# Double Markov Random Fields and Bayesian Image Segmentation

Dina E. Melas and Simon P. Wilson

***Abstract*—Markov random fields are used extensively in model-based approaches to image segmentation and, under the Bayesian paradigm, are implemented through Markov chain Monte Carlo (MCMC) methods. In this paper, we describe a class of such models (the double Markov random field) for images composed of several textures, which we consider to be the natural hierarchical model for such a task. We show how several of the Bayesian approaches in the literature can be viewed as modifications of this model, made in order to make MCMC implementation possible. From a simulation study, conclusions are made concerning the performance of these modified models.**

***Index Terms*—Bayesian statistics, hierarchical model, image segmentation, Markov random field, remote sensing.**

## I. Introduction

THE MARKOV random field has been used in many model-based solutions to image analysis problems, including that of image segmentation. In image segmentation, a digital image is to be divided into regions that are deemed to possess similar local properties, which, here, are taken to be texture. Applications include

- land-use estimation from satellite images;
- computer-aided medical diagnosis;
- content-based image retrieval;
- image compression;
- recovery of shape information from an image.

We consider so-called supervised and semi-unsupervised segmentation, where the number of texture classes in the image is known but information about their properties is either known or unknown, respectively. In a Bayesian approach, the goal is to infer the posterior distribution of possible segmentations and, where necessary, any unknown model parameters. This approach is implemented through Markov chain Monte Carlo (MCMC) methods, usually the Gibbs sampler. Although computationally more expensive than many other approaches, the MCMC approach has the advantage that the analysis also yields, through the posterior distribution, information on uncertainty in the segmentation and properties of the texture classes.

There are three objectives in this paper. The first is to describe the general notion of a double Markov random field model for images composed of regions of different texture. Our contention is that this represents a natural model for the task of segmentation. Our second objective is to show that many of the Markov random field models used for segmentation can be viewed as adaptations of this model so that MCMC can be used. Five such general adaptations are described. The final objective is to compare, by a simulation study, the performance of these adaptations. We conclude that one of the models seems to perform better overall than the others, and this is applied to segment a satellite image.

The paper is organized as follows. Section II defines the double Markov random field and its application to Bayesian image segmentation. Section III describes five modifications to the model to enable segmentation by MCMC. Section IV is a comparison of these modifications by a simulation study, and Section V applies the most successful model to land-use estimation from satellite radar images. Section VI completes the paper with some concluding remarks.

## II. Double Markov Random Field

Consider a rectangular lattice of pixel sites $\mathcal{S}$. An image consists of an array of grey values $(x_s)_{s\in\mathcal{S}}$ and labels $(y_s)_{s\in\mathcal{S}}$, identifying the texture type present. We assume that there are $R$ textures in the image and that each texture, defined on all of $\mathcal{S}$, is a Markov random field $T^r$, parameterized by $\theta_r$, with neighborhood system having a set of cliques $\mathcal{C}_r$. The label process is another Markov random field $Y$, parameterized by $\beta$ and with neighborhood system represented by the set of cliques $\mathcal{C}_Y$. All the fields are independent, conditional on model parameters, and their distributions have the following Gibbs representation:

$$P(T^r = t|\theta_r) = \frac{\exp\left[-\sum_{c\in\mathcal{C}_r} V_{r,c}(t;\,\theta_r)\right]}{Z_r(\theta_r)} \tag{1}$$

$$P(Y = y|\beta) = \frac{\exp\left[-\sum_{c\in\mathcal{C}_Y} V_{Y,c}(y;\,\beta)\right]}{Z_Y(\beta)} \tag{2}$$

where $V_{r,c}(t;\,\theta_r)$ and $V_{Y,c}(y;\,\beta)$ are the clique potentials, and $Z_r$ and $Z_Y$ are the partition functions. The observed image is

Manuscript received December 7, 2000; revised October 9, 2001. This work was supported by the European Science Foundation Program on Highly Structured Stochastic Systems and also forms part of the work under the European Union Research Network MOUMIR. This paper was partly written while S. P. Wilson was at Projet Ariana, INRIA-Sophia Antipolis, France. The associate editor coordinating the review of this paper and approving it for publication was Prof. Simon J. Godsill.

D. E. Melas is with Interoperability Systems International, Athens, Greece (e-mail: dina@isihellas.com).

S. P. Wilson is with the Department of Statistics, Trinity College, Dublin, Ireland (e-mail: simon.wilson@tcd.ie).

Publisher Item Identifier S 1053-587X(02)00560-3.

the collage $X = (T_s^{Y_s})_{s \in \mathcal{S}}$, and the joint distribution for $X$ and $Y$ is

$$P(X = x, Y = y|\theta_1, \ldots, \theta_R, \beta) = P(Y = y|\beta) \prod_{r=1}^{R} P\left(T_{S_r}^r = x_{S_r}|\theta_r\right) \quad (3)$$

where $S_r = \{s \in \mathcal{S} : y_s = r\}$ and $T_{S_r}^r$, $x_{S_r}$ denote $T^r$ and $x$ restricted to $S_r$.

In order to evaluate $P(T_{S_r}^r = x_{S_r}|\theta_r)$, it is necessary to identify the interaction structure of pixels that have missing neighbors. This is generally intractable, leading to the use of various boundary assumptions. The marginal distribution on the regions is then approximated by the joint distribution of the sites in $S_r$ conditioned on the assumed boundary values. One common boundary approximation is the free boundary, where pixels on the edges of the region have fewer neighbors than those in the interior. Another possibility is a fixed boundary, where the missing neighboring values of the pixels on the edges are replaced by arbitrary fixed values. The often-used toroidal structure is not applicable in this case because the sets of pixels $S_r$ are not necessarily rectangular. From (1), under a free boundary structure, we would have

$$P\left(T_{S_r}^r = x_{S_r}|\theta_r\right) = \frac{\exp\left[-\sum_{c \in \mathcal{C}_r : c \subset S_r} V_{r,c}(x_c; \theta_r)\right]}{Z_{S_r}(\theta_r)}. \quad (4)$$

Equations (2)–(4) define the double Markov random field: a phrase that was, to our knowledge, first used in [1]. Inference is now based on the posterior density

$$P(y, \theta_1, \ldots, \theta_R, \beta|x) \propto P(X = x, Y = y|\theta_1, \ldots, \theta_R, \beta)\pi(\theta_1, \ldots, \theta_R, \beta) \quad (5)$$

in the semi-supervised case for a prior distribution $\pi$ or $P(y|\theta_1, \ldots, \theta_R, \beta, x)$ in the supervised case. A best single segmentation $y^*$ is then selected from the posterior distribution. Two are considered in the literature: the maximum *a posteriori* (MAP) $y^* = \arg\max_y P(Y = y|x)$, or the marginal posterior mode $y_s^* = \arg\max_{y_s} P(Y_s = y_s|x)$. These are motivated by decision theory; they are the segmentations that minimize posterior expected loss, when the loss function is 0–1 (for MAP) or the number of misclassified pixels (for MPM). The former is found by stochastic maximization—usually simulated annealing in tandem with Gibbs sampling (see [2]—and the latter by Gibbs sampling of the posterior, and picking the most frequently generated label at each site after convergence is deemed to have occurred (see, for a recent example, [3]).

