Man-Made Structure Detection in Natural Images using a Causal Multiscale Random Field

Sanjiv Kumar and Martial Hebert
The Robotics Institute, Carnegie Mellon University
Pittsburgh, PA 15213, USA, {skumar, hebert}@ri.cmu.edu

Abstract

This paper presents a generative model based approach to man-made structure detection in 2D natural images. The proposed approach uses a causal multiscale random field suggested in [3] as a prior model on the class labels on the image sites. However, instead of assuming the conditional independence of the observed data, we propose to capture the local dependencies in the data using a multiscale feature vector. The distribution of the multiscale feature vectors is modeled as mixture of Gaussians. A set of robust multiscale features is presented that captures the general statistical properties of man-made structures at multiple scales without relying on explicit edge detection. The proposed approach was validated on real-world images from the Corel data set, and a performance comparison with other techniques is presented.

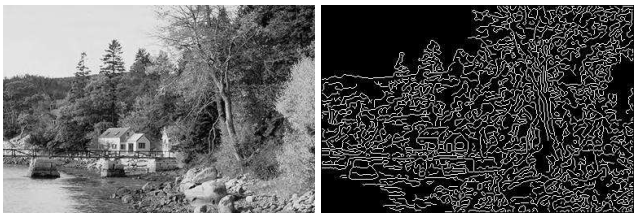
1. Introduction

Automatic detection of man-made structure in ground-level images is useful for scene understanding, robotic navigation, surveillance, image indexing and retrieval etc. This paper focuses on the detection of man-made structures, which can be characterized primarily by the presence of linear structures. The detection of such a constrained set of man-made structures from a single static ground-level image is still a non-trivial problem due to three main reasons. First, the realistic views of a structured object captured from a ground-level camera are unconstrained unlike the aerial views, which complicates the use of predefined models or model-specific properties in detection. Second, no motion or stereo information is available, precluding the use of geometrical information pertaining to the structure. Finally, the images of natural scenes contain large amount of clutter, and the edge extraction is very noisy. This makes the computation of the image primitives such as junctions, angles etc., which rely on explicit edge or line detection, prone to errors.

Buildings are one possible instance of man-made structures and some of the related work on structure detection exists for buildings [13][12][9][7][4]. A majority of the techniques for building detection from aerial imagery try to generate a hypothesis on the presence of building roof-tops in the scene [13]. This is usually attained by first detecting low-level image primitives, e.g. edges, lines or junctions, and then grouping these primitives using either geometric-model based heuristics [12], or a statistical model, e.g. Markov Random Field (MRF) [9]. For the ground-level images, the detection of roof-tops is not feasible and shadows do not constrain the detection problem unlike the aerial images.

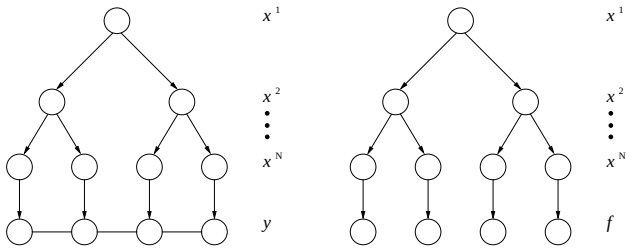
Perceptual Organization based building detection has been presented in [7] for image retrieval. In [17] a technique was proposed to learn the parameters of a large perceptual organization using graph spectral partitioning. However, these techniques also require the low-level image primitives to be computed explicitly, and to be relatively noise-free. There has been some recent research work regarding the classification of a whole image as a landscape or an urban scene [14][18]. Oliva and Torralba [14] obtain a low-dimensional holistic representation of the scene using principal components of the power spectra. We found the power spectra based features to be noisy for our images, which contain a mixture of both the landscape and man-made regions within the same image. It might be due to the fact that a 'single' image (or a region contained in it) may not follow the assumption that the power spectra falls with a form $f^{-\alpha}$ where f is spatial frequency [10]. Vailaya et al. [18] use the edge coherence histograms over the whole image for the scene classification, using edge pixels at different orientations. Olmos and Trucco [15] have recently proposed a system to detect the presence of man-made objects in underwater images using properties of the contours in the image. The techniques which classify the whole image in a certain class implicitly assume the image to be exclusively containing either man-made or natural objects, which is not true for many real-world images.

