PERGAMON

Pattern Recognition 33 (2000) 1919–1925

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

MRF parameter estimation by MCMC method

Lei Wang, Jun Liu*, Stan Z. Li

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore

Received 13 January 1999; received in revised form 28 July 1999; accepted 28 July 1999

Abstract

Markov random field (MRF) modeling is a popular pattern analysis method and MRF parameter estimation plays an important role in MRF modeling. In this paper, a method based on Markov Chain Monte Carlo (MCMC) is proposed to estimate MRF parameters. Pseudo-likelihood is used to represent likelihood function and it gives a good estimation result. © 2000 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: MRF; MCMC; Least-squares fit; Parameter estimation; Pseudo-likelihood

1. Introduction

The objective of mathematical modeling in pattern analysis is aimed to extract the intrinsic characteristics of the pattern in a few parameters so as to represent the pattern effectively. Markov random field modeling is a very popular pattern analysis method and it plays an important role in pattern recognition and computer vision. Markov random field models were popularized by Besag to model spatial interactions on lattice system [1]. It can be used in texture classification and segmentation as well as image restoration [2]. The most important characteristic of MRF modeling is that the global patterns can be formed via stochastic propagation of local interactions. MRF parameter estimation is necessary in MRF modeling after the form of model is given. During the past years, many authors presented methods to estimate MRF parameters. Simulated annealing [3], maximum likelihood [4], coding method [1], mean field approximations [5], Bayesian estimation [6] and least-squares (LS) fit [7] are discussed to estimate MRF parameters.

Least-squares (LS) methods and maximum likelihood methods are often used. However, LS is not accurate in estimation and maximum likelihood method is time-consuming. Here a method based on MCMC is used to estimate the parameters which can give a good solution to the estimation.

The general parameter estimation principle is as follows. Let F denote any finite set which comprises of a random field and $f \in F$ is an observation of F. On F a family of distributions

$$\Pi = \{\Pi(F; \theta) : \theta \in \Theta\}$$

is considered where $\Theta \subset R^d$ is a set of parameters. The 'true' parameter $\theta^* \in \Theta$ is not known and needs to be determined or at least approximated. The only available information is hidden in the observation f which is a realization of F. Now, the problem is how to choose $\hat{\theta}$ as a substitute for θ^* if f is picked at random from $\Pi(F; \theta^*)$.

In this paper, the estimation of parameters is based on deriving posterior distribution calculated using the Metropolis–Hastings algorithm. This is a Markov chain Monte Carlo (MCMC) technique [8].

The paper is arranged as follows. MRF image model is discussed in Section 2. MCMC parameter estimation is proposed in Section 3. The experiments are shown in Section 4 and conclusion is given in Section 5.

2. MRF modeling

This section introduces some notations related to MRF modeling which will be used in the following sections of the paper.

* Corresponding author.

E-mail addresses: elwang@263.net (L. Wang), ejliu@ntu.edu.sg (J. Liu), szli@szli.eee.ntu.edu.sg (S.Z. Li).

PII: S0031-3203(99)00178-8

A lattice is a square array of pixels, or sites, $\{(j,k): 0 \leqslant j \leqslant N-1, 0 \leqslant k \leqslant N-1\}$. We adopt a simple numbering of sites by assigning sequence number $i = k + Nj$ to site (j, k). Letting $M = N^2$ denote the number of sites,

$$S = \{0, 1, \dots, M-1\}$$

index the set of sites. A random field model is a distribution for the M-tuple random vector F, which contains a random variable $F(i)$ for the value of site i. The sites in S are related to one another via a neighborhood system. A neighborhood system for S is defined as

$$\mathcal{N} = \{\mathcal{N}_i | \forall i \in S\},$$

where $\mathcal{N}_i$ is the set of sites neighboring i. The neighboring relationship has the following properties:

(1) a site is not neighboring to itself;
(2) the neighboring relationship is mutual.

A clique c is a set of sites in which all pairs of sites are mutual neighbors. The set of all cliques in a neighborhood system is denoted as Q.