## III. Modifying the Double Markov Random Field Model

Although (3) represents our ideal model, it cannot be readily used because the partition functions $Z_{S_r}(\theta_r)$ cannot be computed. Even if one were to approximate them, their dependence on the labels would not usually have the convenient local interaction structure introduced by the Markov property in the label field model. This means that Gibbs sampling directly from (5) is not possible. As a result, several computationally feasible models have been proposed that are in the same spirit but allow Gibbs sampling. In this section, we specify five such models. They can be viewed as approximations to the general model of (2)–(4) made to resolve two issues: dependence of the partition functions on the labels and the assumed boundary structure between regions.

The first, which we call Model I, recovers the local interaction structure on the labels by redefining the texture models to be causal, thus creating partition functions of a simple form. The other four modify (3) to define posterior full conditional distributions on $Y$ that admit a Gibbs formulation. Model II uses noncausal texture models, but the partition function is ignored. In Models III–V, local noncausal texture models are combined in order to simplify the dependence of the partition function on the labels. These five models cover many of the Markov random field approaches that have been proposed in the literature.

In Models I, III, IV, and V, we will see that the modifications to the original double Markov random field have the effect of introducing an external field to the prior model for the labels. The effect in Model II is less clear, but it can be interpreted as placing a prior on parameters that is proportional to the partition functions. We also note that only in Model I does the full conditional equation define a probability model; the others do not give a consistent model and should be seen as an approximation to the true full conditionals of the double Markov random field.

At the end of the section, we discuss an auxiliary variable approach that would directly sample the posterior from the double Markov random field model but at a potentially large increase in computation.

### A. Model I: Causal Texture Model

The $T^r$ are modeled as causal Gaussian autoregressive (AR) processes. Assume that sites in $\mathcal{S}$ are labeled lexicographically. Then

$$T_s^r = \mu_r + \sum_{j:\, \langle s, j \rangle} \phi_{r_{sj}}(T_j^r - \mu_r) + \sigma_r \epsilon_s^r \quad (6)$$

$$\phi_{r_{sj}} = \begin{cases} \phi_{r_k}, & \text{if } \langle s, j \rangle_k,\ s > j \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

for $r = 1, \ldots, R$, where $\langle s, j \rangle$ means $s$ and $j$ are neighbors, $\langle s, j \rangle_k$ means neighbors of clique type $k$ (vertically, horizontally, or diagonally neighboring, etc.), $\sigma_r$ is the standard deviation in the texture, and the $\epsilon_s^r$ are independent standard Gaussian errors. Condition $s > j$ imposes the causality. Fig. 1 shows a realization of a causal AR process with the second-order neighborhood system and a strong horizontal correlation, along with the Gaussian Markov random field with the same parameters for comparison. We see that the two are similar, but the MRF has a better defined texture. Simulations with other parameter values have supported this observation.

For segmentation, the outcome is that the posterior of $Y$ maintains a local interaction structure and can be simulated with

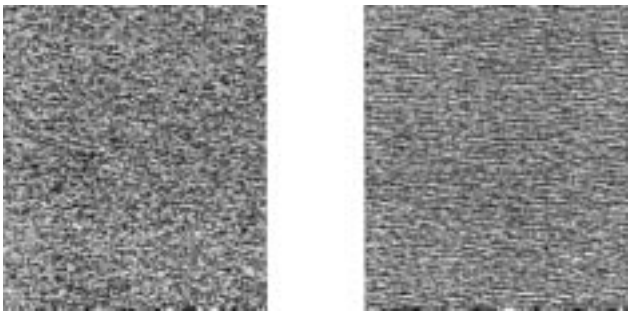

Fig. 1. Realization of a causal Gaussian AR process on the left, with a realization of the Gaussian Markov random field having the same parameters on the right.

MCMC. Under a first-order Potts model for $Y$, the log full conditionals are, up to an additive constant

$$\begin{aligned}\log(P(Y_s = r|x,\, y_j,\, j \neq s,\, \theta_r,\, \beta)) &= -\tfrac{1}{2}\log(2\pi\sigma_r^2) \\ &\quad - \frac{1}{2\sigma_r^2}\left\{x_s - \mu_r - \sum_{k=1}^{K}\sum_{\substack{j:\,\langle x,\, j\rangle_k \\ s>j}} \phi_{r_k}(x_j - \mu_r)\right\}^2 \\ &\quad + \beta\sum_{k=1}^{2}\sum_{j:\,\langle s,\, j\rangle_k} \delta(r - y_j) \end{aligned} \tag{8}$$

where $\delta(u) = 1$ if $u = 0$, and is 0 otherwise, $K$ is the number of different clique types in the neighborhood system, and the types $k = 1$ and $k = 2$ are the first-order doubleton cliques (horizontally and vertically neighboring pairs). Although causal models have computational advantages, the directionality implied in the definition means that they have less discriminating power than a Markov random field. However, in supervised segmentation, the model will still perform well in many situations, as we demonstrate later in Section IV. An early discussion of such causal models is in [4]. This model is an example of a Markov mesh model, further examples of which are in [5].

### B. Model II: Ignore the Partition Function Term

Model II is defined by ignoring the partition function terms $Z_{S_r}(\theta_r)$ in (4). Doing this, we obtain the model of [6], which is used as a template model to divide an image into two regions. Again, a Markov structure on the posterior of $Y$ is recovered. When a Potts model is assumed for $Y$ and Gaussian conditional autoregressive (CAR) for the $T^r$, that is, the $T^r$ follow the noncausal AR model $T_s^r = \mu_r + \sum_{j:\,\langle s,\, j\rangle} \phi_{r_{sj}}(T_j^r - \mu_r) + \sigma_r \epsilon_s^r$, where $\phi_{r_{sj}} = \phi_{r_k}$ if $\langle s,\, j\rangle_k$, we obtain the following log full conditional, up to an additive constant

$$\begin{aligned}\log(P(Y_s = r|x,\, y_j,\, j \neq s,\, \theta_r,\, \beta)) &= -\frac{(x_s - \mu_r)^2}{2\sigma_r^2} \\ &\quad + \sum_{k=1}^{K}\sum_{j:\,\langle s,\, j\rangle_k}\left(\frac{(x_s - \mu_r)\phi_{r_k}(x_j - \mu_r)}{\sigma_r^2}\right)\delta(r - y_j) \\ &\quad + \sum_{k=1}^{K}\sum_{\substack{t=1 \\ t\neq r}}^{R}\sum_{j:\,\langle s,\, j\rangle_k}\left(\frac{(x_s - \mu_r)\phi_{r_k}(x_j - \mu_r)}{\sigma_r^2} + \frac{(x_s - \mu_t)\phi_{t_k}(x_j - \mu_t)}{\sigma_t^2}\right)\delta(t - y_j) \\ &\quad + \beta\sum_{k=1}^{2}\sum_{j:\,\langle s,\, j\rangle_k}\delta(r - y_j). \end{aligned} \tag{9}$$

We observe that this would be the posterior obtained from the double Markov random field, were the priors on the parameters to be proportional to the partition functions. However, there seems no argument, other than computational convenience, why one should specify such a prior.