The techniques described in [5][8] perform classifica-



(a) Input image (b) Edge image

Figure 1. A natural image and the corresponding edge image obtained using Canny edge detector to illustrate that reliable extraction of low-level image primitives, $e.g.$ lines, edges or junctions for man-made structure detection is hard in natural images.



(a) Proposed model (b) Approximated model

Figure 2. A 1-D representation of the quad-tree structured MSRF based image generative model. See text for more.

tion in outdoor images using color and texture features, but employ different classification schemes. These papers report poor performance on the classes containing man-made structures since color and texture features are not very informative for these classes [18]. In addition, in comparison to the Sowerby database used by them, we use a more diverse set of images from the Corel database for training as well as testing.

In this paper, we propose to detect man-made structures in a 2D image, located at medium to long distances from the camera. To visualize the problems with low-level primitives using edges, an input image and the corresponding edge image obtained from the Canny edge detector are shown in Figure 1. It is clear that detection based on these primitives is going to be a daunting task for this type of images. Instead, in the present work we propose a hybrid approach which uses the bottom-up approach of extracting generic features from the image blocks, followed by the top-down approach of classifying image blocks based on statistical distribution of the features learned from the training data.

2. Image Generative Model

Given an input image, the detection problem can be posed as a classification problem where each site (a block or a pixel) in the image is classified into the *structured* class or the *nonstructured* class. Let $\boldsymbol{y}$ be the observed data associated with the input image, where $\boldsymbol{y} = \{y_m\}_{m=1}^{M}$, y_m be the data from m^{th} site. Let the corresponding labels at the image sites be given by $\boldsymbol{x}^N = \{x_m^N\}_{m=1}^{M}$, where $x_m^N \in \{0, 1\}$.

In the Bayesian framework, given $\boldsymbol{y}$, we are interested in finding the predictive posterior over the labels $\boldsymbol{x}^N$, which can be written as $P(\boldsymbol{x}^N|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{x}^N)P(\boldsymbol{x}^N)$. Here $P(\boldsymbol{y}|\boldsymbol{x}^N)$ is the observation (or likelihood) model and $P(\boldsymbol{x}^N)$ is the prior model on the labels. For vision applications, MRF has been a popular choice for modeling the prior over the labels. However, there are several disadvantages of using the MRF models [3]. In the standard MRF formulation, computation of exact Maximum a Posteriori (MAP) or Modes of Posterior Marginal (MPM) estimates is in general NP-hard, and the approximate estimates are expensive to compute. The parameter estimation in MRF is difficult due to the presence of the partition function. To alleviate these problems, in the present work we use a causal Multiscale Random Field (MSRF) as a prior model as proposed by Bouman and Shapiro [3] and further used by [5] for semantic image segmentation.

In a MSRF model, labels over an image are generated using Markov chains defined over coarse to fine scales. Such a hierarchical structure is also known as Tree-Structured Belief Network (TSBN) [5]. It can facilitate easy incorporation of long-range correlations in the image. We use the standard singly-connected quad-tree representation of MSRF to model the prior distribution over labels. One big advantage of such MSRF models is that the MAP/MPM inference is noniterative and time complexity is linear in the number of image sites. However, these models suffer from the nonstationarity of the induced random field, leading to 'blocky' smoothing of the image labels [5]. According to the overall image generative model, the image data $\boldsymbol{y}$ is generated from an underlying process $\boldsymbol{x}$, where $\boldsymbol{x}$ is a MSRF. For simplicity, a 1-D representation of the overall image generative model is given in figure 2 (a). The labels at N levels of the causal tree are denoted by $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N$ with $P(\boldsymbol{x}) = P(\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N)$. It can be noted that the observed image labels are nodes of the layer $\boldsymbol{x}^N$. In the MSRF model, the Markov assumption over scales implies $P(\boldsymbol{x}^n|\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{n-1}) = P(\boldsymbol{x}^n|\boldsymbol{x}^{n-1})$ for $n = 2, \ldots, N$. Further, from the conditional independence assumption for the directed graphs, $P(\boldsymbol{x}^n|\boldsymbol{x}^{n-1}) = \prod_{i \in S^n} P(x_i^n|z_i^{n-1})$, where x_i^n is i^{th} node at level n, z_i^{n-1} is its parent at level $(n-1)$, and S^n is the set containing all the nodes at level n. Each node in the MSRF model is a bernoulli variable in our case.