Suppose F is an MRF. Let $f \in F$ be a realization of F. A clique function, or potential function, $V_c(f)$, is associated with each clique and the energy function, $U(f)$, of MRF can be expressed as the sum of clique functions.

$$U(f) = \sum_{c \in Q} V_c(f).$$

To a homogeneous MRF, the potential function is independent of locations. Thus, the number of clique potentials can be reduced to the number of clique types, that is, each potential corresponding to a clique type.

Consider a multi-level logistic (MLL) model [4]. Let $\mathcal{L} = \{1, \dots, m\}$ be the label set and $\theta = (\theta_1, \dots, \theta_n)$ be the parameter vector for clique potentials where each component corresponds to a clique type. Consider the distribution of Gibbs form,

$$P(F = f, \theta) \propto P(F = f|\theta) = Z(\theta)^{-1} \exp(-U(f, \theta)), \quad (1)$$

where $U(f, \theta)$ is energy function and depends linearly on θ. Suppose $H(f) = (H_1, \dots, H_n)$ is the histogram of cliques of f, n denotes the index of clique type. Let

$$\delta(z) = \begin{cases} 1 & \text{if } z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

$$H_i = \sum_{j \in S} \left(2 \sum_{j' \in c_i(j)} \delta(f_j - f_{j'}) - 1 \right), \quad i = 1, \dots, n,$$

where $c_i(j)$ denotes the neighbor of site j in the ith clique c_i.

Then the distributions have the form

$$P(f, \theta) = Z(\theta)^{-1} \exp(\langle \theta, H(f) \rangle), \ \theta \in \Theta$$

where $\langle \theta, H(f) \rangle = \sum_{i=1}^{n} \theta_i H_i$ is the inner product of θ and H, and $Z(\theta) = \sum_f \exp(\langle \theta, H(f) \rangle)$ is the normalizing partition function. The conditional probability is as follows:

$$P(f_j | f_{N_j}) = \frac{\exp(-\langle \theta, H(f_j) \rangle)}{\sum_{z_j \in \mathcal{L}} \exp(-\sum \langle \theta, H(z_j) \rangle)}, \quad (2)$$

where $H(f_j)$ is the local histogram only calculated in the neighborhood of site j. $H(z_j)$ denotes the local histogram replacing f_j with z_j and the neighborhood of j is fixed. The computation of $Z(\theta)$ is infeasible because there are a combinatorial number of elements in the configuration space. In order to avoid using the partition function $Z(\theta)$, the pseudo-likelihood function

$$PL(f|\theta) = \log\left(\prod_{j \in \mathcal{S}} P(f_j | f_{N_j}) \right)$$

$$= \prod_{j \in \mathcal{S}} (\langle \theta, H(f_j) \rangle) - \log \sum_{z_j \in \mathcal{L}} \exp(\langle \theta, H(z_j) \rangle) \quad (3)$$

can be used to replace likelihood function. The pseudo-likelihood does not involve the partition function $Z(\theta)$. Hence it is much easier to be calculated.

3. MCMC estimation of MRF parameters

According to Bayesian theorem, the posterior distribution of θ conditional on f is

$$P(\theta | f) = \frac{P(\theta) P(f|\theta)}{\int P(\theta) P(f|\theta) \, d\theta} \propto P(\theta) P(f|\theta). \quad (4)$$

According to Gilks et al. [8], any features of the posterior distribution are legitimate for Bayesian inference: moments, quantiles, highest posterior density regions, etc. All these quantiles can be expressed in terms of posterior expectations of functions of θ. The posterior expectation of a function $g(\theta)$ is

$$E[g(\theta)|f] = \frac{\int g(\theta) P(\theta) P(f|\theta) \, d\theta}{\int P(\theta) P(f|\theta) \, d\theta}. \quad (5)$$