### C. Model III: Overlapping Window

Another possibility is to consider the grey levels to be composed of a set of overlapping square windows $(W_s)_{s\in\mathcal{S}}$ of size $n \times n$ centered at $s$, whose values are those of the corresponding window in $X$. Within each window, the texture $Y_s$ is assumed, and the windows are assumed independent given $Y$. This reduces (4) to

$$P\left(T_{S_r}^r = x_{S_r}|\theta_r\right) = \prod_{s\in S_r}\frac{\exp\left[-\sum_{c\in\mathcal{C}_r:\, c\subset W_s} V_{r,\,c}(x_c|\theta_r)\right]}{Z_r(\theta_r)} \tag{10}$$

where the partition function $Z_r(\theta_r)$ is that over $W_s$. In the case of a CAR model with a toroidal boundary assumption, $Z_r(\theta_r)$ would be the normalizing constant of an $n^2$-dimensional multivariate normal, with mean and variance structure specified by $\theta_r$; see [7] for the form of the covariance matrix as a function of $\theta_r$. Thus, $Z_r(\theta_r)$ is a function of the determinant of a $n^2 \times n^2$ matrix, and readily computable for small $n$, although the computational effort grows quickly with $n$. In supervised segmentation, we need compute these only once (one for each texture), whereas in semi-unsupervised segmentation, they would have to be recomputed at each MCMC iteration. This model is used in [8]–[10], with windows of size $3 \times 3$; a similar model is presented in [11].

For Gaussian CAR texture models and a Potts model for the labels, and with fixed boundary conditions, the log full conditionals of the posterior of $Y$ are, up to an additive constant

$$\begin{aligned}\log(P(y_s = r|x,\, y_j,\, j \neq s,\, \theta_r,\, \beta)) &= -\log(Z_r(\theta_r)) - \frac{1}{2\sigma_r^2}\sum_{j\in W_s}\left\{(x_j - \mu_r)^2\right. \\ &\quad \left. - \frac{1}{\sigma_r^2}\sum_{k=1}^{K}\sum_{\langle j,\, l\rangle_k}(x_j - \mu_r)\phi_{r_k}(x_l - \mu_r)\right\} \\ &\quad + \beta\sum_{k=1}^{2}\sum_{j:\,\langle s,\, j\rangle_k}\delta(r - y_j). \end{aligned} \tag{11}$$

Unfortunately, calculation of $Z_r(\theta_r)$ for other MRF models can still be a lengthy task, even for small $n$, and is only practical in supervised segmentation, where their calculation is only required once.

### D. Model IV: Use the Pseudo-Likelihood

Model III is susceptible to considerable boundary effects at the edges between different textures because a single texture model is assumed in each window. This effect increases with increasing window size. The minimum size of window that can be considered without losing textural information is one that contains the neighborhood of the central pixel. Model IV is then defined to be Model III but where the window size is the neighborhood of the central pixel. Thus

$$P\left(T^r_{S_r} = x_{S_r}|\theta_r\right) = \prod_{s\in S_r} \frac{\exp\left[-\sum_{c\in\mathcal{C}_r:\, s\in c} V_{r,\,c}\,(x_c;\theta_r)\right]}{Z_r}. \quad (12)$$

This model corresponds to using the pseudo-likelihood of the double Markov random field model for the posterior distribution, as defined in [12]. The pseudo-likelihood has also been used in a Bayesian approach in [13]. The boundary effect still exists in that at the boundary this term is a function of some grey levels that are from another texture. Under Gaussian CAR texture models and a Potts model for the labels, the full conditionals of the posterior of $Y$ are

$$\begin{aligned}\log(P(y_s = r|x,\, y_j,\, j\neq s,\, \theta_r,\, \beta)) &= -\tfrac{1}{2}\log(2\pi\sigma_r^2) \\ &\quad - \frac{\left\{x_s - \mu_r - \sum_{k=1}^{K}\sum_{j:\,\langle s,\,j\rangle_k}\phi_{r_k}(x_j-\mu_r)\right\}^2}{2\sigma_r^2} \\ &\quad + \beta\sum_{k=1}^{2}\sum_{j:\,\langle s,\,j\rangle_k}\delta(r-y_j). \end{aligned} \quad (13)$$

### E. Model V: Pseudo-Likelihood, but Ignore Grey Levels from a Different Texture

A modified version of (13) was used in [14], where only neighboring grey levels corresponding to pixels with the same label as pixel $s$ were included in the full conditional. Under Gaussian CAR textures and an Ising model prior for the labels, the full conditionals for the posterior of $Y$ are

$$\begin{aligned}\log(P(y_s = r|x,\, y_j,\, j\neq s,\, \theta_r,\, \beta)) &= -\tfrac{1}{2}\log(2\pi\sigma_r^2) \\ &\quad - \frac{\left\{x_s - \mu_r - \sum_{k=1}^{K}\sum_{\substack{j:\,\langle s,\,j\rangle_k \\ y_j = r}}\phi_{r_k}(x_j-\mu_r)\right\}^2}{2\sigma_r^2} \\ &\quad + \beta\sum_{k=1}^{2}\sum_{j:\,\langle s,\,j\rangle_k}\delta(r-y_j). \end{aligned} \quad (14)$$

### F. MCMC from the Double Markov Random Field

Since $P(X = x, Y = y|\theta_1, \ldots, \theta_R, \beta) \propto P(Y = y|\beta)\prod_{i=1}^{R} P(T^r|\theta_r)$, an MCMC scheme is then to sample the $T^r$ on the complement of $S_r$ given $T^r_{S_r} = x_{S_r}$, and then, sample $Y$ from the full conditional proportional to $P(Y = y|\beta)\prod_{i=1}^{R} P(T^r|\theta_r)$. In the supervised case, this would allow for the sampling from the double MRF posterior. For the Gaussian MRF case, such conditional distributions on the $T^r$ are multivariate Gaussian and can be calculated (see [15] for an efficient method). However, such an approach has its own problems. For texture classes with only a few pixels assigned, one is faced with simulating a realization of the texture across almost the entire image conditional on this small sample. Although not applied to image analysis, experience in [16] of simulation of Gaussian Markov random fields showed that convergence problems can easily arise. This would be particularly true in the semi-unsupervised case, where parameters from classes with only a few pixels assigned would not be well estimated. We do not pursue this idea further here.

