can be seen as a parallel to the traditional MRF models where one can use arbitrary local generative classifier to model the unary potential. One possible choice of $P'(.)$ can be Generalized Linear Models (GLM), which are used extensively in statistics to model the class posteriors given the observations (McCullagh and Nelder, 1987). In this work we used the logistic function[5] as a *link* in the GLM. Thus, the local class conditional can be written as,

$$P'(x_i{=}1 | \boldsymbol{f}_i(\boldsymbol{y})) {=} \frac{1}{1{+}e^{-(w_0 + \boldsymbol{w}_1^T \boldsymbol{f}_i(\boldsymbol{y}))}} {=} \sigma(w_0 {+} \boldsymbol{w}_1^T \boldsymbol{f}_i(\boldsymbol{y})), \tag{4}$$

where $\boldsymbol{w} = \{w_0, \boldsymbol{w}_1\}$ are the model parameters. This form of $P'(.)$ will yield a linear decision boundary in the feature space spanned by vectors $\boldsymbol{f}_i(\boldsymbol{y})$. To extend the logistic model to induce a nonlinear decision boundary, a transformed feature vector at each site i is defined as $\boldsymbol{h}_i(\boldsymbol{y}) = [1, \phi_1(\boldsymbol{f}_i(\boldsymbol{y})), \ldots, \phi_R(\boldsymbol{f}_i(\boldsymbol{y}))]^T$ where $\phi_k(.)$ are arbitrary nonlinear functions. These functions can be seen as kernel mapping of the original feature vector into a high dimensional space. The first element of the transformed vector is kept as 1 to accommodate the bias parameter w_0. Further, since $x_i \in \{-1, 1\}$, the probability in (4) can be compactly expressed as,

$$P'(x_i | \boldsymbol{y}) = \sigma(x_i \boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y})). \tag{5}$$

Finally, for this choice of $P'(.)$, the association potential can be written as,

$$A(x_i, \boldsymbol{y}) = \log(\sigma(x_i \boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y}))) \tag{6}$$

This transformation ensures that the DRF is equivalent to a logistic classifier if the interaction potential in (2) is set to zero. Note that the use of logistic function

[5] One can use other choices of link such as *probit* link.

to model the discriminative classifier yields $A(.)$ that is linear in features. This is similar to the original form of the 1-D sequential CRFs of (Lafferty et al., 2001) with the difference that we use kernels to define this potential. Parallel to our work, researchers have proposed the use of kernels in CRF-type of models (Taskar et al., 2003)(Lafferty et al., 2004). Moreover, while designing graph potentials, recently other researchers have explored the use of different classifiers such as probit classifier (Qi et al., 2005)(Szummer and Qi, 2004) which will not yield a linear form of $A_i(.)$. Similarly, in Boosted Random Fields (BRFs) proposed by Torralba et at. (Torralba et al., 2005), the authors design unary potential using boosting. They show good results on the application of contextual object detection using BRFs.

Note that in (6), the transformed feature vector at *each* site i i.e., $\boldsymbol{h}_i(\boldsymbol{y})$ is a function of the whole set of observations $\boldsymbol{y}$. This allows one to pool arbitrarily complex dependencies in the observed data for the purpose of classification. On the contrary, the assumption of conditional independence of the data in the traditional MRF framework allows one to use the data only from a particular site, i.e., $\boldsymbol{y}_i$ to get the log-likelihood, which acts as the association potential as shown in (1).

In related work, a neural network based discriminative classifier was used by Feng et al. (Feng et al., 2002) to model the observations in a generative tree-structured belief network model. Since the model required generative data likelihood, the discriminative output of the neural network was used to approximate the actual likelihood of the data in an ad-hoc fashion. On the contrary, in the DRF model, the discriminative class posterior is an integral part of the full conditional model in (2), and all the models parameters are learned simultaneously.

3.2. Interaction Potential

In the DRF framework, the interaction potential can be seen as a measure of how the labels at neighboring sites i and j should interact given the observed image y. To model the interaction potential, I, we first analyze the form commonly used in the MRF framework. For the isotropic, homogeneous Ising model, the interaction potential is given as $I = \beta x_i x_j$, which penalizes every dissimilar pair of labels by the cost β (Ising, 1925). This form of interaction favors piecewise constant smoothing of the labels without considering the discontinuities in the observed data explicitly. Geman and Geman (Geman and Geman, 1984) have proposed a line-process model which allows discontinuities in the labels through piecewise continuous smoothing. Other discontinuity models have also been proposed for adaptive smoothing (Li, 2001), but all of them require the labels to be either continuous or ordered. On the contrary, in the classification task there is no natural ordering in the labels. Also, these discontinuity adaptive models do not use the observed data to model the discontinuities.

In contrast, in the DRF formulation, the interaction potential is a function of all the observations $\boldsymbol{y}$. We propose to model I in DRFs using a data-dependent term along with the constant smoothing term of the Ising model. In addition to modeling arbitrary pairwise relational information between sites, the data-dependent smoothing can compensate for the errors in modeling the association potential. To model the data-dependent term, the aim is to have similar labels at a pair of sites for which the observed data supports such a hypothesis. In other words, we are interested in learning a pairwise discriminative model.

Suppose, $\boldsymbol{\psi}(.)$ is a function that maps an arbitrary patch in an image to a feature vector such that $\boldsymbol{\psi} : \mathcal{Y}_p \rightarrow \Re^{\gamma}$. Let $\Omega_i(\boldsymbol{y})$ be an arbitrary patch in the neighborhood of site i in image $\boldsymbol{y}$ from which we want to extract a feature vector $\boldsymbol{\psi}(\Omega_i(\boldsymbol{y}))$. Note that the neighborhood used for the patch $\Omega_i(\boldsymbol{y})$ need not be the same as the label neighborhood $\mathcal{N}_i$. For clarity, with slight abuse of notation, we will denote the feature vector $\boldsymbol{\psi}(\Omega_i(\boldsymbol{y}))$ at each site i by $\boldsymbol{\psi}_i(\boldsymbol{y})$. Similarly, we define a feature vector $\boldsymbol{\psi}_j(\boldsymbol{y})$ for site j. Again, to emphasize, the subscripts i and j indicate the difference just in the feature vectors at different sites, *not* in the functional form of $\boldsymbol{\psi}(.)$. Given the features at two different sites, we want to learn a pairwise discriminative model $P''(x_i = x_j | \boldsymbol{\psi}_i(\boldsymbol{y}), \boldsymbol{\psi}_j(\boldsymbol{y}))$. Note that by choosing the function $\boldsymbol{\psi}_i$ to be different from $\boldsymbol{f}_i$, used in (4), information different from $\boldsymbol{f}_i$ can be used to model the relations between pairs of sites.

Let t_{ij} be an auxiliary variable defined as

$$t_{ij} = x_i x_j,$$

and let $\boldsymbol{\mu}(\boldsymbol{\psi}_i(\boldsymbol{y}), \boldsymbol{\psi}_j(\boldsymbol{y}))$ be a new feature vector such that $\boldsymbol{\mu} : \Re^{\gamma} \times \Re^{\gamma} \rightarrow \Re^{q}$. Denoting this feature vector as $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$ for simplification, we model the pairwise discriminatory term similar to the one defined in (5) as,

$$P''(t_{ij} | \boldsymbol{\psi}_i(\boldsymbol{y}), \boldsymbol{\psi}_j(\boldsymbol{y})) = \sigma(t_{ij} \boldsymbol{v}^T \boldsymbol{\mu}_{ij}(\boldsymbol{y})), \tag{7}$$

where $\boldsymbol{v}$ are the model parameters. Note that the first component of $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$ is fixed to be 1 to accommodate the bias parameter. Now, the interaction potential in DRFs is modeled as a convex combination of two terms, i.e.

$$I(x_i, x_j, \boldsymbol{y}) = \beta \left(K x_i x_j + (1 - K)(2\sigma(t_{ij} \boldsymbol{v}^T \boldsymbol{\mu}_{ij}(\boldsymbol{y})) - 1) \right) \tag{8}$$

where $0 \leq K \leq 1$. The first term is a data-independent smoothing term, similar to the Ising model. The second term is a $[-1, 1]$ mapping of the pairwise logistic function defined in (7). This mapping ensures that both terms have the same range. Ideally, the data-dependent term will act as a discontinuity adaptive model that will moderate smoothing when the data from two sites is 'different'. The parameter K gives flexibility to the model by allowing the learning algorithm to adjust the relative contributions of these two terms according to the training data. Finally, β is the interaction coefficient that controls the degree of smoothing. Large values of β encourage smoother solutions. Note that even though the model seems to have some resemblance to the line process suggested in (Geman and Geman, 1984), K in (8) is a global weighting parameter unlike the line process where a discrete parameter is introduced for each pair of sites to facilitate discontinuities in smoothing. Anisotropy can be easily included in the DRF model by parameterizing the interaction potentials of different directional pairwise cliques with different sets of parameters $\{\beta, K, \boldsymbol{v}\}$.

To summarize the roles of the two potentials in DRFs, the association potential acts as a complex nonlinear classifier for individual sites, while the interaction potential can be seen as a data-dependent discriminative label interaction.