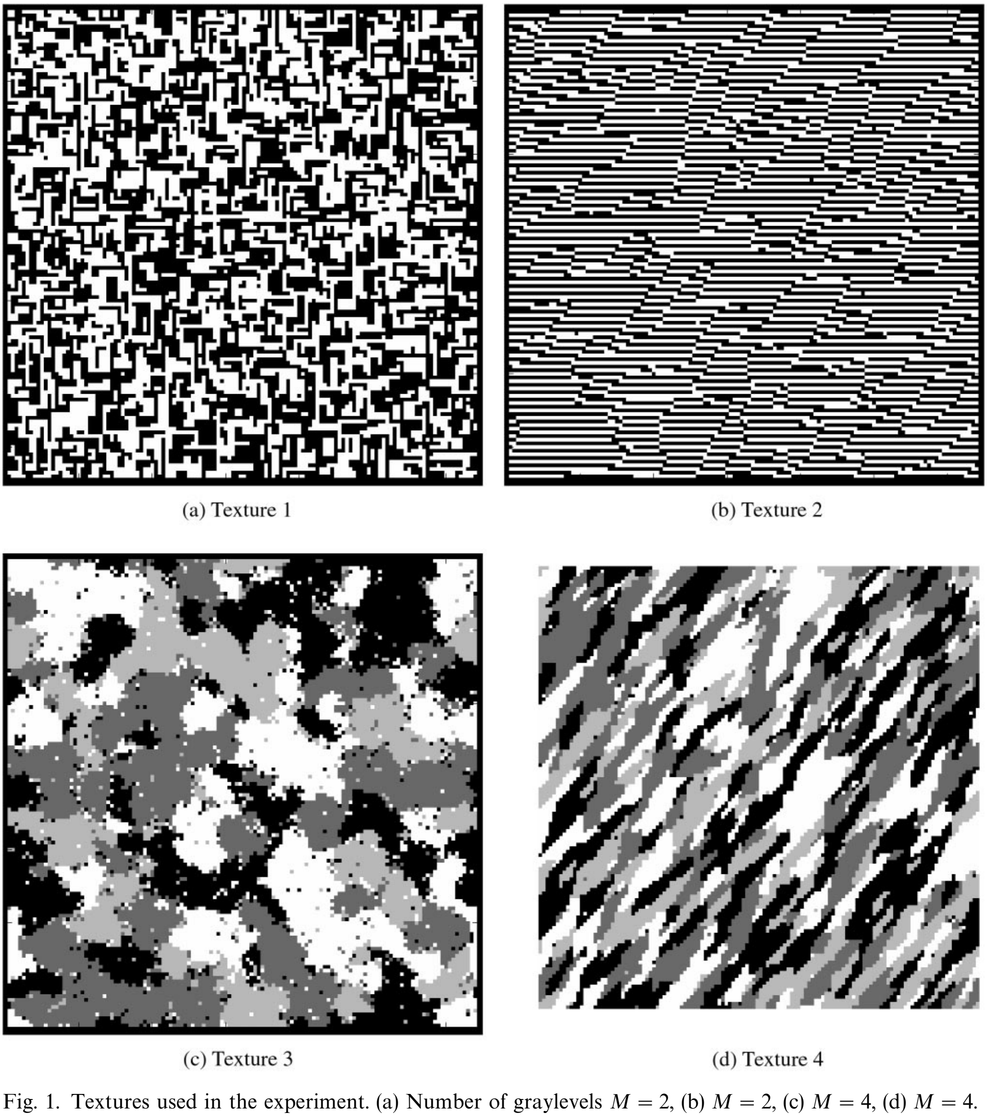
The integrations in this expression are difficult to be solved in Bayesian inference. Monte Carlo integration including Markov Chain Monte Carlo (MCMC) approach [6] can be used to deal with the difficulty [8]. The task is to evaluate the expectation

$$E(g(\theta)) = \frac{\int g(\theta)P(\theta)\,\mathrm{d}\theta}{\int P(\theta)\,\mathrm{d}\theta}. \tag{6}$$