### G. Estimation of Model Parameters

In a Gibbs sampling scheme to simulate from (5), in a semi-unsupervised approach, we also require the full conditionals of all model parameters, which are a function of the partition function. It is possible to approximate the partition function, but it is computationally expensive; see [17] for an example using the scheme of [18].

This approximation can only be practically evaluated for the texture parameters of Model I, where the dependence on the label parameter is from the partition function of the label model, and therefore, the partition function does not need to be re-evaluated at each iteration of the sampler. In Models IV and V, we can make an approximation to the full conditional by restricting the dependence of the parameters to terms in (12). With uniform prior distributions over some suitable range on all texture parameters, the full conditionals for $\theta_r$ are then proportional to the pseudo-likelihood function

$$PL(\theta_r;\, x,\, y,\, \beta) = \prod_{s\in\mathcal{S}_r} P(y_s = r|x,\, y_j,\, j\neq s,\, \theta_r,\, \beta). \quad (15)$$

In the case of Gaussian CAR parameters, the range of allowable values is determined in [19], and the full conditionals are given in [20, App. 2].

In all cases, where a Potts model is assumed for the labels, the full conditional of $\beta$ is not available. We either consider it fixed or sample from the pseudo-likelihood of the model; this latter case we call the *adaptive* algorithm.

We note that the pseudo-likelihood estimate of $\beta$ tends to overestimate its value in MPM segmentation, leaving it above the "critical" temperature for the Ising model of $\beta \approx 0.88$. Since values of $\beta$ above this value place most probability on segmentations with large regions of one class, this implies the possibility of oversmooth segmentations [21, ch. 5]. In MAP segmentation, $\beta$ is confounded with the temperature parameter; therefore, its estimate is meaningless. Our experience here is that the mean and variance parameters of the CAR model are estimated well, whereas the correlation parameters may not be. It has also been observed that this approach underestimates uncertainty in

the full conditionals [22]. Thus, this approach cannot be recommended if parameter estimates are the objective. However, we are interested merely in differentiating between classes and that $\beta$ be calibrated to permit a reasonable segmentation.

### H. Implementation Issues

To estimate the MPM segmentation using any of the models, one samples from the full conditional of the labels as specified; then, in the semi-unsupervised case, one also samples parameters; for Models IV and V, one can use (15). This is repeated until "convergence" occurs, whereupon samples are stored for each label. When "enough" have been collected, the most-often sampled label at each pixel after convergence is said to be the class at that point. The MAP segmentation is obtained using simulated annealing; therefore, it requires in addition a sequence of temperatures $T_1, T_2, \ldots$ decreasing to 0, and at the $n$th iteration of the MCMC sampler, labels are sampled from the distribution proportional to $P(y_s = r|x, y_j, j \neq s, \theta_r, \beta)^{1/T_n}$ and parameters from that proportional to $PL(\theta_r; x, y, \beta)^{1/T_n}$. Once the temperature is near 0, the sampling stops, and the current segmentation is taken to be the MAP. Several temperature sequences can be considered, such as geometric ($T_n = \rho T_{n-1}$, for $0 < \rho < 1$) or logarithmic ($T_n = 1/(a + b\log(n))$); theoretical discussion of the merits of these is found in [2].

We always initialize with a random segmentation, and on average, this seems to work well. One can, of course, start from an initial crude segmentation, but it is well documented that these MCMC approaches are liable to remain in local posterior maxima (see [23] for a simple example in the case of image restoration); therefore, the segmentation is sensitive to the starting conditions. For example, starting with a segmentation where all pixels are in one class, our experience is, even for the simple images to be analyzed in the next section, that the algorithm may only move slowly from this state. Other MCMC approaches to sampling labels, such as the Swendsen–Wang algorithm, can improve this situation [24].

Another issue is the number of iterations. This depends on image size, complexity, and the number of classes, and there are clearly no absolute rules that one can follow. For MPM segmentation, we require that the MCMC has reached equilibrium before using the sampled labels to determine the MPM; monitoring $\beta$ or the number of pixels assigned to each class is an indicator of when the method has reached an equilibrium. In general, we then again run the algorithm for as many iterations as it took to reach equilibrium and use these to determine the MPM segmentation. For the MAP segmentation, too few iterations implies a quickly decreasing temperature and a risk of being caught in a local maximum far from the global. Our experience is that 100 iterations is an absolute lower bound for either MAP or MPM, and for most images, many more will be needed. For example, for images of size 500 $\times$ 500, segmented into about five classes, 1000–2000 iterations are usually required for MPM segmentation. We emphasize that these numbers are no substitute for monitoring and evaluation of the algorithm for each image that is segmented.

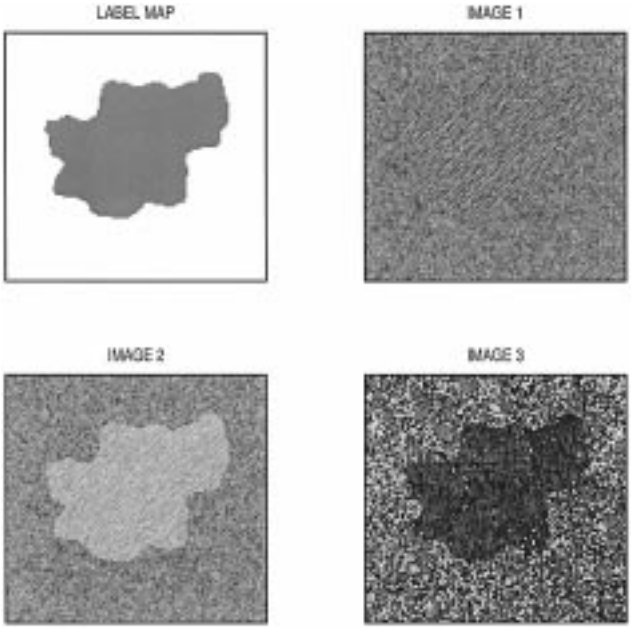


Fig. 2. Label map and the 3 textured images to be used in the simulation study.

## IV. Simulation Experiments

In this section, we compare the five models with a simulation study. We assume second-order Gaussian CAR models for the textures and a Potts model for the labels, with the exception of Model I, where the causal AR model is assumed, that is, (8), (9), (11), (13), and (14) are the relevant full conditionals for $Y$. Three different 128 $\times$ 128 images are used, as displayed in Fig. 2. Each is composed of two textures according to the true label map in the figure. Images 1 and 2 are realizations of Gaussian CAR models, and image 3 is a composition of images of leaves and grass. In image 1, the mean and variance of both classes is the same, whereas in image 2, they are not.