For the observation model, it is generally assumed that the data is conditionally independent given the class labels

[5][3]. However, this assumption is not correct for man-made structures, since the neighboring sites containing a man-made structure exhibit strong dependencies. In other words, the lines and edges at such spatially adjoining sites follow some underlying organization rules rather than being random. In this work, instead, we assume that given the class label x_m^N at site m, the data y_m is dependent only on its neighbors. This imposes a MRF-like noncausal dependency among the data $\boldsymbol{y}$, which is shown by undirected links in Figure 2 (a). Thus, the generative model has two random fields, one on the labels, and the other on the data given the labels. Hierarchical MRFs have been used to perform texture segmentation by defining separate MRFs over the texture labels and the data given the labels [19]. To avoid dealing with intractable true joint conditional $P(\boldsymbol{y}|\boldsymbol{x}^N)$, we assume a factored form of the observation model similar to [19] such that,

$$P(\boldsymbol{y}|\boldsymbol{x}^N) \approx \prod_{m \in S^N} P(y_m|y_{\omega_m}, x_m^N) \tag{1}$$

where ω_m is the neighborhood set of site m, and $y_{\omega_m} = \{y_{m'}|m' \in \omega_m\}$. The above approximation is known as Pseudo Likelihood (PL) in the MRF literature [11]. Thus, the overall generative model of the image can be expressed as,

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(x^1) \prod_{i \in S} P(x_i|z_i) \prod_{m \in S^N} P(y_m|y_{\omega_m}, x_m^N) \tag{2}$$

where S is the set containing all the nodes in the tree $\boldsymbol{x}$ except the root node x^1, and z_i is the parent of node x_i. To simplify the notations, we have denoted a generic node at any level of the tree by x_i, and its parent by z_i.

We further assume the field over the data $\boldsymbol{y}$ to be homogeneous, and approximate the conditional $P(y_m|y_{\omega_m}, x_m^N)$ by $p(f_m|x_m^N)$, where f_m is a multiscale feature vector which encodes the dependencies of data at site m given its neighbors. This approximation is driven by the fact that the conditional distribution $P(y_m|y_{\omega_m}, x_m^N)$ has a very limited power of structure description because it is about the datum on a single site m given the neighborhood [19]. In [19], the authors used the distribution of the data contained in the neighborhood of site m to approximate the conditional distribution in the context of texture segmentation. For our application, this issue becomes even more important as we need a rich representation of the data for man-made structure detection, which is inherently contained over multiple scales. Generic texture features have been shown to be inadequate due to wide variations in the appearance of man-made structures [8]. The need for rich data representation becomes crucial in the case of limited training data. The idea of multiscale feature vector is similar to the concept of parent vector defined by De Bonet [2], with the distinction that we compute features at a particular site by varying the size of the window around it so that the dependencies on the neighbors could be encoded explicitly. This kind of scale is also known as integration or artificial scale in the vision literature. Using the above assumptions, we can now approximate the overall image generative model as given in Figure 2 (b). Note that the original observation layer $\boldsymbol{y}$ has been replaced by a multiscale observation layer $\boldsymbol{f}$. The topology of the approximated generative model is similar to the one used in [3], and the benefits of that model in terms of exact noniterative inference can now be reaped.

Finally, exploiting the assumption of homogeneity, the likelihood of the multiscale feature vector was modeled using a Gaussian Mixture Model (GMM) for each class, as $P(f_m|x_m^N) = \sum_{\gamma=1}^{\Gamma} P(f_m|x_m^N, \gamma)P(\gamma|x_m^N)$, where $P(f_m|x_m^N, \gamma) \sim \mathcal{N}(\mu_\gamma, \Sigma_\gamma)$, μ_γ is the mean and Σ_γ is the covariance of the γ^{th} Gaussian, and Γ is the total number of Gaussians in GMM.