3.3. Parameter Estimation

Let θ be the set of parameters of the DRF model where $\theta = \{\boldsymbol{w}, \boldsymbol{v}, \beta, K\}$. The form of the DRF model resembles the posterior for the MRF framework given in (1). However, in the MRF framework, the parameters of the class generative models, $p(\boldsymbol{s}(\boldsymbol{y}_i)|x_i)$ and the parameters of the prior random field on labels, $P(\boldsymbol{x})$ are generally assumed to be independent and are learned separately (Li, 2001). In contrast, we make no such assumption and learn all the parameters of the DRF model simultaneously. Nevertheless, the similarity of the forms allows for most of the techniques used for learning the MRF parameters to be utilized for learning the DRF parameters with a few modifications.

We take the standard maximum-likelihood approach to learn the DRF parameters which, similar to the conventional MRF learning, involves the evaluation of the partition function Z. The evaluation of Z is, in general, a NP-hard problem. One could use either sampling techniques or resort to some approximations e.g. mean-field or pseudolikelihood to estimate the parameters (Li, 2001). As a preliminary choice, we used the pseudolikelihood formulation due to its simplicity. According to this,

$$\widehat{\theta}^{ML} \approx \arg\max_{\theta} \prod_{m=1}^{M} \prod_{i \in S} P(x_i^m | \boldsymbol{x}_{\mathcal{N}_i}^m, \boldsymbol{y}^m, \theta) \tag{9}$$

$$\mathit{Subject\ \ to}\ \ 0 \leq K \leq 1$$

where m indexes over the training images and M is the total number of training images, and

$$P(x_i | \boldsymbol{x}_{\mathcal{N}_i}, \boldsymbol{y}, \theta) = \frac{1}{z_i} \exp\{A(x_i, \boldsymbol{y}) + \sum_{j \in \mathcal{N}_i} I(x_i, x_j, \boldsymbol{y})\},$$

$$z_i = \sum_{x_i \in \{-1,1\}} \exp\left(A(x_i, \boldsymbol{y}) + \sum_{j \in \mathcal{N}_i} I(x_i, x_j, \boldsymbol{y}) \right)$$

The pseudolikelihood given in (9) can be maximized by using line search methods for constrained maximization with bounds (Gill et al., 1981). Since the pseudolikelihood in (9) is not a convex function of the parameters, good initialization of the parameters is important to avoid bad local maxima. To initialize the parameters $\boldsymbol{w}$ in $A(x_i, \boldsymbol{y})$, we first learn these parameters using standard maximum likelihood logistic regression assuming all the labels x_i^m to be independent given the data $\boldsymbol{y^m}$ for each image m (Minka, 2001). Using (5), the log-likelihood can be expressed as,

$$L(\boldsymbol{w}) = \sum_{m=1}^{M} \sum_{i \in S} \log(\sigma(x_i^m \boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y^m}))) \tag{10}$$

The Hessian of the log-likelihood is given as,

$$\nabla^2_{\boldsymbol{w}} L(\boldsymbol{w}) = -\sum_{m=1}^{M} \sum_{i \in S} \left\{ \sigma(\boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y^m}))(1 - \sigma(\boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y^m}))) \right\} \boldsymbol{h}_i(\boldsymbol{y^m}) \boldsymbol{h}_i^T(\boldsymbol{y^m})$$

Note that the Hessian does not depend on how the data is labeled and is non-positive definite. Hence the log-likelihood in (10) is convex (convex downward or concave), and any local maximum is the global maximum. Newton's method was used for maximization which has been shown to be much faster than other techniques for correlated features (Minka, 2001). The initial estimates of the parameters $\boldsymbol{v}$ in data-dependent term in $I(x_i, x_j, \boldsymbol{y})$ were also obtained similarly.

3.4. Inference

Given a new test image $\boldsymbol{y}$, our aim is to find the optimal label configuration $\boldsymbol{x}$ over the image sites where optimality is defined with respect to a cost function. Maximum A Posteriori (MAP) solution is a widely used estimate that is optimal with respect to the zero-one cost function defined as $C(\boldsymbol{x}, \boldsymbol{x}^*) = 1 - \delta(\boldsymbol{x} - \boldsymbol{x}^*)$, where $\boldsymbol{x}^*$ is the true label configuration, and $\delta(\boldsymbol{x} - \boldsymbol{x}^*)$ is 1 if $\boldsymbol{x} = \boldsymbol{x}^*$, and 0 otherwise. For binary classifications, MAP estimate can be computed exactly for an undirected graph using the max-flow/min-cut type of algorithms if the probability distribution meets certain conditions (Greig et al., 1989)(Kolmogorov and Zabih, 2002). For the DRF model, exact MAP solution can be computed if $K \geq 0.5$ and $\beta \geq 0$. However, in the context of MRFs, the MAP solution has been shown to perform poorly for the Ising model when the interaction parameter, β takes large values (Greig et al., 1989)(Fox and Nicholls, 2000). Our results in Section 3.5.3 corroborate this observation for the DRFs too.

An alternative to the MAP solution is the Maximum Posterior Marginal (MPM) solution for which the cost function is defined as $C(\boldsymbol{x}, \boldsymbol{x}^*) = \sum_{i \in S}(1 - \delta(x_i - x_i^*))$, where x_i^* is the true label at the i^{th} site. The MPM computation requires marginalization over a large number of variables which is generally NP-hard. One can use either sampling procedures (Fox and Nicholls, 2000) or use Belief Propagation to obtain an estimate of the MPM solution. On the other hand, one can obtain local MAP estimates using the algorithm Iterated Conditional Modes (ICM), proposed by Besag (Besag, 1986) which is equivalent to zero-temperature simulated annealing. Given an initial label configuration, ICM maximizes the local conditional probabilities iteratively, i.e.

$$x_i \leftarrow \arg\max_{x_i} P(x_i | \boldsymbol{x}_{\mathcal{N}_i}, \boldsymbol{y})$$

ICM yields local maximum of the posterior and has been shown to give reasonably good results even when exact MAP performs poorly for large values of β (Greig et al., 1989)(Fox and Nicholls, 2000). In the first set of experiments, we primarily used ICM to get local MAP estimates of the labels and also compared the ICM results with the MAP results (Section 3.5.3). In our ICM implementation, the image sites were divided into coding sets to speed up the sequential updating procedure (Besag, 1986). A coding sets is a set of image pixels such that each pixel in this set is a non-neighbor of any other pixel in the set. This type of update provides a useful compromise between the synchronous and the asynchronous schemes (Besag, 1986).

3.5. Man-made Structure Detection Task

The proposed DRF model was applied to the task of detecting man-made structures in natural scenes. Automatic detection of man-made structures in ground-level images is useful for scene understanding, robotic navigation, surveillance, image indexing and retrieval etc. This section focuses on the detection of man-made structures, which can be characterized primarily by the presence of linear structures. A detailed account of the main issues related to this application and comparison with other approaches is given in (Kumar and Hebert, 2003c).

The training and the test set contained 108 and 129 images respectively, each of size 256×384 pixels, from the Corel image database. Each image was divided in nonoverlapping 16×16 pixels blocks, and we call each such block an image site. The ground truth was generated by hand-labeling every site in each image as a *structured* or *nonstructured* block. The whole training set contained $36,269$ blocks from the *nonstructured* class, and $3,004$ blocks from the *structured* class. The detailed explanation of the features used for the structure detection application is given in our previous work (Kumar and Hebert, 2003c). Here we briefly describe the features to set the notations.

For each image we compute two different types of feature vectors at each site. Using the same notations as introduced in Section 3, first a *single-site* feature vector at the site i, $\boldsymbol{s}_i(\boldsymbol{y_i})$ is computed using the histogram from the data $\boldsymbol{y}_i$ at that site (i.e., 16×16 block) such that $\boldsymbol{s}_i : \boldsymbol{y_i} \rightarrow \Re^d$. Obviously, this vector does not take into account influence of the data in the neighborhood of that site. Next, a *multiscale* feature vector at the site i, $\boldsymbol{f}_i(\boldsymbol{y})$ is computed which explicitly takes into account the dependencies in the data contained in the neighboring sites. It should be noted that the neighborhood for the data interaction need not be the same as for the label interaction. In the following section we describe the details of feature extraction at each image site.

3.5.1. *Feature Set Description*

In this section, first we describe *multiscale* feature vector that captures the general statistical properties of the man-made structures by combining observed data at multiple adjoining sites. For each site in the image, we compute the features at different scales, which capture intrascale as well as interscale dependencies. The multiscale feature vector at site i is then given as: $\boldsymbol{f}_i = \big[\{f_i^\lambda\}_{\lambda=1}^{\Lambda}, \{f_i^\rho\}_{\rho=1}^{R}\big]$, where f_i^λ is λ^{th} intrascale feature and f_i^ρ is ρ^{th} interscale feature.

A. Intrascale Features

As mentioned earlier, here we focus on those man-made structures which are primarily characterized by straight lines and edges. To capture these characteristics, at first, the input image is convolved with the derivative of Gaussian filters to yield the gradient magnitude and orientation at each pixel. Then, for an image site i, the gradients contained in a window W_c at scale c ($c\!=\!1,\ldots,C$) are combined to yield a histogram over gradient orientations. However, instead of incrementing the counts in the histogram, we weight each count by the gradient magnitude at that pixel as in (Barrett and Petersen, 2001). It should be noted that the weighted histogram is made using the raw gradient information at *every* pixel in W_c without any thresholding. Let E_δ be the magnitude of the histogram at the δ^{th} bin, and Δ be the total number of bins in the histogram. To alleviate the problem of hard binning of the data, we smoothed the histogram using kernel smoothing. The smoothed histogram is given as,

$$E'_\delta = \frac{\sum_{i=1}^{\Delta} K((\delta - i)/h)E_i}{\sum_{i=1}^{\Delta} K((\delta - i)/h)} \tag{11}$$

where K is a kernel function with bandwidth h. The kernel K is generally chosen to be a non-negative, symmetric function.