Two sets of simulations are conducted, comprising a total of 49 experiments. The first set is supervised, with the main goal of comparing the performance of the five models. Within each set, there are various options we might select:

- model;
- image;
- window size in the case of Model III;
- whether to fix $\beta$ or sample it.

For the latter, we choose values of $\beta$ around and above the critical value of 0.88 to compare with the sampled values, which were generally in the range 1.0–1.5 (see Fig. 3 for a typical trace of sampled values of $\beta$). The true texture parameter values are used for images 1 and 2 and the maximum likelihood estimates for image 3 (see [12]). A uniform prior for $\beta$ on the range (0, 4) is used, which gives support over critical temperature values.

In the semi-unsupervised case, we only consider Model IV. This is because we conclude from the first set of experiments that Models IV and V give better segmentations than Models I, II, or III. Models II and III are also computationally more expensive: Model II because the label full conditionals are determined by the current label of each particular neighbor and cannot be pre-computed for each sweep of all labels, as for Models IV and V, and Model III because the computation of the partition

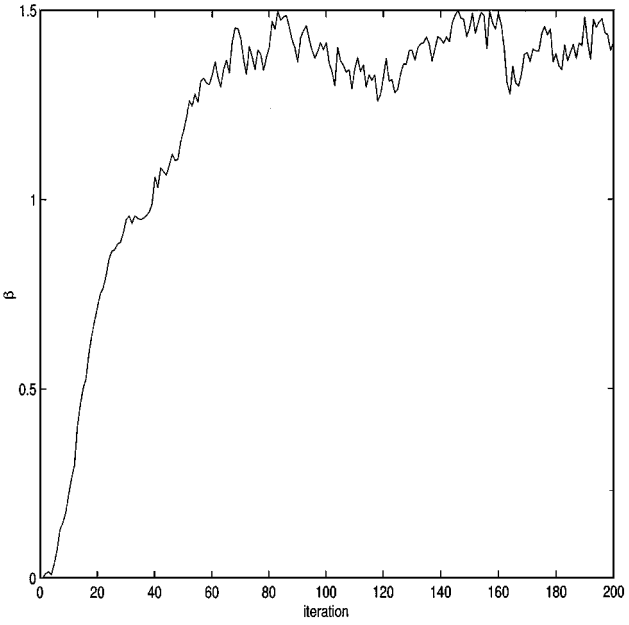


Fig. 3. Sampled values of $\beta$ for semi-unsupervised segmentation of Image 1 using Model IV.

functions on the window $W_s$ has to be redone at each iteration. Model V then fails to give a good segmentation at all if one starts at an initially random configuration of labels because many sites would have few neighbors of the same label; therefore, the full conditional of the texture parameters contains no information from the data. In this set, therefore, we concentrate on the effect of sampling $\beta$ on the segmentation. Eight different experiments are run. Uniform priors are assumed for all CAR parameters: over [0, 255] for $\mu$, [0, $255^2$] for $\sigma^2$, and over the allowable range for the clique parameters.

For both supervised and semi-unsupervised experiments, Gibbs sampling was used, with an initial burn-in of 100 iterations, and the MPM taken to be the most observed label at each site on the next 100 iterations. This is a small number of iterations, but it is adequate for such small and simple images; see Fig. 3, which shows sampled values of $\beta$ for segmentation of image 1 using Model IV, which converges after about 80 iterations. Note that the value of $\beta$ is considerably over the "critical" temperature of the Ising model, at $\beta \approx 0.88$, supporting the observation that the pseudo-likelihood overestimates $\beta$. The adequacy of the number of iterations was also determined by pilot runs of 500 iterations. Each experiment consisted of 100 separate segmentation runs. Since the MPM is that segmentation that minimizes the expected number of misclassified labels, we adopt as our performance measure the number of misclassified labels in the MPM segmentation, compared with the true label map in Fig. 2.

### A. Supervised Segmentation

Table I lists each of the 41 experiments conducted under supervised segmentation. In almost all comparable cases, the use of Model IV or V gives better results than for Models II or III. The performance of Model I can be good but is very sensitive to the choice of $\beta$, and further, it performs very badly when $\beta$ is sampled. Between Models IV and V, no strong conclusions can

TABLE I
MEDIAN AND INTERQUARTILE RANGE (IN PARENTHESES) OF MISCLASSIFIED PIXEL PERCENTAGE IN SUPERVISED SEGMENTATION. WINDOW SIZE OF MODEL III IS ALSO GIVEN

| Exp. No. | Im. No. | Model | $\beta$ | % pixels misclassified |
|---|---|---|---|---|
| 1 | 1 | I | 0.8 | 4.6 (0.69) |
| 2 | | I | 1.1 | 3.2 (1.4) |
| 3 | | I | 1.3 | 6.3 (3.4) |
| 4 | | I | 1.4 | 9.1 (4.4) |
| 5 | | I | 1.5 | 11.5 (6.7) |
| 6 | | I | ad. | 37.1 (0.35) |
| 7 | 1 | II | 0.8 | 16.7 (0.43) |
| 8 | | II | 1.1 | 9.7 (0.91) |
| 9 | | II | 1.3 | 8.2 (1.2) |
| 10 | | II | 1.4 | 7.8 (1.2) |
| 11 | | II | 1.5 | 7.7 (1.4) |
| 12 | | II | ad. | 13.3 (0.81) |
| 13 | 1 | III/15 | ad. | 5.8 (0.20) |
| 14 | | III/11 | ad. | 3.9 (0.11) |
| 15 | | III/9 | ad. | 3.0 (0.079) |
| 16 | | III/7 | ad. | 3.0 (0.085) |
| 17 | | III/5 | ad. | 4.2 (0.088) |
| 18 | 1 | IV | 0.8 | 5.8 (0.54) |
| 19 | | IV | 1.1 | 2.3 (0.33) |
| 20 | | IV | 1.3 | 1.9 (0.28) |
| 21 | | IV | 1.4 | 1.9 (0.35) |
| 22 | | IV | 1.5 | 1.82 (0.32) |
| 23 | | IV | ad. | 1.81 (0.28) |
| 24 | 1 | V | 0.8 | 6.4 (0.80) |
| 25 | | V | 1.0 | 2.9 (0.57) |
| 26 | | V | 1.1 | 2.5 (0.56) |
| 27 | | V | 1.3 | 2.7 (0.70) |
| 28 | | V | 1.5 | 3.4 (1.8) |
| 29 | | V | ad. | 2.5 (0.57) |
| 30 | 2 | III/9 | ad. | 5.1 (0.038) |
| 31 | | III/7 | ad. | 4.0 (0.044) |
| 32 | | III/5 | ad. | 2.9 (0.022) |
| 33 | | III/3 | ad. | 1.8 (0.0031) |
| 34 | 2 | IV | ad. | 0.54 (0.038) |
| 35 | 2 | V | ad. | 0.32 (0.032) |
| 36 | 3 | III/9 | ad. | 4.8 (0.066) |
| 37 | | III/7 | ad. | 3.9 (0.041) |
| 38 | | III/5 | ad. | 2.9 (0.050) |
| 39 | | III/3 | ad. | 3.2 (0.095) |
| 40 | 3 | IV | ad. | 1.4 (0.15) |
| 41 | 3 | V | ad. | 1.1 (0.14) |