2.1. Parameter Estimation

The full image generative model has two different sets of parameters: Θ_p in the prior model, and Θ_o in the observation model. The observation model parameters consist of mean and covariance matrices of the Gaussians, which are estimated through standard maximum likelihood formulation for GMM using Expectation Maximization (EM). The prior model parameter set consists of conditional transition probabilities over different links in the tree, and the prior probabilities over the root node. Let θ_{ikl} be the transition probability for node $i \in S$, defined as, $\theta_{ikl} = P(x_i = l|z_i = k)$, with the constraint $\sum_l \theta_{ikl} = 1$, where $k, l \in \{0, 1\}$. It simply defines the conditional distribution at i^{th} node in MSRF given the label of its parent in the previous layer. The prior model parameters were learned using the Maximum Likelihood (ML) approach [5] by maximizing the probability of the labeled training images as,

$$\widehat{\Theta}_p^{ML} = \arg\max_{\Theta_p} \prod_{t=1}^{T} P(\boldsymbol{x}^{Nt}, \boldsymbol{y}^t|\Theta_p, \Theta_o)$$

where t indexes over the training images, and T is the total number of training images. Assuming the observation model to be fixed, the ML estimate of Θ_p is simply obtained using the labeled images $\boldsymbol{x}^{Nt}$ as, $\widehat{\Theta}_p^{ML} = \arg\max_{\Theta_p} \prod_{t=1}^{T} P(\boldsymbol{x}^{Nt}|\Theta_p)$. This maximization is carried out using EM, where all the nodes of MSRF from root to level $(N-1)$ are interpreted as the hidden variables. Denoting the hidden variables by $\boldsymbol{x}_h = \{\boldsymbol{x} \backslash \boldsymbol{x}^N\}$, in the E-step the lower bound is computed for the likelihood function at the current estimate of the parameters Θ_p' as the following expectation:

$$Q(\Theta_p, \Theta_p') = \sum_{t=1}^{T} \boldsymbol{E}_{\boldsymbol{x}_h^t|\Theta_p'} \left[\log P(\boldsymbol{x}_h^t, \boldsymbol{x}^{Nt}|\Theta_p)\right]$$

Computing the lower bound simply amounts to estimating the posterior probabilities over each parent-child pair,

$$P(x_i^t{=}l, z_i^t{=}k|\boldsymbol{x}^{Nt}, \Theta'_p) = \frac{\lambda(x_i^t{=}l)\ \theta'_{ikl}\ \pi(z_i^t{=}k)}{\sum_{k'}\pi(z_i^t{=}k')\lambda(z_i^t{=}k')} \prod_{u\in U(x_i^t)} \lambda_u(z_i^t{=}k) \quad (3)$$

where $U(x_i)$ is the set containing all the siblings of x_i, $\lambda(x_i)$ is the λ-value at node x_i, $\pi(z_i)$ is the π-value at node z_i, and $\lambda_u(z_i)$ is the λ-message sent from node u to z_i. All these notations are the same as defined in [16] in the context of belief propagation on singly-connected causal trees. In the M-step, new parameter values are obtained by maximizing the bound. In the case of limited training data, computing a different θ_{ikl} for each link is not practical. Thus, all the θ_{ikl} at each level n were forced to be the same as suggested in [5], and denoted as θ_{nkl}. Maximizing the bound defined above, subject to the constraint $\sum_l \theta_{nkl}{=}1$ yields for level n,

$$\theta_{nkl} = \frac{\sum_{t=1}^{T}\sum_{x_i\in S^n} P(x_i^t{=}l, z_i^t{=}k|\boldsymbol{x}^{Nt}, \Theta'_p)}{\sum_{t=1}^{T}\sum_{x_i\in S^n}\sum_{l'} P(x_i^t{=}l', z_i^t{=}k|\boldsymbol{x}^{Nt}, \Theta'_p)} \quad (4)$$

The prior probabilities over the root node are simply given by the belief at that node obtained through λ-π message passing scheme of Pearl [16].

2.2. Inference

Given a new test image $\boldsymbol{y}$, the aim is to find the optimal class labels over the image sites where the optimality is evaluated with respect to a particular cost function. The MAP estimate can be excessively conservative since it maximizes the probability that *all* the sites in the image are correctly classified [3]. In the present work, the labels are obtained through Maximum Posterior Marginals (MPM) such that the optimal labels maximize $P(x_m^N|\boldsymbol{f})$ for $m = 1, \cdots, M$. This can be achieved noniteratively by computing the belief at each node of the tree at level N using Pearl's λ-π message passing scheme [16] in one upward and one downward pass over the tree.

