Unsupervised Parallel Image Classification Using a Hierarchical Markovian Model*

Zoltan KATO, Josiane ZERUBIA and Marc BERTHOD

INRIA - 2004 Route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex - FRANCE

Tel (33) 93 65 78 57 - Fax (33) 93 65 76 43 - email:<name>@sophia.inria.fr

Abstract

This paper deals with the problem of unsupervised classification of images modeled by Markov Random Fields (MRF). If the model parameters are known then we have various methods to solve the segmentation problem (simulated annealing, ICM, etc...). However, when they are not known, the problem becomes more difficult. One has to estimate the hidden label field parameters from the only observable image. Our approach consists of extending a recent iterative method of estimation, called Iterative Conditional Estimation (ICE) to a hierarchical markovian model. The idea resembles the Estimation-Maximization (EM) algorithm as we recursively look at the Maximum a Posteriori (MAP) estimate of the label field given the estimated parameters then we look at the Maximum Likelihood (ML) estimate of the parameters given a tentative labeling obtained at the previous step. We propose unsupervised image classification algorithms using a hierarchical model. The only parameter supposed to be known is the number of regions, all the other parameters are estimated. The presented algorithms have been implemented on a Connection Machine CM200. Comparative tests have been done on noisy synthetic and real images (remote sensing).

Key Words: hierarchical Markovian model, parameter estimation, unsupervised image classification.

1 Introduction

In real life applications, the model parameters are usually unknown, one has to estimate [1] them only from the observable image. From a statistical viewpoint, this means that we want to estimate parameters from random variables whose joint distribution is a mixture of distributions. If we have a realization of the label field then the problem is relatively easy, we have many standard methods to do parameter estimation (Maximum Likelihood, Coding method [2], etc...). Unfortunately, such a realization is not known, so the direct use of such estimation algorithm is impossible. We have to approximate it by some function of the image data, which is the only observable attribute.

Some nowadays used algorithms are iterative [3, 14, 15], subsequently generating a labeling, estimating parameters from it, then generating a new labeling using these parameters, etc ... For such a method, we need a reasonably good initial value for each parameter. Since the classes of a labeling problem are mostly represented by a Gaussian distribution, the initialization of the mean and the variance of each class is very important because of its influence on subsequent labelings and hence on the final estimates. On the other hand, it is a classical problem, namely the determination of the modes of a Gaussian mixture without any a priori information. There are many approaches in this domain: Method of moments [6], Prony's Method [5] or geometrical analysis of the histogram [16], for instance.

Herein, we will present a parameter estimation method applied to hierarchical MRF models. The proposed algorithm has been tested on image segmentation problems. Comparative test have been done on noisy synthetic and real satellite images.

2 The Parameter Estimation Problem

Let us briefly review some notations. $\mathcal{F} = \{F_s : s \in \mathcal{S}\}$ denotes a set of image data on the sites (or pixels) $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$. Furthermore, each of these sites may take a label from $\Lambda = \{0, 1, \ldots, L-1\}$. The configuration space Ω is the set of all global discrete labelings $\omega = (\omega_{s_1}, \ldots, \omega_{s_N}), \omega_s \in \Lambda$. The label process is denoted by $\mathcal{X}$.

In parameter estimation problems, $\mathcal{F}$ is also called the *observed image* and $\mathcal{X}$ denotes the *unobserved* image attributes (labels). Furthermore, we are given n parameters forming a vector Θ which appears in the MRF model:

$$\Theta = \begin{pmatrix} \vartheta_1 \\ \vdots \\ \vartheta_n \end{pmatrix} \tag{1}$$

Usually, Θ is considered to be known. Therefore, one is looking for the labeling which maximizes the a posteriori distribution

$$\widehat{\omega} = \arg\max_{\omega \in \Omega} P_\Theta(\omega \mid \mathcal{F}, \Theta). \tag{2}$$

where $\widehat{\omega}$ is the MAP estimate of the label field, given $\mathcal{F}$, under the model P_Θ (in the followings, the index Θ will be omitted). If both Θ and ω are unknown, the