be made based on these results. However, the computational cost of Model IV is considerably less since, in the label sampling step at each pixel, terms in the full conditional probabilities [see (13)] can be precomputed for each texture and stored in a look-up table. Each run with Model IV took, on average, 18% of the time of Model V. This contrast in computational cost allows us to recommend the use Model IV, where this factor is important. For Image 1, when Models IV and V are used, the adaptive version of the algorithm can be contrasted with the case where the label field parameter $\beta$ is fixed. The adaptive algorithm gives results that are at least as good as the best choice of the $\beta$ value. We conclude that, although the pseudo-likelihood does not estimate its value well, incorporating $\beta$ in the sampling process gives a better segmentation on the average at a cost of a slight increase in computation time. These results also emphasize the sensitivity of these methods to the choice of $\beta$ when it is fixed.

One can also see that images 2 and 3 were segmented more successfully than image 1 (experiments 13–41). This is not surprising, given that both classes in image 1 have the same mean and variance, whereas in 2 and 3, they do not. Another interesting result is that Model V performed better than Model IV in

TABLE II
MEDIAN AND INTERQUARTILE RANGE (IN PARENTHESES) OF THE PERCENTAGE OF MISCLASSIFIED PIXELS OVER 100 RUNS OF THE SEMI-UNSUPERVISED SEGMENTATION EXPERIMENTS WITH MODEL IV

| Exp. No. | Image No. | $\beta$ | % pixels misclassified |
|---|---|---|---|
| 42 | | 0.8 | 8.2 (0.49) |
| 43 | | 1.0 | 3.4 (0.43) |
| 44 | 1 | 1.1 | 2.7 (0.50) |
| 45 | | 1.3 | 3.1 (2.5) |
| 46 | | 1.5 | 9.2 (13.6) |
| 47 | | ad. | 2.0 (0.47) |
| 48 | 2 | ad. | 0.46 (0.038) |
| 49 | 3 | ad. | 2.27 (0.77) |

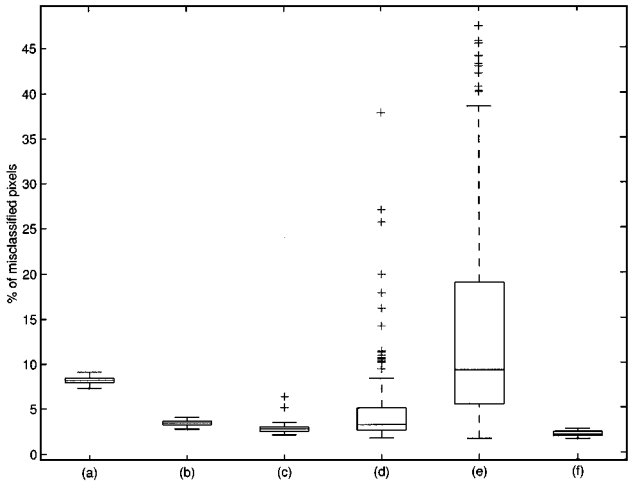


Fig. 4. Results of Experiments 42 to 49. Boxplots of misclassified pixels under Model IV with Image 1 over 100 segmentations with (a) $\beta = 0.8$, (b) $\beta = 1.0$, (c) $\beta = 1.1$, (d) $\beta = 1.3$, (e) $\beta = 1.5$, and (f) $\beta$ adaptive.

images 2 and 3 but worse in image 1. We believe that because the classes in images 2 and 3 are so distinct, Model V has the advantage over Model IV because it eliminates any boundary effects by excluding pixels in the other class from the full conditional. However, in image 1, the boundary effect is much less because the classes have the same mean and variance. In this case, what dominates is the greater uncertainty (in the sense that the full conditional probability of each class is nearer 0.5) with Model V than Model IV at boundaries because the full conditional is not based on the full neighborhood. We therefore conclude that Model V is preferable over Model IV when the classes have distinct means but not when classes are close in mean.

There is also considerable difference in the variability in performance, as indicated by the interquartile range. This seems to be mainly an image effect, with the most challenging (image 1) showing the largest variability.

### B. Semi-Unsupervised Segmentation

Table II lists each of the eight experiments conducted under semi-unsupervised segmentation. For Image 1, the segmentation algorithm with fixed $\beta$ is compared with the adaptive version, where the latter gives better results (experiments 42–47, see also Fig. 4). Each run took, on average, only 2% longer for the adaptive case than the fixed $\beta$ case. Therefore, we conclude that sampling $\beta$ improves the segmentation.

Fig. 5. Radar image of an agricultural region of Holland.

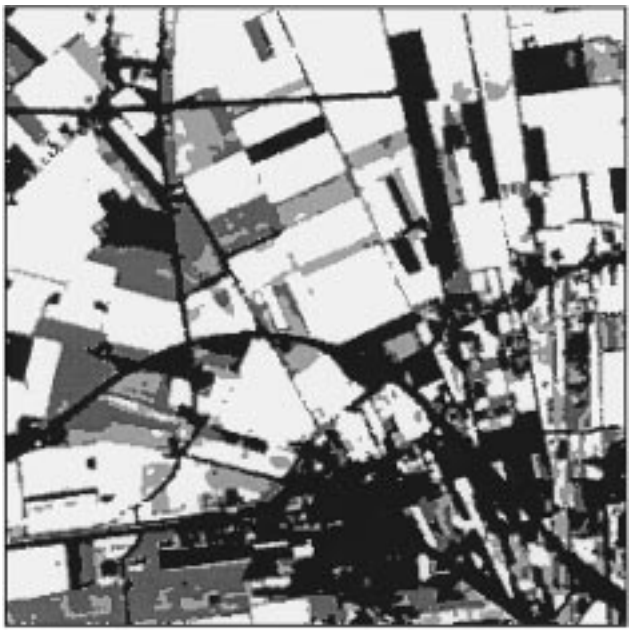

Fig. 6. MPM segmentation of the image in Fig. 5 into four classes.