To summarize, we have proposed MSRF based image generative model that takes into account the spatial dependencies of not only the class labels but also the observed data. After making some common approximations, learning of the model parameters and inference over the model can be carried out using efficient techniques.

3. Feature Set Description

The choice of appropriate features without relying on ad hoc heuristics is important for a generic structure detection system. On the other hand, given a small training set, task dependent feature extraction becomes unavoidable to efficiently encode the relevant task information in a limited number of features. There is currently no formal solution to deriving optimal task-dependent features. In this section, we propose a set of multiscale features that captures the general statistical properties of the man-made structures over spatially adjoining sites.

For each site in the image, we compute the features at multiple scales, which capture intrascale as well as interscale dependencies. The multiscale feature vector at site m is then given as: $f_m = \left[\{f_m^j\}_{j=1}^J, \{f_m^\rho\}_{\rho=1}^R\right]$ where, f_m^j is j^{th} intrascale feature and f_m^ρ is ρ^{th} interscale feature.

3.1. 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 m, 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 [1]. 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,

$$E'_\delta = \frac{\sum_{i=1}^{\Delta} K((\delta - i)/h)E_i}{\sum_{i=1}^{\Delta} K((\delta - i)/h)} \quad (5)$$

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} or-

der moment ν_p is given as,

$$\nu_p = \frac{\sum_{\delta=1}^{\Delta}(E'_\delta - \nu_0)^{p+1} H(E'_\delta - \nu_0)}{\sum_{\delta=1}^{\Delta}(E'_\delta - \nu_0) H(E'_\delta - \nu_0)} \qquad (6)$$

where $H(x)$ is the unit step function such that $H(x) = 1$ for $x > 0$, and 0, otherwise. The moment computation in Eq. (6) 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 [9]. 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.

3.2. 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.

4. Experimental Results

The proposed detection scheme was trained and tested on two different datasets drawn randomly from the Corel Photo Stock. The training set consisted of 108 images while the testing set contained 129 images, each of size 256×384 pixels. Most of the images in both the datasets contained both natural objects and man-made structures captured at medium to long distances from a ground-level camera. The ground truth was generated by hand-labeling each nonoverlapping 16×16 pixels block in each image as a *structured* or *nonstructured* block. This kind of coarse labeling was sufficient for our purpose as we were interested in finding the location of the *structured* blocks without explicitly delineating the object boundary. However, the block quantization introduces noise in the labels of the blocks lying on the object boundary, since a block containing a small part of the structure could be given either of the labels. This makes the quantitative evaluation of the results hard and there is no formal solution to this problem. To circumvent this, we do not count as false positive a misclassification that is adjacent to a block with ground truth label *structured*. In practice, small classification variations at the object boundary do not affect future processing such as grouping blocks into connected regions or extracting bounding boxes. The whole training set contained $36,269$ blocks from the *nonstructured* class, and $3,004$ blocks from the *structured* class.

To train the generative model, a multiscale feature vector was computed for each nonoverlapping 16×16 pixels block in the training images. One of the reasons for choosing this block size is related to the fundamental ambiguity in the structure detection task. If the *structure* is too far, it will become like 'texture', and if it is too near, only a small portion (e.g., a long edge or a smooth patch from a wall) will occupy almost the whole image. The lowest and the highest scales for the feature extraction were chosen to constrain this ambiguity. We are interested in the *structures* which are not smaller than the lowest scale, and are not totally smooth or contain only unidirectional edges at the highest scale. For multiscale feature computation, the number of scales 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. The largest scale implicitly defines the neighborhood ω_m defined in Eq. (1) over which the data dependencies are captured.

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 m, a fourteen component multiscale feature vector f_m was obtained. We used only a limited number of features due to the lack of sufficient training data to reliably

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

Figure 3. The learned parameters for the 2-class, 5-level MSRF model. The brighter intensity indicates a higher probability. (a) Prior probabilities at the root node (right block indicates the *structured* class), (b) through (e) transition probability matrices for the links between adjacent levels starting from level 1 to level 5 (top left block indicates the transition from *structured* to *structured* class).

estimate the GMM parameters. Each feature was normalized linearly over the training set between zero and one for numerical reasons.