A Markov chain can be adopted for the purpose of evaluation. Suppose we generate a sequence of random variables $\{\theta^0, \theta^1, \dots\}$. At each time $t \geqslant 0$, the next state θ^{t+1} is sampled from a distribution $P(\theta^{t+1}|\theta^t)$ which depends only on the current state θ^t of the chain. This Markov chain is assumed to be time-homogeneous. Thus, the sequence will gradually converge to a unique stationary distribution $\phi(.)$. After a sufficient long burn-in of say m iterations, $\{\theta^t, t = m+1, \dots, n\}$ will be dependent samples approximately from $\phi(.)$. Let

$$\bar{\theta} = \frac{1}{n-m} \sum_{t=m+1}^{n} \theta^t. \tag{7}$$

This is an ergodic average. Convergence to the required expectation is ensured by the ergodic theorem. Eq. (7) shows how a Markov chain can be used to estimate $E(\theta|f)$. Such a Markov chain can be constructed by Metropolis–Hastings algorithm [8]. At each time t, the next state θ^{t+1} is chosen by first sampling a candidate point θ' from a proposal distribution $q(\theta'|\theta^t)$. The choice of proposal distribution is almost arbitrary; here a multivariate normal distribution centered on the current value θ^t is adopted. The candidate θ' is accepted


(a) Texture 1

(b) Texture 2

(c) Texture 3

(d) Texture 4

Fig. 1. Textures used in the experiment. (a) Number of graylevels $M = 2$, (b) $M = 2$, (c) $M = 4$, (d) $M = 4$.

with probability

$$\alpha(\theta^t, \theta') = \min\left(1, \frac{P(\theta'|f)q(\theta'|\theta^t)}{P(\theta^t|f)q(\theta^t|\theta')}\right).$$

The transition kernel for the Metropolis-Hastings algorithm is

$$P(\theta^{t+1}|\theta^t) = q(\theta^{t+1}|\theta^t)\alpha(\theta^t, \theta^{t+1}) + I(\theta^{t+1} = \theta^t) \times \left[1 - \int q(\theta'|\theta^t)\alpha(\theta^t, \theta')d\theta'\right]$$

where $I(.)$ denotes the indicator function (taking 1 when its argument is true, and 0 otherwise). If the candidate θ' is accepted, the next state becomes $\theta^{t+1} = \theta'$, otherwise $\theta^{t+1} = \theta^t$. Since $P(\theta|f) \propto P(\theta)P(f|\theta)$ and the prior $P(\theta)$ can be assumed to be flat when the prior information is totally unavailable,

$$\alpha(\theta^t, \theta') = \min\left(1, \frac{P(\theta'|f)q(\theta'|\theta^t)}{P(\theta^t|f)q(\theta^t|\theta')}\right) = \min\left(1, \frac{P(f|\theta')q(\theta'|\theta^t)}{P(f|\theta^t)q(\theta^t|\theta')}\right). \quad (8)$$

Since the choice of proposal distribution here is normal centered on the current value, $q(\theta'|\theta^t) = q(\theta^t|\theta')$ due to the symmetric property of the proposed distribution. Thus, the acceptance probability formula can be reduced to

$$\alpha(\theta^t, \theta') = \min\left(1, \frac{P(f|\theta')}{P(f|\theta^t)}\right). \quad (9)$$

Thus, the Metropolis–Hastings algorithm is switched to Metropolis algorithm. When we use pseudo-likelihood to represent the likelihood function, we get

$$\begin{aligned}\alpha(\theta^t, \theta') &= \min(1, \exp(PL(f|\theta') - PL(f|\theta^t)))\\ &= \min\Bigg(1, \exp\Bigg(\sum_{j\in S}\Bigg[(\langle\theta', H(f_j)\rangle - \langle\theta^t, H(f_j)\rangle)\\ &\quad - \log\Bigg(\sum_{z_j\in\mathscr{L}} \exp(\langle\theta', H(z_j)\rangle)\Bigg)\\ &\quad + \log\Bigg(\sum_{z_j\in\mathscr{L}} \exp(\langle\theta^t, H(z_j)\rangle)\Bigg)\Bigg]\Bigg)\Bigg). \quad (10)\end{aligned}$$

With this acceptance probability, the θ can be approximated effectively.