When comparing the results of the adaptive segmentation algorithm on the three images (experiments 47–49), it can be seen that there is more variability in the results corresponding to the image composed by natural textures, whereas the best results are obtained with Image 2.

## V. APPLICATION TO A SATELLITE IMAGE

As an illustration of the best performing model, we segment a satellite image. The image in Fig. 5 is of an agricultural region of Holland at a resolution of 10 m/pixel. The semi-unsupervised MPM segmentation algorithm was applied to this image with an adaptive $\beta$. The algorithm was run for 1000 iterations, and the results are based on the last 600 iterations. Fig. 6 displays

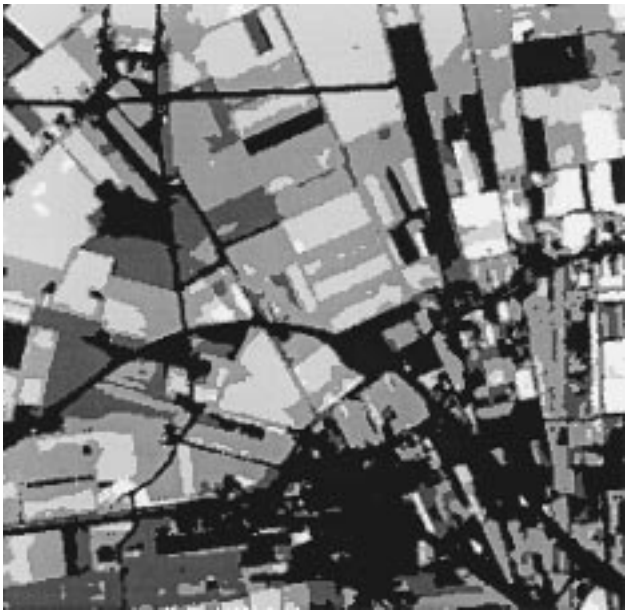

Fig. 7. MPM segmentation of the image in Fig. 5 into five classes.

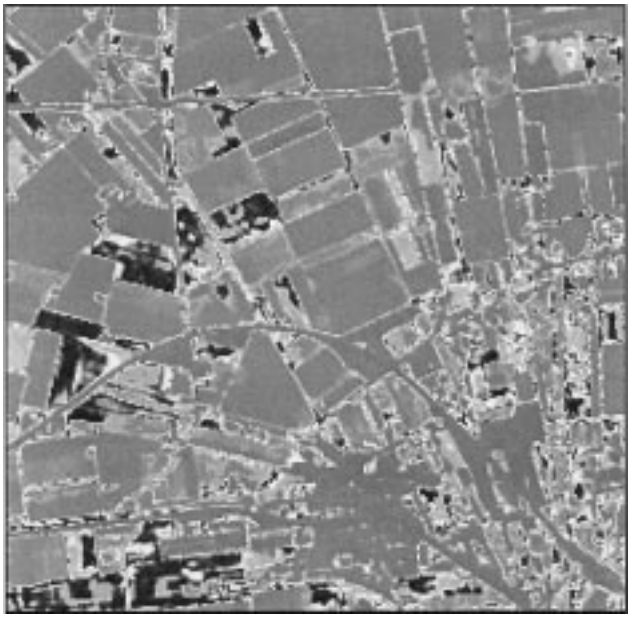

Fig. 8. Entropy of MPM segmentation of the image in Fig. 5 into four classes.

the results for an MPM segmentation into four classes. Fig. 7 then shows the MPM segmentation into five classes, based on the same number of iterations. We see that the classes assigned black and dark gray in Fig. 6 have remained but that the other two classes have been split and divided out very differently into three new classes in Fig. 7. The classification into five classes certainly appears less tidy, but it has distinguished a new class (assigned white) of lighter colored fields.

Some of the additional analyses available, in the four-class case, from the MCMC are given in Figs. 8 and 9. Fig. 8 is the entropy in marginal posterior distribution of each label, that is

$$e_s = -\sum_{\substack{j=1 \\ \hat{p}_{sj}>0}}^{R} \hat{p}_{sj} \log(\hat{p}_{sj})$$

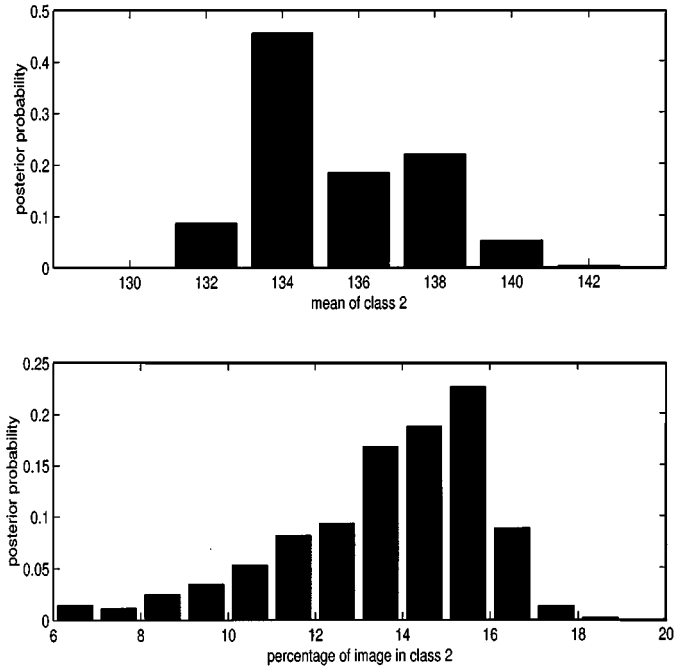


Fig. 9. Estimates of the posterior density of (top) the mean intensity for class 2, which is colored dark gray in Fig. 6, and (bottom) the percentage of pixels in class 2.

where $\hat{p}_{sj}$ is the proportion of times $y_s$ was sampled as class $j$, with lighter colors indicating higher entropy. This gives a measure of uncertainty in the class of each pixel, and we see that class 2 (colored dark gray in the segmentation) has the lowest uncertainty in general, and the highest uncertainty occurs at the borders between regions.

By looking at the relative proportion of values sampled, estimates of posterior distributions of parameters can be made. The top of Fig. 9 is an estimate of the posterior distribution of the mean of class 2. Similar plots can be made for all other model parameters, although we recall that the pseudo-likelihood may not be a good approximation for the posterior of correlation parameters. Having recorded at each iteration the number of pixels in each class, we can construct the lower plot, which is an estimate of the posterior distribution of the percentage of pixels that are in this class: something that might be of interest in applications to land-use estimation.

## VI. Concluding Remarks

Any simulation study of the type we have described cannot hope to address all the interesting issues and must restrict itself in some way. In this study, we have concentrated on how the five methods perform on simple images and on the value of adapting $\beta$ using the pseudo-likelihood approximation. Other interesting issues that we have not addressed include MPM versus MAP segmentation, performance on larger images with more classes, the effect of assuming different order neighborhoods for labels and textures, and the performance of the MCMC method for the true double Markov random field model.