*This work has been partially funded by CNES, AFIRST and DRED/GdR-TdSI

maximization problem in Equation (2) becomes [7, 12]

$$(\widehat{\omega}, \widehat{\Theta}) = \arg\max_{\omega,\Theta} P(\omega, \mathcal{F} \mid \Theta). \tag{3}$$

The pair $(\widehat{\omega}, \widehat{\Theta})$ is the global maximum of the joint probability $P(\omega, \mathcal{F} \mid \Theta)$. If we regard Θ as a random variable, the above maximization is an ordinary MAP estimation in the following way [7]: Let us suppose, that Θ is restricted to a finite volume domain $\mathcal{D}_\Theta$ and suppose that Θ is uniform on $\mathcal{D}_\Theta$ (that is $P(\Theta)$ is constant). Then, we get [7]:

$$\begin{aligned}
& \arg\max_{\omega,\Theta} P(\omega, \Theta \mid \mathcal{F}) \\
= \; & \arg\max_{\omega,\Theta} \frac{P(\omega, \mathcal{F} \mid \Theta) P(\Theta)}{P(\mathcal{F})} && (4) \\
= \; & \arg\max_{\omega,\Theta} \frac{P(\omega, \mathcal{F} \mid \Theta)}{\int_{\mathcal{D}_\Theta} \sum_{\omega \in \Omega} P(\omega, \mathcal{F} \mid \Theta) d\Theta} && (5) \\
= \; & \arg\max_{\omega,\Theta} P(\omega, \mathcal{F} \mid \Theta). && (6)
\end{aligned}$$

However, this maximization is very difficult, having no direct solution. Even Simulated Annealing (SA) is not implementable because the local characteristics with respect to the parameters Θ cannot be computed from $P(\omega, \mathcal{F} \mid \Theta)$. One possible solution is to adopt the following criterion instead [7, 12]:

$$\begin{aligned}
\widehat{\omega} &= \arg\max_{\omega} P(\omega, \mathcal{F} \mid \widehat{\Theta}) && (7) \\
\widehat{\Theta} &= \arg\max_{\Theta} P(\widehat{\omega}, \mathcal{F} \mid \Theta) && (8)
\end{aligned}$$

Clearly, Equation (7) is equivalent to Equation (3) for $\Theta = \widehat{\Theta}$ and Equation (8) is equivalent to Equation (3) with $\omega = \widehat{\omega}$. Furthermore, Equation (7) is equivalent to the MAP estimate of ω in the case of known parameters:

$$\begin{aligned}
\arg\max_{\omega} P(\omega, \mathcal{F} \mid \widehat{\Theta}) &= \arg\max_{\omega} P(\omega \mid \mathcal{F}, \widehat{\Theta}) P(\mathcal{F} \mid \widehat{\Theta}) \\
&= \arg\max_{\omega} P(\omega \mid \mathcal{F}, \widehat{\Theta}).
\end{aligned}$$

3 Parameter Estimation from Incomplete Data

In real life applications, labeled samples are usually not available. We have to estimate the parameters from an *unlabeled* sample. In statistics, the problem is known as the *incomplete data* problem. A broadly applicable algorithm has been proposed by Dempster *et al.* [4], called *Expectation – Maximization* (EM). The algorithm aims at determining the ML estimate of the parameters Θ by making use of the estimation of the missing data (i.e. the label field $\mathcal{X}$). A few other iterative estimation methods [7, 14, 15] are available when dealing with incomplete data.

Another EM-like algorithm has been proposed by Geman in [7], which is called *Adaptive Simulated Annealing* (ASA). The algorithm was adapted to image segmentation problems in [12], where the convergence of ASA has also been proved. The ASA algorithm is very similar to the SEM, it may be seen as a special case where the S-step is implemented by a Simulated Annealing.

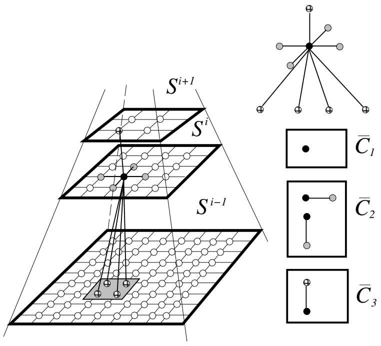

Figure 1: Hierarchical MRF model.