The Metropolis–Hastings algorithm can be summarized in the following procedures:

```
Initialize $\theta^0$; set $t = 0$ and $T$ = maximum number of iteration
While $t < T$
BEGIN
Sample a point $\theta'$ from $q(.|\theta^t)$
Sample a uniform (0, 1) random variable $v$
If $v \leqslant \alpha(\theta^t, \theta')$, set $\theta^{t+1} = \theta'$. Otherwise set $\theta^{t+1} = \theta^t$
Increment $t$
END
```

4. Experiments

In order to inspect the performance of the method proposed in this paper, a Gibbs sampler [4] is used to sample textures with the specified parameters. Here a second-order neighborhood system is used and four double-site cliques $\theta = \{\beta_1, \beta_2, \beta_3, \beta_4\}$ corresponding to 0°, 90°, 45° and 135° individually are adopted as non-zero parameters. Fig. 1 shows four 128×128 textures generated from the Gibbs Sampler. The first two textures are sampled with two graylevels and the next two textures are sampled with four graylevels. The parameters of the four textures are listed in Table 1. In order to get acceptable parameters, the MCMC procedure described in the previous section should be repeated until stability of the Markov chain is reached. The choice of starting values θ^0 will not affect the stationary distribution if the chain is irreducible. In our experiments, θ^0 are chosen randomly. The usual informal approach to detection of convergence is visual inspection of plots of the Monte-Carlo output $\{\theta^t, t = 1, \ldots, n\}$. From Figs. 2–5, three independent samples of Markov chains for texture 4 are given. From the

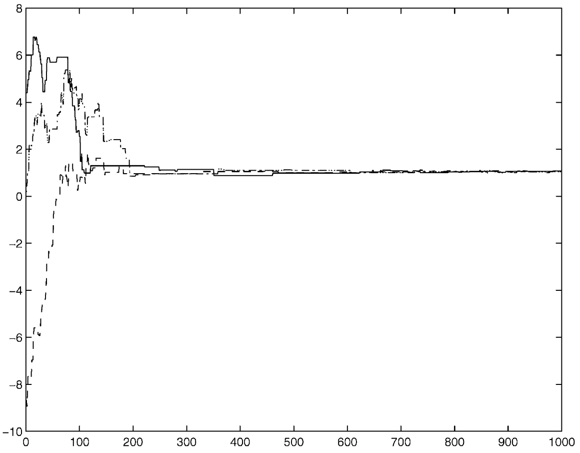


Fig. 2. 1000 iterations with different starting values for estimating β_1 for texture 4.

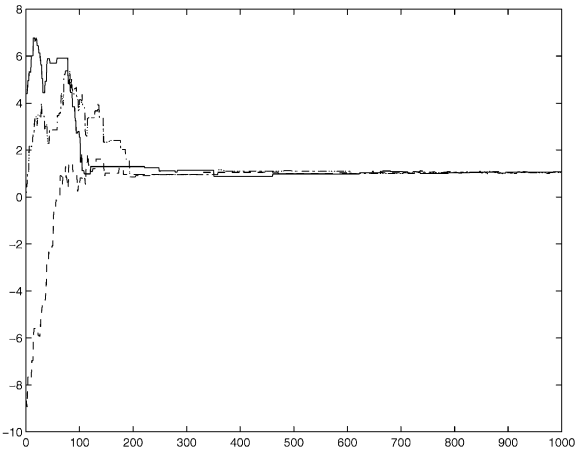


Fig. 3. 1000 iterations with different starting values for estimating β_2 for texture 4.

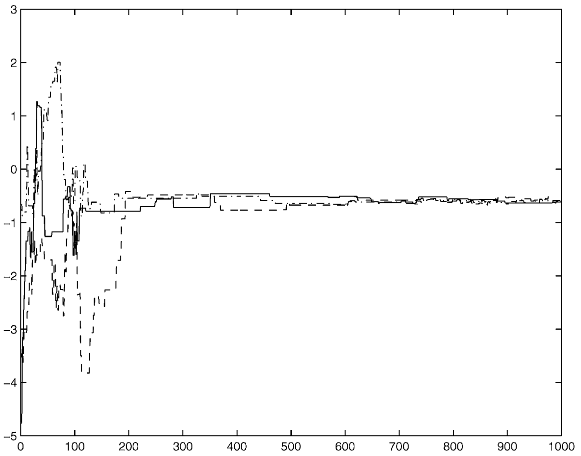


Fig. 4. 1000 iterations with different starting values for estimating β_3 for texture 4.