According to our performance measure and considering the computational complexity involved, Model IV showed the best performance and, indeed, in semi-unsupervised segmentation,

is the only one that we have been able to implement satisfactorily. Where classes have distinct means, Model V may be better at the expense of greater computing time. We also conclude that in general, sampling of $\beta$ from the pseudo-likelihood improves the segmentation, in spite of the fact that it tends to be overestimated.

Some future developments of these techniques include fully unsupervised segmentation using Markov random field models, where $R$ is unknown. This is possible under the Bayesian approach by MCMC if one uses reversible jump methods, but this is at the expense of considerable additional computational effort [25], [26]. The issue of what constitutes a reasonable prior for $R$, or what loss function is appropriate for such segmentation, still needs to be addressed; this latter issue clearly depends on the objective of the segmentation. Indeed, development of techniques that allow a wider range of loss functions to be used, other than 0–1 and number of misclassified pixels, would allow the method to be specialized for particular applications. Some discussion of possible alternative loss functions in image analysis more generally are given in [27] and [28]. Such developments will add further computational costs to the approach, but we emphasize that the power of the MCMC approach is not in its speed but in the additional information on uncertainties in the segmentation that can be obtained.

## ACKNOWLEDGMENT

The authors would like to thank the Centre National d'Etudes Spatiales, France, for the satellite image.

## REFERENCES

[1] J. Zhang, J. W. Modestino, and D. A. Langan, "Maximum likelihood estimation for unsupervised model-based image segmentation," *IEEE Trans. Image Processing*, vol. 3, pp. 404–419, July 1994.

[2] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721–741, 1984.

[3] M. L. Comer and E. J. Delp, "Segmentation of textured images using a multiresolution Gaussian autoregressive model," *IEEE Trans. Image Processing*, vol. 8, pp. 408–420, Mar. 1999.

[4] L. Baum, T. Petire, G. Souiles, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, 1970.

[5] M. Titterington, A. J. Gray, and J. W. Kay, "An empirical study of the simulation of various models used for images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, pp. 507–513, Apr. 1994.

[6] D. B. Phillips and A. F. M. Smith, "Dynamic image analysis using Bayesian shape and texture models," *Adv. Appl. Statist. (supplement to J. Appl. Statist.)*, vol. 20, no. 5/6, pp. 299–322, 1993.

[7] R. L. Kashyap and R. Chellappa, "Choice of neighbors in spatial-interaction models of images," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 60–72, Feb. 1983.

[8] H. Derin and W. S. Cole, "Segmentation of textured images using Gibbs random fields," *Comput. Graph. Image Process.*, vol. 35, pp. 72–98, 1986.

[9] H. Derin and H. Elliott, "Modeling and segmentation of noisy and textured images using random fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 39–55, Jan. 1987.

[10] B. S. Manjunath, T. Simchony, and R. Chellappa, "Stochastic and deterministic networks for texture segmentation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1039–1049, June 1990.

[11] S. Geman and C. Graffigne, "Markov random field models and their application to computer video," in *Proc. Int. Congr. Math.*, M. Gleason, Ed. Providence, RI: Amer. Math. Soc., 1987, pp. 1496–1517.

[12] J. Besag, "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *J. R. Statist. Soc. B*, vol. 36, pp. 192–236, 1974.

[13] J. Heikkinen and H. Högmander, "Fully Bayesian approach to image restoration with an application in biogeography," *Appl. Stat.*, vol. 43, no. 4, pp. 569–583, 1994.

[14] F. S. Cohen and D. V. Cooper, "Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 195–219, Feb. 1987.

[15] H. Rue, "Fast sampling of Gaussian Markov random fields," *J. R. Statist. Soc. Ser. B*, vol. 63, no. 2, pp. 325–338, 2001.

[16] L. Knorr-Held and H. Rue. (2001) On block updating in Markov random field models for disease mapping. Tech. Rep.. [Online]. Available: www.stat.uni-muenchen.de/~leo/publikationen.html.

[17] I. S. Weir, "Fully Bayesian reconstructions from single-photon emission computed tomography data," *J. Amer. Statist. Assoc.*, vol. 92, no. 437, pp. 49–60, 1997.

[18] C. J. Geyer and E. A. Thompson, "Constrained Monte Carlo maximum likelihood for dependent data (with discussion)," *J. R. Statist. Soc. B*, vol. 54, pp. 657–699, 1992.

[19] S. Lakshmanan and H. Derin, "Valid parameter space for 2-D Gaussian Markov random fields," *IEEE Trans. Inform. Theory*, vol. 39, pp. 703–709, Apr. 1993.

[20] D. E. Melas, "A Bayesian approach to the segmentation of textured images," Ph.D. dissertation, Dept. Statist., Univ. Dublin, Trinity College, Dublin, Ireland, 1999.

[21] A. M. Higdon, "Spatial applications of Markov chain Monte Carlo for Bayesian inference," Ph.D. dissertation, Dept. Statist., Univ. Washington, Seattle, 1994.

[22] C. J. Geyer, "Markov chain Monte Carlo maximum likelihood," in *Proc. 23rd Symp. Interface*, 1991, pp. 156–163.

[23] D. M. Greig, B. T. Porteus, and A. H. Seheult, "Exact maximum *a posteriori* estimation for binary images," *J. R. Statist. Soc. B*, vol. 51, pp. 271–279, 1989.

[24] R. H. Swendsen and J.-S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," *Phys. Rev. Lett.*, vol. 58, pp. 86–88, 1987.

[25] Z. Kato, "Bayesian color image segmentation using reversible jump Markov chain Monte Carlo," Eur. Res. Consort. Inform. Math., Res. Rep. 01/99-R055, 1999.

[26] S. A. Barker and P. J. W. Rayner, "Unsupervised image segmentation using Markov random field models," *Pattern Recogn.*, vol. 33, pp. 587–602, 2000.

[27] H. Rue, "New loss functions in Bayesian imaging," *J. Amer. Statist. Assoc.*, vol. 90, pp. 900–908, 1995.

[28] H. Rue and A. Frigessi, "Bayesian image classification using Baddeley's delta loss," *J. Comput. Graphic. Stat.*, vol. 6, no. 1, pp. 55–73, 1997.

**Dina E. Melas** received the Ph.D. degree from the Department of Statistics, Trinity College, Dublin, Ireland.

She is currently with Interoperability Systems International, Athens, Greece.

**Simon P. Wilson** received the Ph.D. degree in stochastic modeling from the George Washington University, Washington, DC.

He is a Lecturer with the Department of Statistics, Trinity College, Dublin, Ireland.