If the window W_c contains a smooth patch, the gradients will be very small and the mean magnitude of the histogram over all the bins will also be small. On the other hand, if W_c contains a textured region, the histogram will have approximately uniformly distributed bin magnitudes. Finally, if W_c contains a few straight lines and/or edges embedded in smooth background, as is the case for the *structured* class, a few bins will have significant peaks in the histogram in comparison to the other bins. Let ν_0 be the mean magnitude of the histogram such that $\nu_0 = \frac{1}{\Delta}\sum_{\delta=1}^{\Delta} E'_\delta$. We aim to capture the average *spikeness*, of the smoothed histogram as an indicator of the *structuredness* of the patch. For this, we propose heaved central-shift moments for which p^{th} order moment ν_p is given as,

$$\nu_p = \frac{\sum_{\delta=1}^{\Delta}(E'_\delta - \nu_0)^{p+1} H(E'_\delta - \nu_0)}{\sum_{\delta=1}^{\Delta}(E'_\delta - \nu_0) H(E'_\delta - \nu_0)} \tag{12}$$

where $H(x)$ is the unit step function such that $H(x) = 1$ for $x > 0$, and 0, otherwise. The moment computation in (12) considers the contribution only from the bins having magnitude above the mean ν_0. Further, each bin value above the mean is linearly weighted by its distance from the mean so that the peaks far away from the mean contribute more. The moments ν_0 and ν_p at each scale c form the gradient magnitude based intrascale features in the multiscale feature vector.

Since the lines and edges belonging to the *structured* regions generally either exhibit parallelism or combine to yield different junctions, the relation between the peaks of the histograms must contain useful information. The peaks of the histogram are obtained simply by finding the local maxima of the smoothed histogram. Let δ_1 and δ_2 be the ordered orientations corresponding to the two highest peaks such that $E'_{\delta_1} \geq E'_{\delta_2}$. Then, the orientation based intrascale feature β^c for each scale c is computed as $\beta^c = |\sin(\delta_1 - \delta_2)|$. This measure favors the presence of near right-angle junctions. The sinusoidal nonlinearity was preferred to the Gaussian function because sinusoids have much slower fall-off rate from the mean. The sinusoids have been used earlier in the context of perceptual grouping of prespecified image primitives (Krishnamachari and Chellappa, 1996). We used only the first two peaks in the current work but one can compute more such features using the remaining peaks of the histogram. In addition to the relative locations of the peaks, the absolute location of the first peak from each scale was also used to capture the predominance of the vertical features in the images taken from upright cameras.

B. Interscale Features

We used only orientation based features as the interscale features. Let $\{\delta_1^c, \delta_2^c, \ldots, \delta_P^c\}$ be the ordered set of peaks in the histogram at scale c, where the set elements are ordered in the descending order of their corresponding magnitudes. The features between scales i and j, β_p^{ij} were computed by comparing the p^{th} corresponding peaks of their respective histograms, *i.e.* $\beta_p^{ij} = |\cos 2(\delta_p^i - \delta_p^j)|$, where $i, j = 1, \ldots, C$. This measure favors either a continuing edge/line or near right-angle junctions at multiple scales.

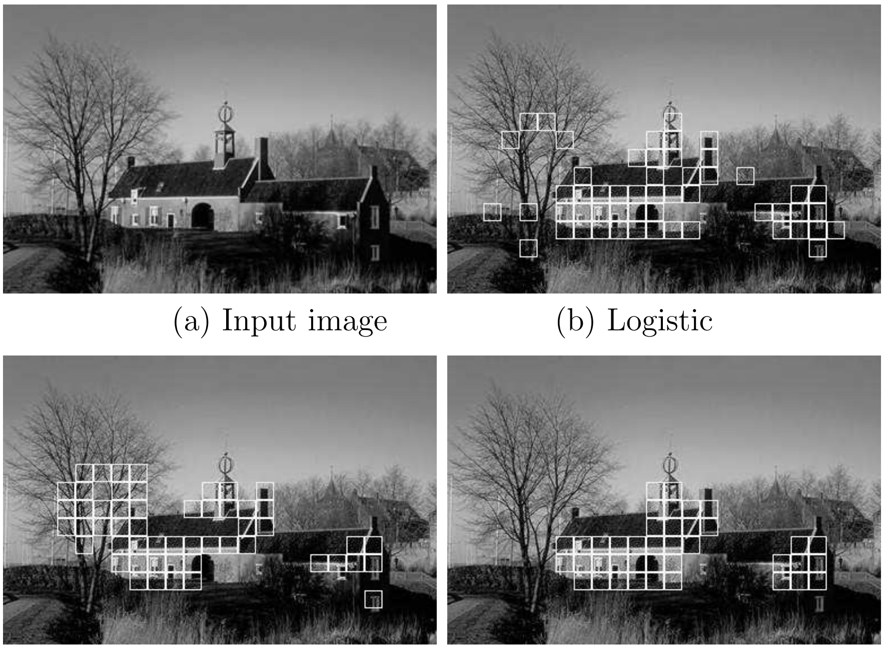
For the multiscale feature vector, the number of scales, C was chosen to be 3, with the scales changing in regular octaves. The lowest scale was fixed at 16×16 pixels, and the highest scale at 64×64 pixels. For each image block, a Gaussian smoothing kernel was used to smooth the weighted orientation histogram at each scale. The bandwidth of the kernel was chosen to be 0.7 to restrict the smoothing to two neighboring bins on each side. The moment features for orders $p \geq 1$ were found to be correlated at all the scales. Thus, we chose only two moment features, ν_0 and ν_2 at each scale. This yielded twelve intrascale features from the three scales including one orientation based feature for each scale. For the interscale features, we used only the highest peaks of the histograms at each scale, yielding two features. Hence, for each image block i, a 14 dimensional multiscale feature vector $\boldsymbol{f}_i$ was obtained. For the *single-site* feature vector, $\boldsymbol{s}_i(\boldsymbol{y_i})$, no interactions between data at multiple sites were allowed (i.e. $C = 1$). This vector was composed of first three moments and two orientation based intrascale features described above.

3.5.2. *Learning*

The parameters of the DRF model $\theta = \{\boldsymbol{w}, \boldsymbol{v}, \beta, K\}$ were learned from the training data using the maximum pseudolikelihood method described in Section 3.3. For the association potentials, a transformed feature vector $\boldsymbol{h}_i(\boldsymbol{y})$ was computed at each site i. In this work we used the quadratic transforms such that the functions $\phi_k(\boldsymbol{f}_i(\boldsymbol{y}))$ include all the l components of the feature vector $\boldsymbol{f}_i(\boldsymbol{y})$, their squares and all the pairwise products yielding $l + l(l+1)/2$ features (Figueiredo and Jain, 2001). This is equivalent to the kernel mapping of the data using a polynomial kernel of degree two. Any linear classifier in the transformed feature space will induce a quadratic boundary in the original feature space. Since l is 14, the quadratic mapping gives a 119 dimensional vector at each site. In this work, the function $\boldsymbol{\psi}_i$, defined in Section 3.2 was chosen to be the same as $\boldsymbol{f}_i$. The pairwise data vector $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$ can be obtained either by passing the two vectors $\boldsymbol{\psi}_i(\boldsymbol{y})$ and $\boldsymbol{\psi}_j(\boldsymbol{y})$ through a distance function, e.g. absolute component wise difference, or by concatenating the two vectors. We used the concatenated vector in the present work which yielded slightly better results. This is possibly due to wide within class variations in the *nonstructured* class. For the interaction potential, first order neighborhood (i.e. four nearest neighbors) was considered similar to the Ising model.

First, the parameters of the logistic functions, $\boldsymbol{w}$ and $\boldsymbol{v}$, were estimated separately to initialize the pseudolikelihood maximization scheme. Newton's method was used for logistic regression and the initial values for all the parameters were set to 0. Since the logistic log-likelihood given in (10) is convex, initial values are not a concern for the logistic regression. Approximately equal number of data points were used from both classes. For the DRF learning, the interaction parameter β was initialized to 0, i.e. no contextual interaction between the labels. The weighting parameter K was initialized to 0.5 giving equal weights to both the data-independent and the data-dependent terms in $I(x_i, x_j, \boldsymbol{y})$. All the parameters θ were learned by using gradient descent for constrained maximization. The final values of β and K were found to be 0.77, and 0.83 respectively. The learning took 100 iterations to converge in 627 s on a 1.5 GHz Pentium class machine.

To compare the results from the DRF model with those from the MRF framework, we learned the MRF parameters using the pseudolikelihood formulation. Each class conditional density was modeled as a mixture of Gaussian.



(a) Input image (b) Logistic

(c) MRF (d) DRF

Figure 2. Structure detection results on a test example for different methods. For similar detection rates, DRF reduces the false positives considerably.