To learn the parameters of the MSRF model (Θ_p), a quad-tree was constructed considering each 16×16 pixels nonoverlapping block in the image to be a node at the leaf level N. This arrangement resulted in 16×24 nodes at the leaf level and five levels ($N = 5$) in the tree. To take into account the $2 : 3$ aspect ratio of the images, we modified the quad-tree as suggested in [5] such that the root node had six children. Since we had assumed the conditional transition probability to be the same for each link within a level, we needed to estimate four transition probability matrices, θ_{nkl}, and the prior probability distribution over the root node. For the ML learning described in section 2.1, the parameter values were initialized by building the empirical trees over the image labels in the training images using the max-voting over the nodes. The training took 8 iterations to converge in 773 s in Matlab 6.5 on a 1.5 GHz Pentium class machine. The learned parameters are shown in Figure 3. The brighter intensity indicates a higher probability. It can be noted that for finer levels, the diagonal probabilities are dominant indicating high probabilities of transition to the same class. The transition matrix between level 1 and level 2 shows a more random transition due to the mixing of blocks at coarser levels. Finally, the prior probability distribution at the root node highly favors the root node to be from the *nonstructured* class. This is reasonable since most of the images have much lesser *structured* blocks compared to the *nonstructured* blocks. For the GMM based observation model, the number of Gaussians in the mixture model was selected to be 8 using cross-validation. The mean vectors, full covariance matrices and the mixing parameters were learned using the standard EM technique.

4.1. Performance Evaluation

In this section we present a qualitative as well as quantitative evaluation of the proposed detection scheme. First

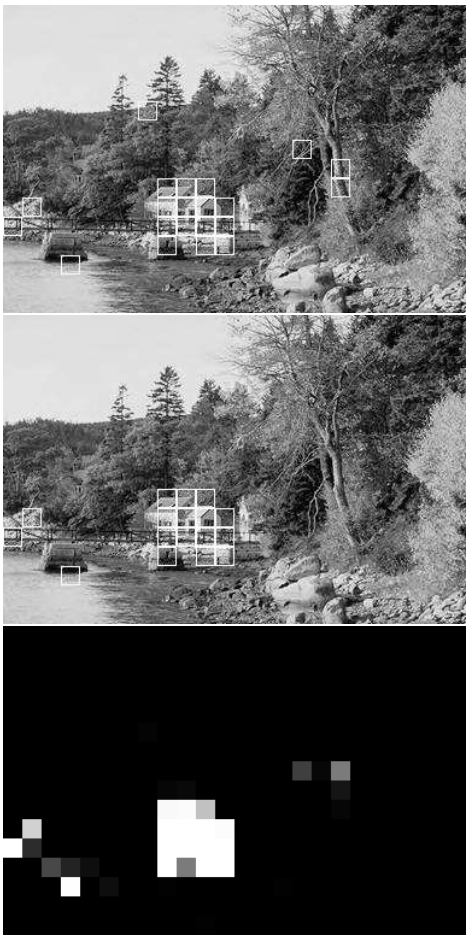
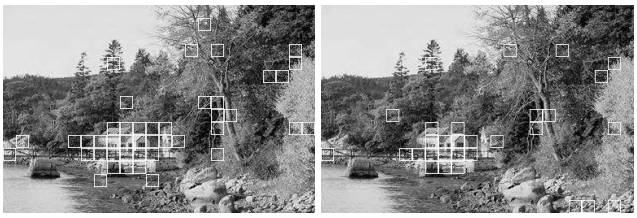


Figure 4. The structure detection results for the input image given in Figure 1 (a). Top: Maximum likelihood results using only GMM. Middle: MPM results using MSRF model. Bottom: The MSRF posterior map displaying the posterior marginals over the image blocks for the *structured* class. The brighter intensity indicates a higher probability.



(a) (b)

Figure 5. The structure detection results using (a) SC, (b) SVM. Both techniques have higher number of false positives in comparison to the MSRF result for a similar detection rate.

	NS	S
NS	42976	188
S	1776	4596

(a) MSRF

	NS	S
NS	42791	373
S	1776	4596

(b) GMM

	NS	S
NS	42587	577
S	1777	4595

(c) SC

	NS	S
NS	42534	630
S	2004	4368

(d) SVM

Figure 6. Confusion matrices for different techniques. S - *structured*, and NS - *nonstructured*. The detection rate was kept nearly the same for all the techniques. The rows contain the ground truth while the columns contain the detection results.