Algorithm 3.1 (ASA)

① *Set $k = 0$ and initialize $\widehat{\Theta}^0$.*

② *Do n iterations ($n \geq 1$) of SA sampling from $P(\omega \mid \mathcal{F}, \widehat{\Theta}^k)$. The resulting labeling is denoted by $\widehat{\omega}^{k+1}$.*

③ *Update the current estimate of the parameters, $\widehat{\Theta}^{k+1}$ to the ML estimate based on the current labeling $\widehat{\omega}^{k+1}$.*

④ *Goto Step ② with $k = k + 1$ until $\widehat{\Theta}$ stabilizes.*

If ML estimate is not tractable, which is often the case when dealing with MRF models, one can use an approximation (Maximum Pseudo Likelihood (MPL), for instance). We remark that a similar algorithm has been reported in [2]. It uses ICM instead of SA in Step ②.

4 Unsupervised Segmentation Using a Hierarchical Model

Considering the segmentation model presented in [9, 10] (see Figure 1), we have the following logarithmic likelihood function:

$$\begin{aligned}
& \sum_{i=0}^{M} \sum_{s^i \in \mathcal{S}^i} \sum_{s \in b^i_{s^i}} \left(-\ln(\sqrt{2\pi}\sigma_{\omega_s}) - \frac{(f_s - \mu_{\omega_s})^2}{2\sigma^2_{\omega_s}} \right) \\
& - \beta \underbrace{\sum_{i=0}^{M} q^i \sum_{C^i \in \mathcal{C}^i} \delta(\widehat{\omega}_{C^i})}_{N^{ih}(\widehat{\omega})} - \gamma \underbrace{\sum_{C \in \bar{\mathcal{C}}_3} \delta(\widehat{\omega}_C)}_{\bar{N}^{ih}(\widehat{\omega})} - \ln(Z(\beta, \gamma))
\end{aligned} \tag{9}$$

where q^i is the number of cliques between two neighboring blocks at scale $\mathcal{B}^i$. $N^{ih}(\widehat{\omega})$ denotes the number of inhomogeneous cliques siting at the same scale and $\bar{N}^{ih}(\widehat{\omega})$ denotes the number of inhomogeneous cliques siting astride two neighboring levels in the pyramid. First, let us consider the first term:

$$\sum_{i=0}^{M} \sum_{s^i \in \mathcal{S}^i} \sum_{s \in b^i_{s^i}} \left(-\ln(\sqrt{2\pi}\sigma_{\omega_s}) - \frac{(f_s - \mu_{\omega_s})^2}{2\sigma^2_{\omega_s}} \right)$$

$$= \sum_{\lambda \in \Lambda} \sum_{i=0}^{M} \sum_{s^i \in \mathcal{S}^i_\lambda} \sum_{s \in b^i_{s^i}} \left(-\ln(\sqrt{2\pi}\sigma_\lambda) - \frac{(f_s - \mu_\lambda)^2}{2\sigma^2_\lambda} \right) \tag{10}$$

where $\mathcal{S}^i_\lambda$ is the set of sites at level i where $\widehat{\omega}_{s^i} = \lambda$. Derivating with respect to μ_λ and σ_λ, we get:

$\forall \lambda \in \Lambda$:

$$\mu_\lambda = \frac{1}{\sum_{i=0}^{M} | \mathcal{S}^i_\lambda |} \sum_{i=0}^{M} \sum_{s^i \in \mathcal{S}^i_\lambda} \sum_{s \in b^i_{s^i}} f_s$$

$$\sigma^2_\lambda = \frac{1}{\sum_{i=0}^{M} | \mathcal{S}^i_\lambda |} \sum_{i=0}^{M} \sum_{s^i \in \mathcal{S}^i_\lambda} \sum_{s \in b^i_{s^i}} (f_s - \mu_\lambda)^2 \tag{11}$$