The number of Gaussians in the mixture was selected to be 5 using cross-validation. The mean vectors, full covariance matrices and the mixing parameters were learned using the standard EM technique. The pseudolikelihood learning algorithm yielded β_m to be 0.68. The learning took 9.5 s to converge in 70 iterations.

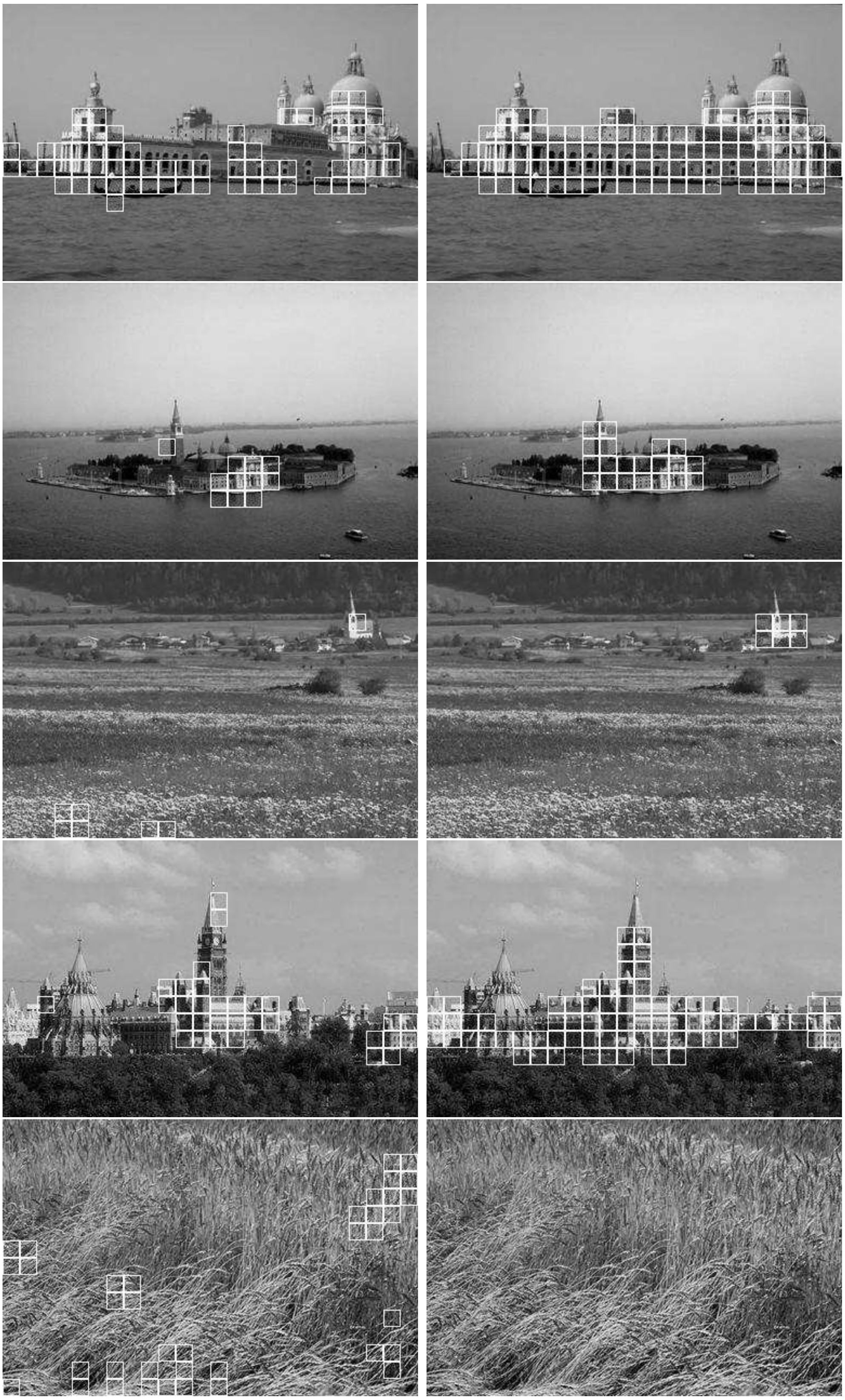
3.5.3. *Performance Evaluation*

In this section we present a qualitative as well as a quantitative evaluation of the proposed DRF model. First we compare the detection results on the test images using three different methods: logistic classifier with MAP inference, and MRF and DRF with ICM inference. The ICM algorithm was initialized from the maximum likelihood solution for the MRF and from the MAP solution of the logistic classifier for the DRF.

For an input test image given in Fig. 2(a), the *structure* detection results for the three methods are shown in Fig. 2. The blocks identified as *structured* have been shown enclosed within an artificial boundary. It can be noted that for similar detection rates, the number of false positives have significantly reduced for the DRF based detection. The logistic classifier does not enforce smoothness in the labels, which led to increased false positives. However, the MRF solution shows a smoothed false positive region around the tree branches because it does not take into account the neighborhood interaction of the data. Locally, different branches may yield features similar to those from the man-made structures. In addition, the discriminative association potential and the data-dependent smoothing in the interaction potential in the DRF also affect the detection results. Some more examples comparing the detection results of the MRF and the DRF are given in Fig. 3 and Fig. 4. The examples indicate that the data interaction is important for both increasing the detection rate as well as reducing the false positives. The ICM algorithm converged in less than 5 iterations for both the DRF and the MRF. The average time taken in processing an image of size 256×384 pixels in Matlab 6.5 on a 1.5 GHz Pentium class machine was 2.42 s for the DRF, 2.33 s for the MRF and 2.18 s for the logistic classifier. As expected, the DRF takes more time than the MRF due to the additional computation of the data-dependent term in the interaction potential in the DRF.

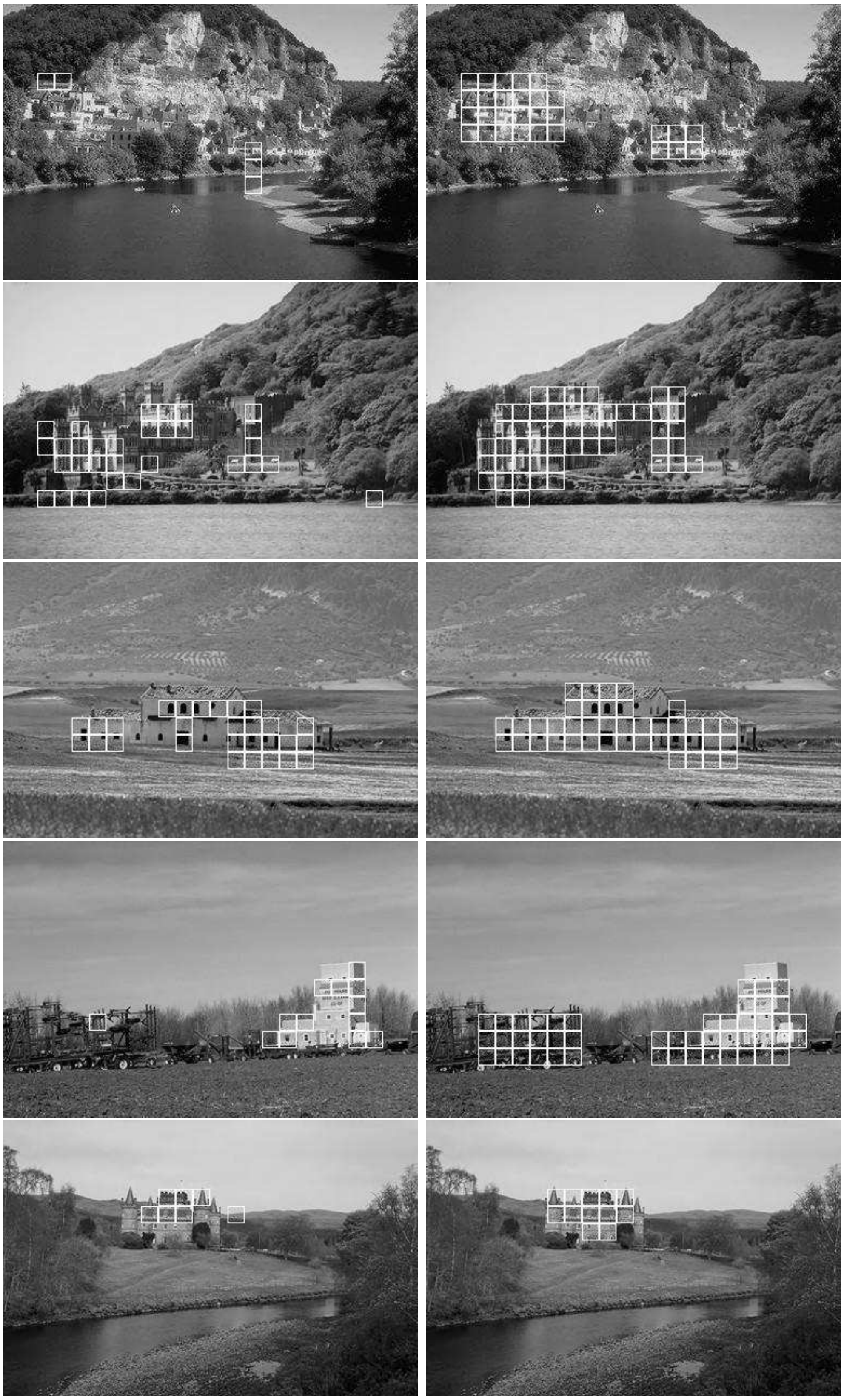
To carry out the quantitative evaluation of our work, we compared the detection rates, and the number of false positives per image for each technique. To avoid the confusion due to different effects in the DRF model, the first set of experiments was conducted using the *single-site* features for all the three methods. Thus, no neighborhood data interaction was used for both the logistic classifier and the DRF, i.e., $\boldsymbol{f}_i(.) = \boldsymbol{s}_i(.)$. The comparative results for the three methods are given in Table I next to 'MRF', 'Logistic$^-$' and 'DRF$^-$'. For comparison purposes, the false positive rate of the logistic classifier was fixed to be the same as the DRF in all the experiments. It can be noted that for similar false positives, the detection rates of the MRF and the DRF are higher than the logistic classifier due to the label interaction. However, higher detection rate of the DRF in comparison to the MRF indicates the gain due to the use of discriminative models in the association and interaction potentials in the DRF.