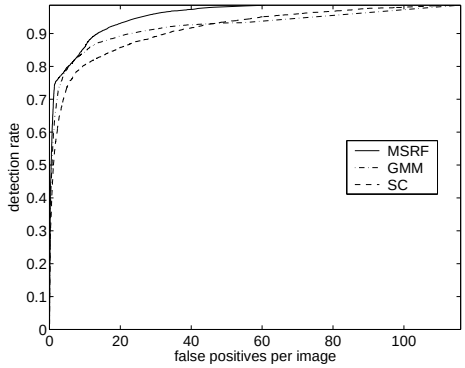


Figure 7. ROC curves for MSRF, GMM, and SC techniques

we compare the detection results on the test images using two different methods: only GMM ($i.e.$ no prior model over the labels) with maximum likelihood inference, and GMM in addition to MSRF prior with MPM inference. For convenience, the former will be referred to as the GMM and the latter as the MSRF model in the rest of the paper. The same set of learned parameters was used in GMM for both the methods. For the input image given in Figure 1 (a), the *structure* detection results from the two methods are given in Figure 4. The blocks identified as *structured* have been shown enclosed within an artificial boundary. It can be noted that for the same detection rate, the number of false positives have significantly reduced for the MSRF based detection. The MSRF model tends to smooth the labels in the image and removes most of the isolated false positives. The bottom image in Figure 4 shows the MSRF posterior map over the input image for the *structured* class, displaying the posterior marginals for each image block. The posterior map exhibits high probability for the *structured* blocks, and the number of *nonstructured* blocks with significant probability is very low. This shows that the MSRF based technique is making fairly confident predictions.

We compare the above results with the results from two other popular classification techniques: Support Vector Machine (SVM) and Sparse Classifier (SC). A Bayesian learning of sparse classifiers was proposed recently by Figueiredo and Jain [6], who have shown good results on the standard machine learning databases. Both classifiers used the multiscale feature vectors defined earlier as the data associated with the image blocks. We implemented a kernel classifier using a symmetric Gaussian kernel of bandwidth 0.1 for both SVM and SC. The cost parameter for SVM was set to be 1000 from cross-validation. The number of support vectors in SVM were found to be 2305, while the number of sparse relevance vectors in SC were 66. The detection results for these two techniques are shown in Figure 5. The results from SC were based on the MAP inference. It can be seen that the detection rate in the image is fairly good for both the techniques. This demonstrates that the multiscale features capture relevant data dependencies for the structure detection. However, the number of false positives for both techniques is significantly higher than that from the MSRF model. Similar to the GMM , SVM and SC do not enforce the smoothness in the labels, which led to increased false positives. 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 2.8 s for MSRF, 2.3 s for GMM, 2.3 s for SC, and 2.8 s for SVM.

To carry out the quantitative evaluation of our work, we first computed the block wise classification accuracy over all the test images. We obtained 94.6% classification accuracy for the $49,536$ blocks contained in 129 test images. However, the classification accuracy is not a very informative criterion here as the number of *nonstructured* blocks $(43,164)$ is much higher than the number of *structured* blocks $(6,372)$, and a high classification accuracy can be obtained even by classifying every block to the *nonstructured* class. Hence, we computed two-class confusion matrices for each technique. The confusion matrix for the MSRF model is given in Figure 6 (a). For an overall detection rate of 72.13%, the false positive rate was 0.43% or 1.46 false positives per image. The main reason for a relatively low detection rate is that the algorithm fails to detect the *structured* blocks that are part of the smooth roofs or walls that have no significant gradients even at larger scales. In fact, it is almost impossible to differentiate these blocks from the smooth blocks contained in natural regions (e.g. sky, land) using any technique without exploiting other auxiliary information such as color. Similarly, too small structures and bad illumination contrast in natural images also make the detection hard. However, it should be noted that this is a significant detection rate at the block level given a low false positive rate. In general we do not require all the

blocks of an structured object to be detected since one could use other postprocessing techniques such as color based region-growing to detect the missing blocks of an object.