Notice that a grey-level value f_s may be considered several times. More precisely, f_s is considered m-times in the above sum for a given λ if there is m scales where $\widehat{\omega}$ assigns the label λ to the site s. m can also be seen as a weight. Obviously, the more s has been labeled by λ at different levels, the more is probable that s belongs to class λ and hence its grey-level value f_s characterizes better the class λ. The derivates of the logarithmic likelihood function with respect to β and γ are given by:

$$\frac{\partial}{\partial \beta} \left(-\beta N^{ih}(\widehat{\omega}) - \ln(Z(\beta,\gamma)) \right) = -N^{ih}(\widehat{\omega}) - \frac{\partial}{\partial \beta} \ln(Z(\beta,\gamma)) \tag{12}$$

$$\frac{\partial}{\partial \gamma} \left(-\gamma \bar{N}^{ih}(\widehat{\omega}) - \ln(Z(\beta,\gamma)) \right) = -\bar{N}^{ih}(\widehat{\omega}) - \frac{\partial}{\partial \gamma} \ln(Z(\beta,\gamma)) \tag{13}$$

From which, we get

$$N^{ih}(\widehat{\omega}) = \frac{\sum_{\omega \in \Omega} N^{ih}(\omega) \exp(-\beta N^{ih}(\omega) - \gamma \bar{N}^{ih}(\omega))}{\sum_{\omega \in \Omega} \exp(-\beta N^{ih}(\omega) - \gamma \bar{N}^{ih}(\omega))} \tag{14}$$

$$\bar{N}^{ih}(\widehat{\omega}) = \frac{\sum_{\omega \in \Omega} \bar{N}^{ih}(\omega) \exp(-\beta N^{ih}(\omega) - \gamma \bar{N}^{ih}(\omega))}{\sum_{\omega \in \Omega} \exp(-\beta N^{ih}(\omega) - \gamma \bar{N}^{ih}(\omega))} \tag{15}$$

The solution of the above equations can be obtained using the following algorithm.

Algorithm 4.1 (Hyperparameter Estimation)

① *Set $k = 0$ and initialize $\widehat{\beta}^0$ and $\widehat{\gamma}^0$. Furthermore, let $N^{ih}(\widehat{\omega})$ denote the number of inhomogeneous cliques at the same scale and $\bar{N}^{ih}(\widehat{\omega})$ denotes the number of inhomogeneous cliques between levels.*

② *Using SA at a fixed temperature T, generate a new labeling η sampling from*

$$P(\mathcal{X} = \omega) = \frac{\exp\left(-\frac{\widehat{\beta}^k}{T} \sum_{i=0}^{M} \sum_{\{s,r\} \in \mathcal{C}^i} \delta(\omega_s, \omega_r)\right)}{Z(\widehat{\beta}^k, \widehat{\gamma}^k)} + \frac{\exp\left(-\frac{\widehat{\gamma}^k}{T} \sum_{\{s,r\} \in \bar{\mathcal{C}}} \delta(\omega_s, \omega_r)\right)}{Z(\widehat{\beta}^k, \widehat{\gamma}^k)}. \tag{16}$$

Compute the number of inhomogeneous cliques $N^{ih}(\eta)$ and $\bar{N}^{ih}(\eta)$ in η.

③ *If $N^{ih}(\eta) \approx N^{ih}(\widehat{\omega})$* and *$\bar{N}^{ih}(\eta) \approx \bar{N}^{ih}(\widehat{\omega})$ then stop, else $k = k + 1$. If $N^{ih}(\eta) < N^{ih}(\widehat{\omega})$ then decrease $\widehat{\beta}^k$, if $N^{ih}(\eta) > N^{ih}(\widehat{\omega})$ then increase $\widehat{\beta}^k$. $\widehat{\gamma}^k$ is obtained in the same way. Continue Step ② with $(\widehat{\beta}^k, \widehat{\gamma}^k)$.*

Hereafter we give the algorithm used for the simulation:

Algorithm 4.2 (Unsupervised Segmentation)

① *Given an image $\mathcal{F}$, initialize μ_λ, σ_λ, β and γ.*

② **(Estimation)** *Using Algorithm 3.1 (ASA), get an estimate $\widehat{\Theta}$ of the parameters.*

③ **(Segmentation)** *Given the parameters $\widehat{\Theta}$, do an ordinary supervised segmentation to get the MAP estimate of the label field given $\mathcal{F}$ and $\widehat{\Theta}$.*

We remark, that in Step ②, the Gaussian parameters were computed considering only the finest level and not the entire pyramid (cf. Equation (11)).