In the next experiment, to take advantage of the power of the DRF framework, data interaction was allowed for both the logistic classifier as well as the DRF. Further, to decouple the effect of the data-dependent term from the data-independent term in the interaction potential in the DRF, the weighting parameter K was set to 0. Thus, only data-dependent smoothing was used for the DRF. The DRF parameters were learned for this setting (Section 3.3) and β was found to be 1.26. The DRF results ('DRF($K\!=\!0$)' in Table I) show significantly higher detection rate than that from the logistic and the MRF classifiers. At the same time, the DRF reduces false positives from the MRF by more than 48%. Finally, allowing all the components of the DRF to act together, the detection rate further increases with a marginal increase in false positives ('DRF' in Table I). However, observe that for the full DRF, the learned value of $K(0.83)$ signifies that the data-independent term dominates in the interaction potential. This indicates that there is some redundancy in the smoothing effects produced by the two different terms in the interaction potential. This is not surprising because the data-dependent term in the interaction potential is based on a pairwise discriminative model which partitions the space of pairwise features $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$ such that all the pairs that are hypothesized to have similar labels lie on one side of



(a) MRF (b) DRF

Figure 3. Some more examples of structure detection from the test set. DRF has higher detection rates and lower false positives in comparison to MRF.



(a) MRF

(b) DRF

Figure 4. Additional structure detection examples from the test set.

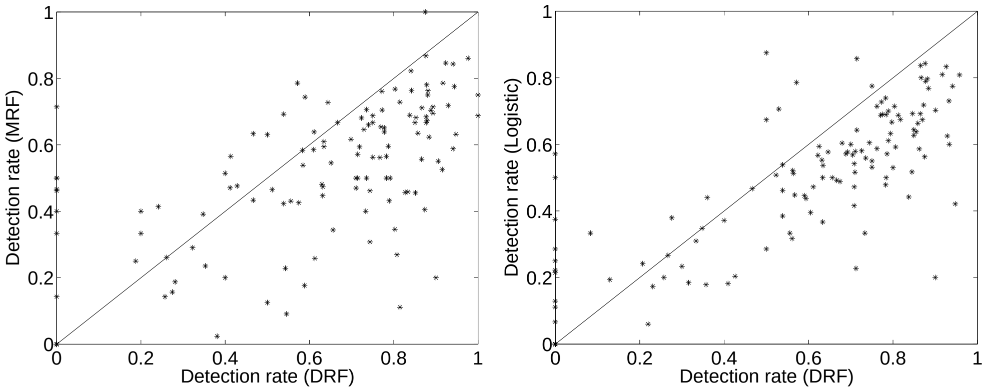


Figure 5. Comparison of the detection rates per image for the DRF and the other two methods for similar false positive rates. For most of the images in the test set, DRF detection rate is higher than others.

Table I. Detection Rates (DR) and False Positives (FP) for the test set containing 129 images. FP for logistic classifier were kept to be the same as for DRF for DR comparison. Superscript $'-'$ indicates no neighborhood data interaction was used. $K = 0$ indicates the absence of the data-independent term in the interaction potential in DRF.

	MRF	Logistic$^-$	DRF$^-$	Logistic	DRF ($K = 0$)	DRF
DR (%)	57.2	45.5	60.9	55.4	68.6	70.5
FP (per image)	2.36	2.24	2.24	1.37	1.21	1.37

the decision boundary. Hence, effectively this term is also implicitly modeling the similarity of the labels at neighboring sites similar to the Ising model of the data-independent term. In Section 4.1 we will describe a modified form of the interaction potential that combines these two terms without duplicating their smoothing effects. To compare per image performance of the DRF with the MRF and the logistic classifier, scatter plots were obtained for the detection rates for each image (Fig. 5). Each point on a plot is an image from the test set. These plots indicate that for a majority of the images the DRF has higher detection rate than the other two methods.

To analyze the performance of the MAP inference for the DRF, a MAP solution was obtained using the min-cut algorithm. The overall detection rate was found to be 24.3% for 0.41 false positives per image. Very low detection rate along with low false positives indicates that MAP is preferring oversmoothed solutions in the present setting. This is because the pseudolikelihood approximation used in this work for learning the parameters tends to overestimate the interaction parameter β. Our MAP results match the observations made by Greig et al. (Greig et al., 1989), and Fox and Nicholls (Fox and Nicholls, 2000) for large

Table II. Results with linear classifiers (See text for more).

	Logistic(linear)	DRF (linear)
DR (%)	55.0	62.3
FP (per image)	2.04	2.04

values of β in MRFs. In contrast, ICM is more resilient to the errors in parameter estimation and performs well even for large β, which is consistent with the results of (Greig et al., 1989), (Fox and Nicholls, 2000), and Besag (Besag, 1986). For MAP to perform well, a better parameter learning procedure than using a factored approximation of the likelihood will be helpful. In addition, one may also need to impose a prior that favors small values of β. These observations lay the foundation for improved parameter learning procedure explained in Section 4.2.

One additional aspect of the DRF model is the use of general kernel mappings to increase the classification accuracy. To assess the sensitivity to the choice of kernel, we changed the quadratic functions used in the DRF experiments to compute $\boldsymbol{h}_i(\boldsymbol{y})$ to one-to-one transform such that $\boldsymbol{h}_i(\boldsymbol{y}) = [1 \;\, \boldsymbol{f}_i(\boldsymbol{y})]$. This transform will induce a linear decision boundary in the feature space. The DRF results with quadratic boundary (Table I) indicate higher detection rate and lower false positives in comparison to the linear boundary (Table II). This shows that with more complex decision boundaries one may hope to do better. However, since the number of parameters for a general kernel mapping is of the order of the number of data points, one will need some method to induce sparseness to avoid overfitting (Tipping, 2000)(Figueiredo and Jain, 2001).

4. Modified Discriminative Random Field

As explained in the previous section, there were two main reasons that prompted us to explore a modified form of the original DRF and a better parameter learning procedure:

1. The form of the interaction potential given in (8) has redundancy in the smoothing effects produced by the data-independent and the data-dependent terms. Also, this form makes the parameter learning a non-convex problem.

2. The pseudolikelihood parameter learning tends to overestimate the interaction coefficients which makes the global MAP estimates to be bad solutions.

In the following sections we discuss the main components of the original DRF formulation that have been modified.[6]

4.1. Interaction potential

For a pair of sites (i,j), let $\boldsymbol{\mu}_{ij}(\boldsymbol{\psi}_i(\boldsymbol{y}), \boldsymbol{\psi}_j(\boldsymbol{y}))$ be a new feature vector such that $\boldsymbol{\mu}_{ij} : \Re^{\gamma} \times \Re^{\gamma} \rightarrow \Re^{q}$, where $\boldsymbol{\psi}_k : \boldsymbol{y} \rightarrow \Re^{\gamma}$. Denoting this feature vector as $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$ for simplification, the interaction potential is modeled as,

$$I(x_i, x_j, \boldsymbol{y}) = x_i x_j \boldsymbol{v}^T \boldsymbol{\mu}_{ij}(\boldsymbol{y}) \tag{13}$$

where $\boldsymbol{v}$ are the model parameters. Note that the first component of $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$ is fixed to be 1 to accommodate the bias parameter. There are two interesting properties of the interaction potential given in (13). First, if the association potential at each site and the interaction potentials of all the pairwise cliques except the pair (i,j) are set to zero in (2), the DRF acts as a logistic classifier which yields the probability of the site pair to have the same labels given the observed data. Of course, one can generalize the form in (13) as,

$$I(x_i, x_j, \boldsymbol{y}) = \log P''(x_i, x_j | \boldsymbol{\psi}_i(.), \boldsymbol{\psi}_j(.)), \tag{14}$$

similar to the association potential in Section 3.1 and can use arbitrary pairwise discriminative classifier to define this term. Recently, a similar idea has been used by other researchers (Qi et al., 2005)(Torralba et al., 2005). The second property of the interaction potential form given in (13) is that it generalizes the Ising model. The original Ising form is recovered if all the components of vector $\boldsymbol{v}$ other than the bias parameter are set to zero in (13). Thus, the form of interaction potential given in (13) effectively combines both the terms of the earlier model in (8). A geometric interpretation of interaction potential is that it partitions the space induced by the relational features $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$ between the pairs that have the same labels and the ones that have different labels. Hence (13) acts as a data-dependent discontinuity adaptive model that will moderate smoothing when the data from the two sites is 'different'. The data-dependent smoothing can especially be useful to absorb the errors in modeling the association potential. Anisotropy can be easily included in the DRF model by parametrizing the interaction potentials of different directional pairwise cliques with different sets of parameters $\boldsymbol{v}$.