figures, we observe that the length of burn-in depends on θ^0. The Markov chains converge in less than 300 iterations in most examples according to visual inspection of the monitoring statistics. Here we set burn-in $m = 500$. More formal methods for convergence diagnostics can be found in Refs. [9,10]. Decision about the iteration number is an important and practical matter. The aim is to run the chain long enough to obtain adequate precision in the estimator. Here three chains are run in parallel with different starting values $\bar{\theta}$ from Eq. (7). If they do not agree adequately, the iteration number n must be increased. Initially, we set $n = 1000$. If the estimates $\bar{\theta}$ do not agree adequately, we increase 500 iterations each time until estimates are similar. We only need to inspect the mean $\bar{\theta}$ and variance σ of the Monte–Carlo output. In our experiments in Table 1, $n = 1000$ is enough. The results of MCMC approach in Table 1 are acceptable where σ denotes the average standard deviation of Markov chains after burn-in. In order to verify the performance of this method, least- squares (LS) fit method proposed by Derin and Elliott [7] is also used in our experiments. From Table 1, it can be seen that LS

Table 1
MRF parameter estimation

Textures	Method	β_1	β_2	β_3	β_4
Texture 1	Specified	1	1	– 0.5	– 0.5
	LS	0.8448	0.8734	– 0.4332	– 0.4382
	MCMC	0.9884	0.9899	– 0.5076	– 0.5078
	σ	0.0436	0.0323	0.0278	0.0400
Texture 2	Specified	1	– 0.8	0.5	– 0.5
	LS	0.9949	– 0.8157	0.4960	– 0.3244
	MCMC	1.0093	– 0.8586	0.5569	– 0.4522
	σ	0.0147	0.0235	0.0168	0.0245
Texture 3	Specified	0.3	0.3	0.3	0.3
	LS	0.1152	0.1520	0.1867	0.1444
	MCMC	0.3478	0.2762	0.2960	0.2877
	σ	0.0266	0.0165	0.0130	0.0086
Texture 4	Specified	0.5	1	– 0.5	0.7
	LS	0.0415	0.4364	0	0.4201
	MCMC	0.5525	1.0394	– 0.5951	0.6810
	σ	0.0321	0.0364	0.0490	0.0156

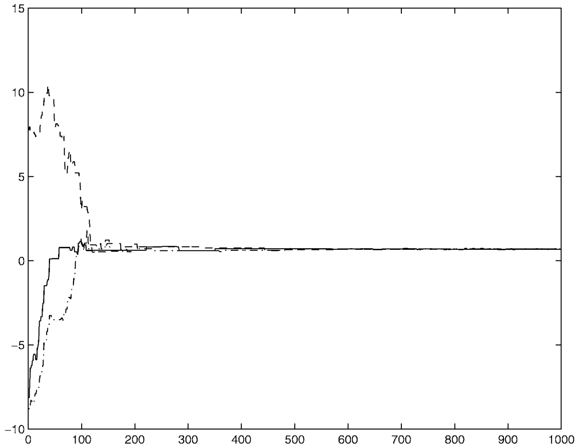


Fig. 5. 1000 iterations with different starting values for estimating β_4 for texture 4.

method is effective only to the textures with two graylevels, while MCMC method is effective to all examples in the experiments even more graylevels are adopted in the model. From the comparison, MCMC method proposed in this paper is much better than LS method. The MCMC routines are run on a Sun Ultra 2 workstation, each analysis takes less than 3 min to perform 1000 iterations.