5 Experimental Results

We have tested the proposed hierarchical unsupervised algorithm on noisy synthetic and real images. The algorithms were implemented on a Connection Machine CM200 [8]. We have compared the obtained parameters and segmentation results to the supervised results already presented in [9]. In general, the quality of unsupervised results are as good, or sometimes slightly better, than the results of supervised segmentation. We observed, however, that the unsupervised algorithm is more sensitive to noise than the supervised one. This is due to the initial conditions. In particular the initialization of the mean and the variance of the classes (the initialization of β and γ are not crucial). For example, in the case of the "triangle" image with SNR= $3dB$ one class has been lost. But with SNR= $5dB$, the result is as good as for the supervised algorithm.

Before evaluating the results, let us explain some important points of the implementation. The only parameter which has to be defined by the user is the number of classes (or regions). All the other parameters are estimated automatically from the data. Essentially, we have followed Algorithm 4.2. First, the initial values of the mean and variance have been estimated: we have used a method, proposed by Postaire and Vasseur [16] which consists of the geometrical analysis of the hystogram, regarded as a Gaussian mixture, in order to determine its modes. For the hyperparameters, we have chosen as initial values $\beta = 0.7$ and $\gamma = 0.1$. Experiments show that these initial values are not vital, practically any value between 0.5 and 1 is good for β and a value close to zero is good for γ.

In the next step of Algorithm 4.2, we use the ASA algorithm (see Algorithm 3.1) to iteratively reestimate the parameters. Using ICM, we maximize the a posteriori probability of ω, given the parameter estimates $\widehat{\Theta}^n$. Then, the ML estimate is computed based on the obtained labeling. Another modification is that the Gaussian parameters were computed considering *only the finest level* and not the entire pyramid as explained in Section 4. This is because the variances obtained with the original algorithm were too large. This modification also reduces the computing time.

Once the sequence $\widehat{\Theta}^n$ becomes steady, the estimation step is finished and one proceeds to the segmentation (with known parameters) using the Gibbs sampler, for instance.

The algorithms were tested on the "checkerboard" (Figure 3) and "triangle" (Figure 2) images. We also give the corresponding histogram, since the initial estimates are based on it. In Table 1 and Table 2, we compare the parameters obtained by the unsupervised algorithm to the ones used for the supervised segmentation. We remark that the parameters of the supervised algorithm are not necessarily correct. They have been computed on training sets selected by an expert. In Table 4, we give the computer time of the estimation and segmentation. As we can see, the estimation requires much more time than the segmentation. The hyperparameter estimation requires the largest part of the computer time since it consists of generating new labelings by SA in Step ② of Algorithm 4.1.

Table 3 provides an objective comparison of supervised and unsupervised segmentation results based on the number of misclassified pixels. The obtained results are practically the same for supervised and unsupervised segmentation. Many tests have also been conducted for unsupervised classification of SPOT satellite images given by CNES (French Space Agency). For more details see [11].

Parameter	Unsupervised Initial	Unsupervised Final	Supervised
μ_0	123.5	126.7	119.2
σ_0^2	256.0	903.4	659.5
μ_1	170.0	151.5	149.4
σ_1^2	169.0	689.3	691.4
β	0.7	0.7	0.7
γ	0.1	0.1	0.3

Table 1: Parameters of the "checkerboard" image.

Parameter	Unsupervised Initial	Unsupervised Final	Supervised
μ_0	83.5	84.3	85.48
σ_0^2	256.0	483.9	446.60
μ_1	100.0	115.5	115.60
σ_1^2	169.0	444.6	533.97
μ_2	152.5	146.7	146.11
σ_2^2	676.0	502.1	540.32
μ_3	181.5	177.9	178.01
σ_3^2	100.0	500.0	504.34
β	0.7	1.0	0.7
γ	0.1	0.1	0.1

Table 2: Parameters of the "triangle" image.