4.2. Parameter learning and inference

Let θ be the set of DRF parameters where $\theta = \{\boldsymbol{w}, \boldsymbol{v}\}$. As shown in Section 3.5.3, pseudolikelihood tends to overestimate the interaction parameters causing the MAP estimates of the field to be very poor solutions. Our experiments in

[6] Early version of this work appeared in Advances in Neural Information Processing Systems (NIPS 03) (Kumar and Hebert, 2003a).

Section 4.3 verify these observations for the interaction parameters $\boldsymbol{v}$ in modified DRFs too. To alleviate this problem, we take a Bayesian approach to get the maximum a posteriori estimates of the parameters. Similar to the concept of weight decay in neural learning literature, we assume a Gaussian prior over the interaction parameters $\boldsymbol{v}$ such that $p(\boldsymbol{v}|\tau) = \mathcal{N}(\boldsymbol{v}; \mathbf{0}, \tau^2\boldsymbol{I})$ where $\boldsymbol{I}$ is the identity matrix. Using a prior over parameters $\boldsymbol{w}$ that leads to weight decay or shrinkage might also be beneficial but we leave that for future exploration. The prior over parameters $\boldsymbol{w}$ is assumed to be uniform. Thus, given M independent training images,

$$\widehat{\theta} = \arg\max_{\theta} \sum_{m=1}^{M} \sum_{i \in S} \left\{ \log \sigma(x_i \boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y})) + \sum_{j \in \mathcal{N}_i} x_i x_j \boldsymbol{v}^T \boldsymbol{\mu}_{ij}(\boldsymbol{y}) - \log z_i \right\} - \frac{1}{2\tau^2} \boldsymbol{v}^T \boldsymbol{v} \tag{15}$$

$$\text{where} \qquad z_i = \sum_{x_i \in \{-1,1\}} \exp \left\{ \log \sigma(x_i \boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y})) + \sum_{j \in \mathcal{N}_i} x_i x_j \boldsymbol{v}^T \boldsymbol{\mu}_{ij}(\boldsymbol{y}) \right\}$$

If τ is given, the penalized log pseudolikelihood in (15) is convex with respect to the model parameters and can be easily maximized using gradient descent.

In related work regarding the estimation of τ, Mackay (Mackay, 1996) has suggested the use of type II marginal likelihood. But in the DRF formulation, integrating the parameters $\boldsymbol{v}$ is a hard problem. Another choice is to integrate out τ by choosing a non-informative hyperprior on τ as in (Williams, 1995) (Figueiredo, 2001). However our experiments showed that these methods do not yield good estimates of the parameters because of the use of pseudolikelihood in our framework. In the present work we choose τ by cross-validation. Alternative ways of parameter estimation include the use of contrastive divergence (Hinton, 2002) and saddle point approximations resembling perceptron learning rules (Collins, 2002). We are currently exploring other possibilities of parameter learning, as discussed in our recent work (Kumar et al., 2005).

To test the efficacy of the penalized pseudolikelihood procedure, we were interested in obtaining the MAP estimates of labels $\boldsymbol{x}$ given an image $\boldsymbol{y}$. Following the discussion in Section 3.4, the MAP estimates for the modified DRFs can also be obtained using graph min-cut algorithms. However, since these algorithms do not allow negative interaction between the sites, the data-dependent smoothing for each clique in (13) is set to be $\boldsymbol{v}^T\boldsymbol{\mu}_{ij}(\boldsymbol{y}) = \max\{0, \boldsymbol{v}^T\boldsymbol{\mu}_{ij}(\boldsymbol{y})\}$, yielding an approximate MAP estimate. This is equivalent to switching the smoothing off at the image discontinuities.

4.3. Man-made Structure Detection Revisited

The modified DRF model was applied to the task of detecting man-made structures in natural scenes. The features were fixed to be the same as used in the tests with the original DRF in Section 3.5. The penalty coefficient τ was chosen to be 0.001 for parameter learning. The detection results were obtained using graph min-cuts for both the MRF and the DRF models.

For a quantitative evaluation, we compared the detection rates and the number of false positives per image for the MRF, the DRF and the logistic classifier. Similar to the experimental procedure of Section 3.5.3, for the comparison of detection rates in all the experiments, the decision threshold of the logistic classifier was fixed such that it yields the same false positive rate as the DRF. The first set of experiments was conducted using the *single-site* features for all the three methods. Thus, no neighborhood data interaction was used for both the logistic classifier and the DRF, i.e. $\boldsymbol{f}_i(\boldsymbol{y}) = \boldsymbol{s}_i(\boldsymbol{y}_i)$. The comparative results for the three methods are given in Table III under 'MRF', 'Logistic^{-}' and 'DRF^{-}'. The detection rates of the MRF and the DRF are higher than the logistic classifier due to the label interaction. However, higher detection rate and lower false positives for the DRF in comparison to the MRF indicate the gains due to the use of discriminative models in the association and interaction potentials in the DRF. In the next experiment, to take advantage of the power of the DRF framework, data interaction was allowed for both the logistic classifier as well as the DRF ('Logistic' and 'DRF' in Table III). The DRF detection rate increases substantially and the false positives decrease further indicating the importance of allowing the data interaction in addition to the label interaction.

Now we compare the results of the modified DRF formulation with those from the original DRF. Comparing the results in table I with those in table III, we find that the original DRF (with ICM inference) gave 70.5% correct detection with 1.37 average false positive per image in comparison to 72.5% correction detection and 1.76 false positives from the modified DRF (with MAP inference). Even though the results seem to be comparable for this application, we have achieved two main advantages in modified DRFs. In comparison to the original DRF formulation, the modified DRF has a much simpler form of interaction potential with comparatively better behaved parameter learning problem (a convex problem). It also overcomes the criticism of the original DRFs that if the global minimum of the energy $(-\log P(x|y))$ is not an acceptable solution, it probably implies that the DRFs are not appropriate models for the purpose of classification. Clearly, the experiments in this section reveal that the bad MAP solutions of DRFs were due to a particular parameter learning scheme (pseudolikelihood) we chose in our earlier experiments. These results also point toward another interesting observation regarding the compatibility of a parameter learning procedure with the inference procedure. Local parameter learning (pseudolikelihood) seems to be yielding acceptable, though usually not the best, results when used with a local inference mechanism (ICM). On the other hand,

Table III. Detection Rates (DR) and False Positives (FP) for the test set containing 129 images ($49,536$ sites). FP for logistic classifier were kept to be the same as for DRF for DR comparison. Superscript $'-'$ indicates no neighborhood data interaction was used.

	MRF	Logistic^{-}	DRF^{-}	Logistic	DRF
DR (%)	58.35	47.50	61.79	60.80	72.54
FP (per image)	2.44	2.28	2.28	1.76	1.76

to make a global inference scheme yield good solutions, it is inevitable to use nonlocal learning procedures. We are currently exploring this duality between parameter learning and inference in a more systematic manner (Kumar et al., 2005).

4.4. Binary Image Denoising Task

The aim of these experiments was to obtain denoised images from corrupted binary images. Four base images, 64 × 64 pixels each, were used in the experiments (top row in Fig. 6). We compare the DRF and the MRF results for two different noise models. For each noise model, 50 images were generated from each base image. Each pixel was considered as an image site and the feature vector $\boldsymbol{s}_i(\boldsymbol{y}_i)$ was simply chosen to be a scalar representing the intensity at i^{th} site. In experiments with the synthetic data, no neighborhood data interaction was used for the DRFs (i.e., $\boldsymbol{f}_i(\boldsymbol{y}) = \boldsymbol{s}_i(\boldsymbol{y}_i)$) to observe the gains only due to the use of discriminative models in the association and interaction potentials. A linear discriminant was implemented in the association potential such that $\boldsymbol{h}_i(\boldsymbol{y}) = [1, \boldsymbol{f}_i(\boldsymbol{y})]^T$. The pairwise data vector $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$ was obtained by taking the absolute difference of $\boldsymbol{s}_i(\boldsymbol{y}_i)$ and $\boldsymbol{s}_j(\boldsymbol{y}_j)$. For the MRF model, each class-conditional density, $p(\boldsymbol{s}_i(\boldsymbol{y}_i)|x_i)$, was modeled as a Gaussian. The noisy data from the left most base image in Fig. 6 was used for training while 150 noisy images from the rest of the three base images were used for testing.

Three experiments were conducted for each noise model. (i) The interaction parameters for the DRF ($\boldsymbol{v}$) as well as for the MRF (β_m) were set to zero. This reduces the DRF model to a logistic classifier and MRF to a maximum likelihood (ML) classifier. (ii) The parameters of the DRF i.e., $[\boldsymbol{w}, \boldsymbol{v}]$, and the MRF i.e., β_m, were learned using pseudolikelihood approach without any penalty i.e., $\tau = \infty$. (iii) Finally, the DRF parameters were learned using penalized pseudolikelihood and the best β_m for the MRF was chosen from cross-validation. The MAP estimates of the labels were obtained using graph-cuts for both models.