5. Conclusion

Markov random fields (MRFs) modeling is a popular pattern analysis method. It can be used in texture classification and segmentation as well as image restoration. MRF parameter estimation plays an important role in MRF modeling. In order to estimate MRF parameter effectively and efficiently, an MRF parameter estimation method based on method MCMC is proposed in this paper. A Markov chain is constructed to sample the MRF parameters via Monte Carlo method. MLL model is used as image model. In order to avoid to calculate the normalizing partition function, pseudo-likelihood function is used to represent likelihood function. Compared to least-squares fit method, the proposed method is more accurate and can be used for multilevel-graylevel texture parameter estimation effectively as seen from the experiments in the paper. This method can be extended to be used in multiresolution analysis of texture modeling and segmentation of textured images.

6. Summary

Markov random fields (MRFs) modeling is a popular pattern analysis method. It can be used in texture classification and segmentation as well as image restoration. MRF parameter estimation plays an important role in MRF modeling. In order to estimate MRF parameter effectively and efficiently, an MRF parameter estimation method based on MCMC is proposed in this paper. A Markov chain is constructed to sample the MRF parameters via Monte Carlo method. MLL model is used as image model. In order to avoid to calculate the normalizing partition function, pseudo-likelihood function is used to represent likelihood function. Compared to least-squares fit method, our method is more accurate and can be used for multi-graylevel texture parameter estimation effectively as seen from the experiments in the paper. This method can be extended to be used in multiresolution analysis of texture modeling and segmentation of textured images.

Acknowledgements

We wish to thank the constructive comments and suggestions of the reviewers.

References

[1] J. Besag, Spatial interaction and the statistical analysis of lattice systems, J. Roy. Statist. Soc. Ser. B 36 (1974) 192–236.

[2] S. Barker, Image segmentation using Markov random field models, Ph.D. Thesis, University of Cambridge, 1998.

[3] S. Geman, D. Geman, Stochastic relaxation, Gibbs distribution and the Bayesian restoratopn of images, IEEE Trans. PAMI 6 (6) (1984) 721–741.

[4] S. Li, Markov Random Field Modeling in Computer Vision, Springer, New York, 1995.

[5] D. Chandler, Introduction to Modern Statistical Mechanics, Oxford University Press, Oxford, 1987.

[6] R. Aykroyd, Bayesian estimation for homogeneous and inhomogeneous Gaussian random fields, IEEE Trans. Pattern Analysis Mach. Intell. 20 (5) (1998) 533–539.

[7] H. Derin, H. Elliott, Modeling and segmentation of noisy and textured images using Gibbs random fields, IEEE Trans. Pattern Analysis Mach. Intell. 9 (1) (1987) 39–55.

[8] W. Gilks, S. Richardson, D. Spiegelhalter, Markov chain Monte Carlo in Practice, Chapman & Hall, London, 1996.

[9] M. Cowles, B. Carlin, Markov Chain Monte Carlo convergence diagnostics: a comparative review, Technical Report, Division of Biostatistics, School of Public Health, University of Minnesota, 1994.

[10] S. Brooks, P. Dellaportas, G. Roberts, An approach to diagnosing total variation convergence of MCMC algorithms, University of Cambridge, http://www.stats.bris.ac.uk/ maspb/mypapers/brodr96.html, 1996.

About the Author—LEI WANG received his B. Eng and M. Eng degrees from Harbin Institute of Technology, China, in 1992 and 1995, respectively. He is currently a Ph.D. candidate in School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include pattern recognition, image compression, image processing, and image retrieval.

About the Author—JUN LIU received his B.S. and M.S. degree from Jiao Tong University, Xian, China in 1982 and 1984, respectively. He obtained his Ph.D. degree from Oakland University, MI, USA in 1989. He is currently an associate professor with Division of Information Engineering, School of Electrical and Electronic Engineering, Nanyang Technological University. His research interests include pattern recognition, image processing and multimedia database.

About the Author—STAN Z. LI received the B.Sc degree from Hunan University, China, in 1982, M.Sc degree from the National University of Defense Technology, China, in 1985 and Ph.D. degree from the University of Surrey, UK, in 1991. All degrees are in EEE. He is currently a senior lecturer at Nanyang Technological University, Singapore. His research interests include computer vision, pattern recognition, image processing and optimization methods.