6 Conclusion

Developing a completely data-driven algorithm for image classification is an extremely difficult problem. We have presented some iterative unsupervised parallel segmentation algorithm for hierarchical Markovian models. The first results are encouraging but unsupervised algorithms require much more computing time due to the hyperparameter estimation (β and γ). In the current implementation, they are computed using Simulated Annealing, which is very time-consuming. Mean Field approximation would probably result in a faster convergence [13, 17].Another important point is the initialization of the Gaussian parameters for each class. We have noticed that unsupervised algorithms are more sensitive to noise than supervised ones. This sensitivity is due to the bad initialization in the case of noisy images.

In summary, the presented unsupervised algorithms provide results comparable to those obtained by supervised segmentations, but they require much more computing time and they are slightly more sensitive to noise. The main advantage is, of course, that unsupervised methods are completely data-driven. The only input parameter is the number of regions. We believe that, for unsupervised methods, the main problem is still the initialization of the Gaussian parameters. Hence, a natural extension of this work would be to look for more efficient initialization techniques.

References

[1] Y. Bard. *Nonlinear Parameter Estimation*. Academic Press, Inc., 1974.

[2] J. Besag. On the statistical analysis of dirty pictures. *Jl. Roy. Statis. Soc. B.*, 1986.

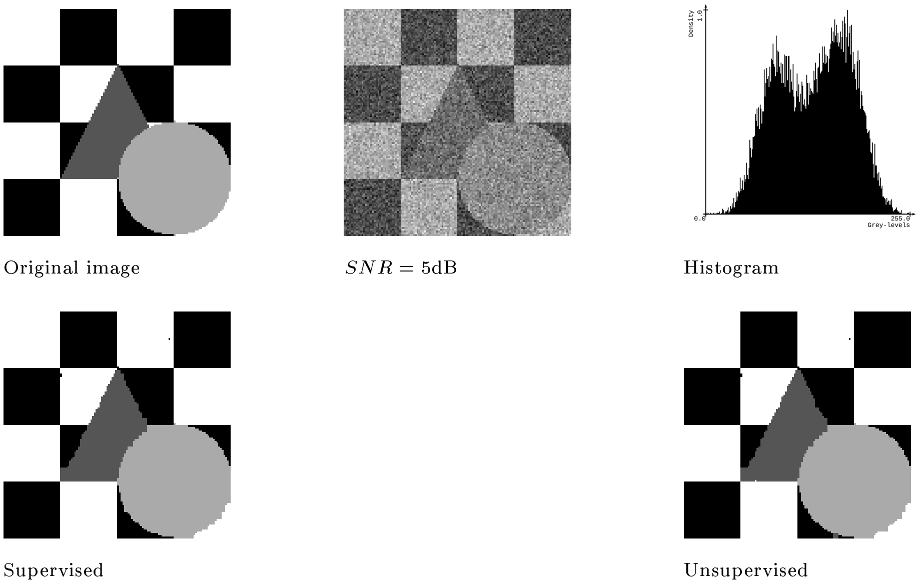

Figure 2: Supervised and unsupervised segmentation results with the Gibbs Sampler.