Under the first noise model, each image pixel was corrupted with independent Gaussian noise of standard deviation 0.3. For the DRF parameter learning, τ

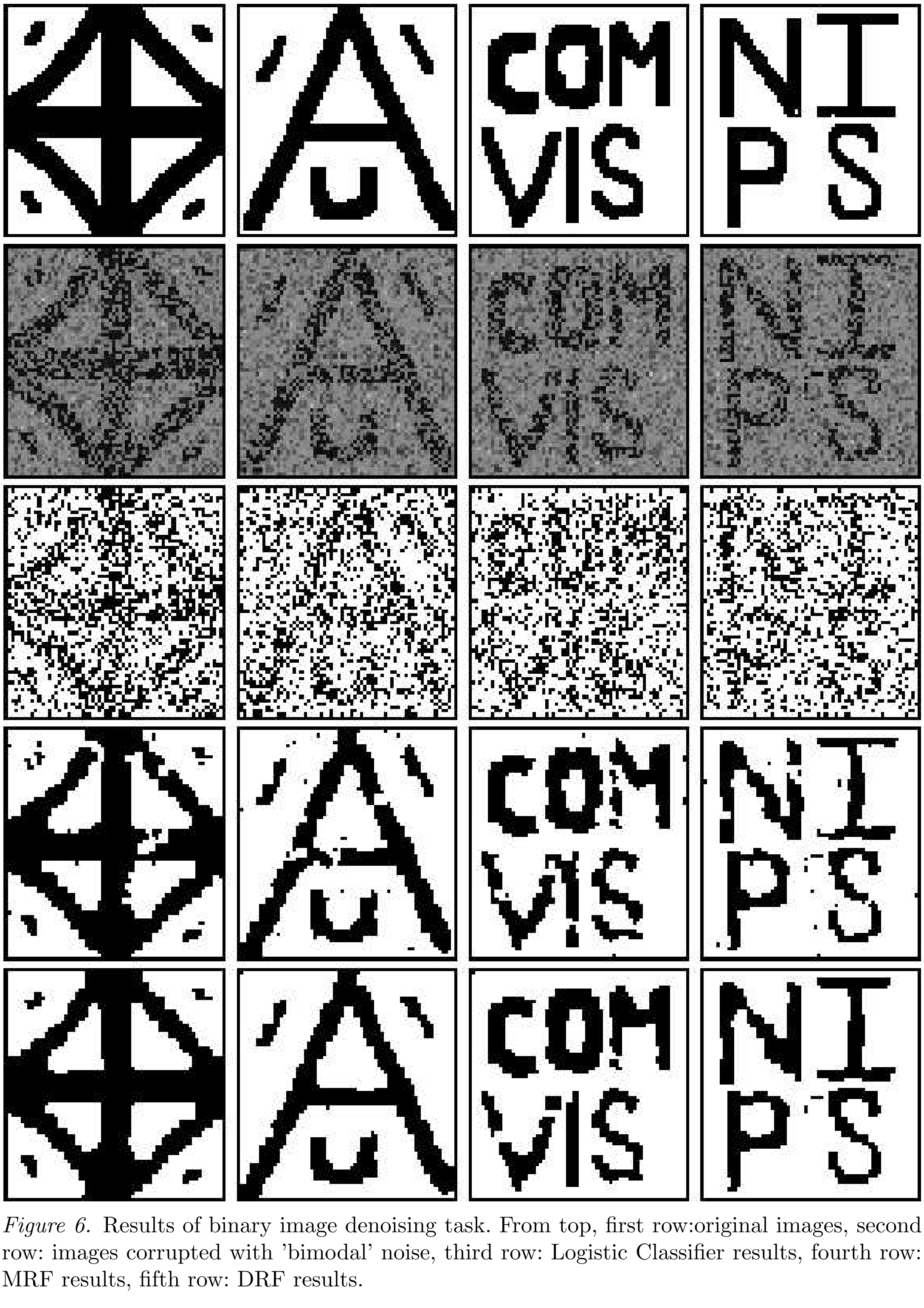


Figure 6. Results of binary image denoising task. From top, first row:original images, second row: images corrupted with 'bimodal' noise, third row: Logistic Classifier results, fourth row: MRF results, fifth row: DRF results.

Table IV. Pixelwise classification errors (%) on 150 synthetic test images. For the Gaussian noise MRF and DRF give similar error while for 'bimodal' noise, DRF performs better. Note that only label interaction (i.e., no data interaction) was used for these tests. 'PL': Pseudo-Likelihood parameter learning, 'PPL': Penalized Pseudo-Likelihood parameter learning.

Noise	ML	Logistic	MRF (PL)	DRF (PL)	MRF	DRF (PPL)
Gaussian	15.62	15.78	2.66	3.82	2.35	2.30
Bimodal	24.00	29.86	8.70	17.69	7.00	6.21

was chosen to be 0.01. The pixelwise classification error for this noise model is given in the top row of Table IV. Since the form of noise is the same as the likelihood model in the MRF, MRF is expected to give good results. The DRF model does marginally better than MRF even for this case. Note that the DRF with penalized pseudolikelihood parameter learning (suffix 'PPL' in Table IV) yields significantly better results than without penalizing the pseudolikelihood (suffix 'PL'). Similarly, the MRF results are also worse when the parameters were learned simply using the pseudolikelihood. With pseudolikelihood parameters, MAP inference yields oversmoothed images. The DRF model is affected more because all the parameters in DRFs are learned simultaneously unlike MRFs.

In the second noise model each pixel was corrupted with independent mixture of Gaussian noise. For each class, a mixture of two Gaussians with equal mixing weights was used yielding a 'bimodal' class noise. The mixture model parameters (mean, std) for the two classes were chosen to be $[(0.08, 0.03), (0.46, 0.03)]$, and $[(0.55, 0.02), (0.42, 0.10)]$ inspired by (Rubinstein and Hastie, 1997). The classification results are shown in the bottom row of Table IV. There was 11.3% relative reduction in pixelwise classification error on the test set with the DRF model over the MRF model. An interesting point to note is that DRF yields lower error than MRF even when the logistic classifier has higher error than the ML classifier on the test data. For a typical noisy version of the four base images, the performance of different techniques in compared in Fig. 6. The logistic classifier gives very poor results because it classifies each pixel independently. It ignores the very basic theme of underlying smoothness of natural images due to which one can hope for recovering the true image from its noisy version. The DRF gives better performance than the MRF model.

5. Conclusion and Future Work

In this work we have introduced Discriminative Random Fields that combine local discriminative classifiers for individual classification of image sites with interaction between neighboring sites. These models allow capturing spatial dependencies in labels and observed data simultaneously in a principled manner on 2D lattices with cycles. The results on various synthetic and real-world images validate the advantages of the DRF model. The proposed DRF framework is general enough to encompass several computer vision tasks varying from low level image denoising to high level object detection. However, there are several extensions required to demonstrate the application of DRFs to high level classification tasks. The first natural step is to extend the proposed binary DRF model to accommodate multiclass classification problems. We have already developed this framework and are presently conducting extended experiments. The most important challenge in the DRF framework is robust and fast learning of the model parameters. Currently we are exploring alternative ways of learning DRF parameters using saddle point approximations which will also be applicable to the conventional MRF models used in classification. Finally, one future aspect of the DRF model is the use of general kernel mappings to increase the classification accuracy. However, we will need some method to induce sparseness in the parameter space to avoid overfitting in a very high dimensional kernel space.

Acknowledgments. We would like to thank John Lafferty, Jonas August and Tom Minka for having several thoughtful discussions on various issues related to random fields and classification problems. Graph min-cut code used in this work was provided by V. Kolmogorov.

References

Barrett, W. A. and K. D. Petersen: 2001, ‘Houghing the Hough: Peak Collection for Detection of Corners, Junctions and Line Intersections’. *In Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition* **2**, 302–309.

Besag, J.: 1986, ‘On the Statistical Analysis of Dirty Pictures’. *Journal of Royal Statistical Soc.* **B-48**, 259–302.

Blake, A., C. Rother, M. Brown, P. Perez, and P. Torr: 2004, ‘Interactive Image Segmentation Using an Adaptive GMMRF Model’. *In Proc. European Conf. on Computer Vision (ECCV).*

Bottou, L.: 1991, *Une Approache theorique de l’Apprentissage Connexionniste Applications a la Reconnaissance de la Parole.* France: PhD thesis, University de Paris.

Bouman, C. A. and M. Shapiro: 1994, ‘A Multiscale Random Field Model for Bayesian Image Segmentation’. *IEEE Trans. on Image Processing* **3**(2), 162–177.

Boykov, Y. and M.-P. Jolly: 2001, ‘Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D images’. *In Proc. International Conference on Computer Vision (ICCV)* **I**, 105–112.

Cheng, H. and C. A. Bouman: 2001, ‘Multiscale Bayesian Segmentation Using a Trainable Context Model’. *IEEE Trans. on Image Processing* **10**(4), 511–525.

Christmas, W. J., J. Kittler, and M. Petrou: 1995, ‘Structural Matching in Computer Vision using Probabilistic Relaxation’. *IEEE Trans. Pattern Anal. Machine Intell.* **17**(8), 749–764.