Keeping the same detection rate as from the MSRF model, we obtain confusion matrices for the GMM and SC. Since SVM does not output probabilities, we varied the cost parameter to obtain the closest possible detection rate. The confusion matrices are given in Figure 6. The average false positives per image for the GMM, SC and SVM are 2.89, 4.47, and 4.88 respectively. The best among these three gives almost twice false positives per image in comparison to the MSRF model. The results from SVM and SC are quite similar with SC having a slight advantage, since the SVM detection rate is 68.55% in comparison to 72.13% of SC for comparable false positives. For a more complete comparison of the detection performance of the MSRF, GMM, and SC techniques, the corresponding ROC curves are shown in Figure 7. The MSRF model performs better than the other two techniques. The GMM performs better than the SC most of the times for our test set. For the regions of low false positive per image (less than 2), the performance of MSRF model is significantly better than the other two techniques.

5. Conclusions

We have presented a technique for man-made structure detection in natural images using a causal MSRF. The proposed generative model captures spatial dependencies of the labels as well as the observed data to yield good results in real-world images. The empirical results support the effectiveness of the proposed multiscale features in capturing neighborhood relationships of the structured objects. However, the price to pay for using a multiscale representation is somewhat degraded localization at the object boundaries. In the future, it will be interesting to explore more powerful models to capture the dependencies in the data by relaxing some of the statistical assumptions made in this paper, and their relation with the prior model over the labels. Finally, beyond the task of structure detection used as a basis of discussion in this paper, the proposed model may potentially be used in many vision tasks in which spatial consistency of class labels as well as the observed data needs to be enforced. Such tasks include object detection, image segmentation, and domain specific image analysis.

Acknowledgments

This work was supported in part by ARL under the Collaborative Technology Alliance Program DAAD19-01-2-0012 and by DARPA NBCH1020014. Our thanks to Jonas August and Tom Minka for helpful discussions.

References

[1] W. A. Barrett and K. D. Petersen. 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, 2001.

[2] J. S. D. Bonet and P. Viola. A non-parametric multi-scale statistical model for natural images. *In Advances in Neural Information Processing*, 10, 1997.

[3] C. A. Bouman and M. Shapiro. A multiscale random field model for bayesian image segmentation. *IEEE Trans. on Image Processing*, 3(2):162–177, 1994.

[4] B. Bradshaw, B. Scholkopf, and J. C. Platt. *Kernel Methods for Extracting Local Image Semantics*. Tech. Report MSR-TR-2001-99, Microsoft Research, 2001.

[5] X. Feng, C. K. I. Williams, and S. N. Felderhof. Combining belief networks and neural networks for scene segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(4):467–483, 2002.

[6] M. A. T. Figueiredo and A. K. Jain. Bayesian learning of sparse classifiers. *In Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition*, 1:35–41, 2001.

[7] Q. Iqbal and J. K. Aggarwal. Applying perceptual grouping to content-based image retrieval: Building images. *In Proc. IEEE Int. Conf. on CVPR*, 1:42–48, 1999.

[8] S. Konishi and A. L. Yuille. Statistical cues for domain specific image segmentation with performance analysis. *In Proc. IEEE Int. Conf. CVPR*, pages 125–132, 2000.

[9] S. Krishnamachari and R. Chellappa. Delineating buildings by grouping lines with mrfs. *IEEE Trans. on Pat. Anal. Mach. Intell.*, 5(1):164–168, 1996.

[10] M. S. Langer. Large-scale failures of $f^{-\alpha}$ scaling in natural image spectra. *Journal of Optical Society of America*, 17(1):28–33, 2000.

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

[12] C. Lin and R. Nevatia. Building detection and description from a single intensity image. *Computer Vision and Image Understanding*, 72:101–121, 1998.

[13] H. Mayer. Automatic object extraction from aerial imagery-a survey focusing on buildings. *Computer Vision and Image Understanding*, 74(2):138–149, 1999.

[14] A. Oliva and A. Torralba. The shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[15] A. Olmos and E. Trucco. Detecting man-made objects in unconstrained subsea videos. *In Proc. British Machine Vision Conference*, pages 517–526, 2002.

[16] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[17] S. Sarkar and P. Soundararajan. Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata. *IEEE Trans. on Pat. Anal. Mach. Intell.*, 22(5):504–525, 2000.

[18] A. Vailaya, A. K. Jain, and H. J. Zhang. On image classification: City images vs. landscapes. *Pattern Recognition*, 31:1921–1936, 1998.

[19] C. S. Won and H. Derin. Unsupervised segmentation of noisy and textured images using markov random fields. *CVGIP*, 54:308–328, 1992.