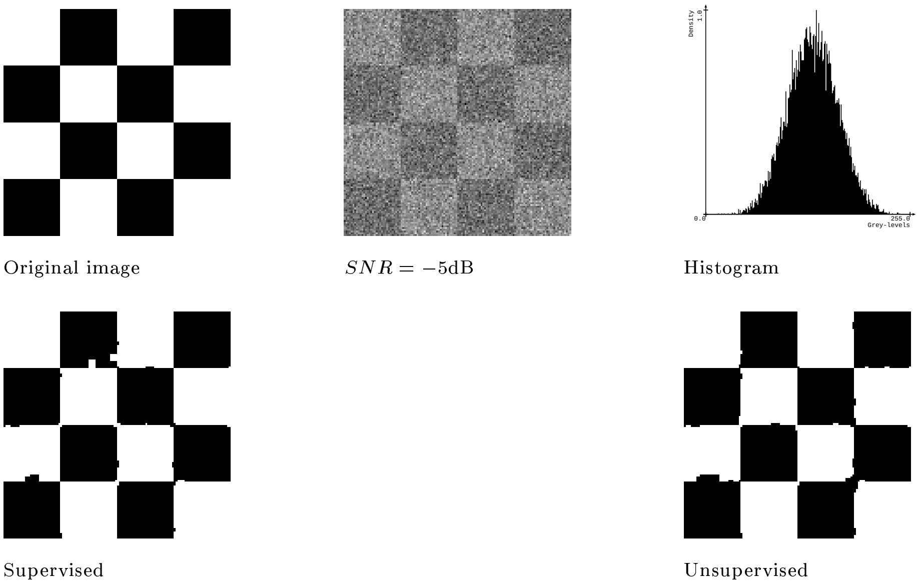

Figure 3: Supervised and unsupervised segmentation results with the Gibbs Sampler.

Image	Supervised	Unsupervised
checkerboard	115 (0.7%)	147 (0.9%)
triangle	104 (0.63%)	111 (0.68%)

Table 3: Comparison of supervised and unsupervised segmentation results (number of misclassified pixels).

Image	VPR	Total CPU time	Estimation	Segmentation
checkerboard	4	1551.93 sec.	1042.46 sec.	446.52 sec.
triangle	4	1762.23 sec.	1232.82 sec.	529.41 sec.

Table 4: Computer times.

[3] H. Caillol, A. Hillion, and W. Pieczynski. Fuzzy Random Fields and Unsupervised Image Segmentation. *IEEE Geoscience and Remote Sensing*, 31(4):801–810, July 1993.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Statist. Soc., ser. B*, vol. 39(1):1–38, 1977.

[5] H. Derin. Estimation Components of Univariate Gaussian Mixtures Using Prony's Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):142–148, January 1987.

[6] K. Fukunaga and T. Flick. Estimation of the Parameters of a Gaussian Mixture Using the Method of Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(4):410–416, July 1983.

[7] D. Geman. Bayesian image analysis by adaptive annealing. In *Proc. IGARSS'85*, pages 269–277, Amherst, USA, Oct. 1985.

[8] W. D. Hillis. The connection machine. *MIT press*, 1985.

[9] Z. Kato, M. Berthod, and J. Zerubia. Multiscale markov random field models for parallel image classification. In *Proc. ICCV*, Berlin, May 1993.

[10] Z. Kato, M. Berthod, and J. Zerubia. Parallel Image Classification using Multiscale Markov Random Fields. In *Proc. ICASSP*, Minneapolis, Apr. 1993.

[11] Z. Kato, J. Zerubia, and M. Berthod. Unsupervised Parallel Image Classification Using a Hierarchical Markovian Model. Research report, INRIA – Sophia Antipolis, April 1995.

[12] S. Lakshmanan and H. Derin. Simultaneous parameter estimation and segmentation of gibbs random fields using simulated annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):799–813, Aug. 1989.

[13] D. A. Langan, K. J. Molnar, J. W. Modestino, and J. Zhang. Use of the Mean-Field Approximation in an EM-Based Approach to Unsupervised Stochastic Model-Based Image Segmentation. In *Proceedings ICASSP'92*, pages III–57–III–60, San Francisco, March 1992.

[14] P. Masson and W. Pieczynski. SEM Algorithm and Unsupervised Statistical Segmentation of Satellite Images. *IEEE Geoscience and Remote Sensing*, 31(3):618–633, May 1993.

[15] W. Pieczynski. Statistical image segmentation. In *Machine Graphics and Vision, GKPO'92*, pages 261–268, Naleczow, Poland, May 1992.

[16] J. G. Postaire and C. P. A. Vasseur. An approximate solution to normal mixture identification with application to unsupervised pattern classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(2):163–179, March 1981.

[17] J. Zerubia and R. Chellappa. Mean field annealing using Compound Gauss-Markov Random fields for edge detection and image estimation. *IEEE Trans. on Neural Networks*, 8(4):703–709, July 1993.