Collins, M.: 2002, ‘Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms’. *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Felzenszwalb, P. F. and D. P. Huttenlocher: 2000, 'Pictorial Structures for Object Recognition'. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 00).*

Feng, X., C. K. I. Williams, and S. N. Felderhof: 2002, 'Combining Belief Networks and Neural Networks for Scene Segmentation'. *IEEE Trans. Pattern Anal. Machine Intelligence* **24**(4), 467–483.

Fergus, R., P. Perona, and A. Zisserman: 2003, 'Object Class Recognition by Unsupervised Scale-Invariant Learning'. *In Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 03)* **2**, 264–271.

Figueiredo, M. A. T.: 2001, 'Adaptive Sparseness using Jeffreys Prior'. *Advances in Neural Information Processing Systems (NIPS).*

Figueiredo, M. A. T. and A. K. Jain: 2001, 'Bayesian Learning of Sparse Classifiers'. *In Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition* **1**, 35–41.

Fox, C. and G. Nicholls: 2000, 'Exact MAP States and Expectations from Perfect Sampling: Greig, Porteous and Seheult Revisited'. *In Proc. Twentieth Int. Workshop on Bayesian Inference and Maximum Entropy Methods in Sci. and Eng.*

Geman, S. and D. Geman: 1984, 'Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images'. *IEEE Trans. on Patt. Anal. Mach. Intelli.* **6**, 721–741.

Gill, P. E., W. Murray, and M. H. Wright: 1981, *Practical Optimization.* San Diego: Academic Press.

Greig, D. M., B. T. Porteous, and A. H. Seheult: 1989, 'Exact Maximum a Posteriori Estimation for Binary Images'. *Journal of Royal Statis. Soc.* **51**(2), 271–279.

Guo, C. E., S. Zhu, and Y. N. Wu: 2003, 'Modeling Visual Patterns by Integrating Descriptive and Generative Models'. *International Journal of Computer Vision* **53**(1), 5–29.

Hammersley, J. M. and P. Clifford, 'Markov Field on Finite Graph and Lattices'. *Unpublished.*

He, X., R. Zemel, and M. Carreira-Perpinan: 2004, 'Multiscale conditional random fields for image labelling'. *IEEE Int. Conf. CVPR.*

Hinton, G. E.: 2002, 'Training Product of Experts by Minimizing Contrastive Divergence'. *Neural Computation* **14**, 1771–1800.

Ising, E.: 1925, 'Beitrag Zur Theorie Der Ferromagnetismus'. *Zeitschrift Fur Physik* **31**, 253–258.

Kittler, J.: 1997, 'Probabilistic Relaxation: Potential, Relationships and Open Problems'. *In Proc. Energy Minimization Methods in Computer Vision and Pattern Recognition* pp. 393–408.

Kittler, J. and E. R. Hancock: 1989, 'Combining Evidence in Probabilistic Relaxation'. *Int. Jour. Pattern Recog. Artificial Intelli.* **3**(1), 29–51.

Kittler, J. and J. Illingworth: 1985, 'Relaxation Labeling algorithms - a review'. *Image and Vision Computing* **3**(4), 206–216.

Kittler, J. and D. Pairman: 1985, 'Contextual Pattern Recognition Applied to Cloud Detection and Identification'. *IEEE Trans.on Geo. and Remote Sensing* **23**(6), 855–863.

Kolmogorov, V. and R. Zabih: 2002, 'What Energy Functions can be Minimized via Graph Cuts'. *In Proc. European Conf. on Computer Vision* **3**, 65–81.

Krishnamachari, S. and R. Chellappa: 1996, 'Delineating Buildings by Grouping Lines with MRFs'. *IEEE Trans. on Pat. Anal. Mach. Intell.* **5**(1), 164–168.

Kumar, S., J. August, and M. Hebert: 2005, 'Exploiting Inference for Approximate Parameter Learning in Discriminative Fields: An Empirical Study'. *Fourth Int. Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR).*

Kumar, S. and M. Hebert: 2003a, 'Discriminative Fields for Modeling Spatial Dependencies in Natural Images'. *in advances in Neural Information Processing Systems (NIPS).*

Kumar, S. and M. Hebert: 2003b, 'Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification'. *in proc. IEEE International Conference on Computer Vision (ICCV)* **2**, 1150–1157.

Kumar, S. and M. Hebert: 2003c, 'Man-made structure detection in natural images using a causal multiscale random field'. *In Proc. IEEE Int. Conf. on Comp. Vision and Pattern Recog. (CVPR)* **1**, 119–126.

Kumar, S., A. C. loui, and M. Hebert: 2003, 'An Observation-Constrained Generative Approach for Probabilistic Classification of Image Regions'. *Image and Vision Computing, Special Issue on Generative Models Based Vision* **21**, 87–97.

Lafferty, J., A. McCallum, and F. Pereira: 2001, 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data'. *In Proc. Int. Conf. on Machine Learning.*

Lafferty, J., X. Zhu, and Y. Liu: 2004, 'Kernel conditional random fields: Representation and clique selection'. *In Proc. Twenty-First International Conference on Machine Learning (ICML).*

Li, S. Z.: 2001, *Markov Random Field Modeling in Image Analysis.* Tokyo: Springer-Verlag.

Mackay, D.: 1996, 'Bayesian Non-linear Modelling for the 1993 Energy Prediction Competition'. *In Maximum Entropy and Bayesian Methods* pp. 221–234.

McCullagh, P. and J. A. Nelder: 1987, *Generalised Linear Models.* London: Chapman and Hall.

Minka, T. P.: 2001, *Algorithms for Maximum-Likelihood Logistic Regression.* Carnegie Mellon University: Statistics Tech Report 758.

Murphy, K., A. Torralba, and W. T. Freeman: 2003, 'Using the Forest to See the Trees:A Graphical Model Relating Features, Objects and Scenes'. *in Advances in Neural Information Processing Systems (NIPS 03).*

Ng, A. Y. and M. I. Jordan: 2002, 'On Discriminative Vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes'. *Advances in Neural Information Processing Systems (NIPS).*

Pieczynski, W. and A. N. Tebbache: 2000, 'Pairwise Markov Random Fields and its Application in Textured Images Segmentation'. *In Proc. 4th IEEE Southwest Symposium on Image Analysis and Interpretation* pp. 106–110.

Qi, Y., M. Szummer, and T. P. Minka: 2005, 'Diagram Structure Recognition by Bayesian Conditional Random Fields'. *In Proc. International Conference on Computer Vision and Pattern Recognition (CVPR).*

Quattoni, A., M. Collins, and T. Darrell: 2004, 'Conditional Random Fields for Object Recognition'. *Neural Information Processing Systems (NIPS).*

Rosenfeld, A., R. Hummel, and S. Zucker: 1976, 'Scene Labeling by Relaxation Operations'. *IEEE Trans System, Man, Cybernatics* **SMC-6**, 420–433.

Rubinstein, Y. D. and T. Hastie: 1997, 'Discriminative vs Informative Learning'. *In Proc. Third Int. Conf. on Knowledge Discovery and Data Mining* pp. 49–53.

Szummer, M. and Y. Qi: 2004, 'Contextual Recognition of Hand-drawn Diagrams with Conditional Random Fields'. *Workshop on Frontiers in Handwriting Recognition.*

Taskar, B., C. Guestrin, and D. Koller: 2003, 'Max-Margin Markov Network'. *Neural Information Processing Systems Conference (NIPS03).*

Tipping, M.: 2000, 'The Relevance Vector Machine'. *Advances in Neural Information Processing Systems-NIPS 12* pp. 652–658.

Torralba, A., K. P. Murphy, and W. T. Freeman: 2005, 'Contextual Models for Object Detection using Boosted Random Fields'. *Adv. in Neural Information Processing Systems (NIPS).*

Waltz, D. L.: 1975, *Understanding Line Drawing of Scenes with Shadows.* New York: The Psychology of Computer Vision, P H Winston, ed. McGraw-Hill.

Wang, Y. and Q. Ji: 2005, 'A Dynamic Conditional Random Field Model for Object Segmentation in Image Sequences'. *In Proc. IEEE Int. Conf. on Comp. Vision and Pattern Recog. (CVPR)* **1**, 264–270.

Weber, M., M. Welling, and P. Perona: 2000, 'Towards Automatic Discovery of Object Categories'. *In Proc.IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 00).*

Weinman, J., A. Hanson, and A. McCallum: 2004, 'Sign Detection in Natural Images with Conditional Random Fields'. *In Proc. of IEEE International Workshop on Machine Learning for Signal Processing.*

Williams, C. K. I. and N. J. Adams: 1999, 'DTs: Dynamic Trees'. *Advances in Neural Information Processing Systems* **11**.

Williams, P.: 1995, 'Bayesian Regularization and Pruning Using a Laplacian Prior'. *Neural Computation* **7**, 117–143.

Wilson, R. and C. T. Li: 2003, 'A Class of Discrete Multiresolution Random Fields and Its Application to Image Segmentation'. *IEEE Trans. on Pattern Anal. and Machine Intelli.* **25**(1), 42–56.

Won, C. S. and H. Derin: 1992, 'Unsupervised Segmentation of Noisy and Textured Images using Markov Random Fields'. *CVGIP* **54**, 308–328.

Xiao, G., M. Brady, J. A. Noble, and Y. Zhang: 2002, 'Segmentation of Ultrasound B-Mode Images with Intensity Inhomogeneity Correction'. *IEEE Trans. on Medical Imaging* **21**(1), 48–57.