The Mean Field Theory in EM Procedures for Markov Random Fields[1]

Jun Zhang[2]

TR-91-001

January, 1991

Abstract

The EM (expectation maximization) algorithm is a maximum-likelihood parameter estimation procedure for incomplete data problems in which part of the data is hidden, or unobservable. In many signal processing and pattern recognition applications, the hidden data are modeled as Markov processes and the main difficulty of using the EM algorithm for these applications is the calculation of the conditional expectations of the hidden Markov processes. In this paper, we show how the mean field theory from statistical mechanics can be used to efficiently calculate the conditional expectations for these problems. The efficacy of the mean field theory approach is demonstrated on the parameter estimation for one-dimensional mixture data and two-dimensional unsupervised stochastic model-based image segmentation. Experimental results indicate that in the 1-D case, the mean field theory apprach provides comparable results to those obtained by Baum's algorithm, which is known to be optimal. In the 2-D case, where Baum's algorithm can no longer be used, the mean field theory provides god parameter estimates and image segmentation for both synthetic and real-world images.

Index Terms: *Mean field theory, EM algorithm, Markov random fields, parameter estimation, image segmentation.*

[1] This work is supported, in part, by the National Science Foundation Research Opportunity Award (ROA) as an extension of grant MIP-8922354.

[2] Permenent address: Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI 53201.

1. Introduction:

Many problems in signal processing and pattern recognition can be formulated as *incomplete data* problems. In an incomplete data problem, part of the data is not observable, or hidden, and it is necessary to estimate the hidden data and its characteristics. Usually, the observed and hidden data are modeled as either random variables or random processes which are characterized by parametric probability distributions.

Recently, Markov processes have been demonstrated as effective models for the hidden random processes in various incomplete data problems in speech recognition, image processing and computer vision (e.g., see [11]-[12] for surveys in these areas). In these applications, the Markov models capture the underlying physical constraints of the problems, such as the transitions of sound units in a phoneme in speech recognition and the continuity of object surfaces in computer vision. The solution of the incomplete data problems often amounts to estimating model parameters of the distributions of the hidden and observed data as well as estimating the hidden data.

An example of particular interest here is that of stochastic model-based image segmentation in which an observed image is separated into disjoint regions of different statistical properties, called *classes*. This is done through assigning pixels of the image to different classes. A number of Markov random field (MRF) model-based image segmentation algorithms have been proposed and demonstrated successfully in segmenting various real-world images (e.g., see [3]-[7]), such as noisy and textured images. MRF's are multi-dimensional Markov processes and are used here to incorporate physical constraints of the segmentation problem, such as "neighboring pixels should belong to the same region except at region boundaries". In these techniques, the pixel gray levels are modeled as an *observable random field*, and the class assignments of the pixels (the segmentation) are modeled as a *hidden random field*, image segmentation is achieved by finding the optimal estimates (in some sense, e.g., maximum a posteriori) of the hidden states.

Most of the MRF based techniques are *supervised* in that they assume that the parameters can be estimated from training data. However, in practice, training data is often not available and one needs *unsupervised* techniques which estimate the model parameters during segmentation. Even when training data is available, training processes often require operator intervention where an automated system is more desirable. Hence, recent research have be directed towards unsupervised techniques [7]-[10].

The EM (expectation-maximization) algorithm [1]-[2] is an iterative maximum-likelihood (ML) procedure for parameter estimation in incomplete data problems. It can also be used to estimate the hidden random variables or processes.[3] Each iteration of the EM procedure consists of two steps, the E-step (expectation) and the M-step (maximization). In the E-step, the conditional expectation of the likelihood of the hidden data (given the observed data and current estimates of parameters) is calculated. In the M-step, the parameter estimates are updated by maximizing the conditional expectation obtained in the E-step.

A main difficulty in using the EM algorithm when hidden data are modeled as Markov processes is in the conditional expectation calculation of the E-step. For one-dimensional (1-D) hidden Markov processes, Baum et al. [13] proposed a forward-backward procedure which requires intensive computation. For two-dimensional (2-D) or N-dimensional ($N > 2$) hidden Markov processes, i.e., Markov random fields, Baum's algorithm can not be used due to the lack of causality in spaces with more than one dimension. Therefore, while the EM procedures are desirable for two-dimensional applications, such as unsupervised segmentation of images, only approximated versions have been studied, where the approximations are made based on more or less heuristic arguments [7]-[10].

Recently, several researchers [14]-[16] have used the mean field theory [17] from statistical mechanics to calculate the mean (expectation) of MRF's in various computer vision applications. The mean field theory provides a mathematically sound (in some sense optimal [17]) approximation to the mean of an MRF. In this paper, we will show how it can be used in EM procedures for MRF's. After a brief review of the mean field theory in Section 2, we will describe its application in EM procedures and show how these EM procedures can be used in unsupervised image segmentation in Section 3. In Sections 4 and 5, the efficacy of the mean field theory approach is demonstrated on 1-D and 2-D signal/image segmentation and parameter estimation, respectively. In the 1-D case, the results are also compared with those obtained by Baum's algorithm. Finally, a summary is provided in

[3] In this case, the estimates are, in general, not ML but optimal in a different sense.

Section 6.

2. The Mean Field Theory:

A clear and comprehensive treatment of the mean field theory can be found in a text on statistical mechanics by Chandler [17]. In this section, we will briefly review the mean field theory based on the materials in [17]. To make it convenient to address our application problems, we will use the notations established in a previous report [18] rather than those of [17].

We will start the discussion with the concept of an MRF. Let **S** be a set of sites with a neighborhood system defined on it. Typical examples of **S** are given in Fig. 1. In Fig. 1a, **S** is a one-dimensional (1-D) lattice with a "first-order" neighborhood system, where the neighbors of a site i are the two sites that immediately proceed and follow it; in Fig. 1b, **S** is a two-dimensional (2-D) lattice with a "second-order" neighborhood system, where the neighbors of site i are the eight sites that immediately surround it. Let c be a set of sites in **S**, then c is called a *clique* of **S** if it either contains a single site or several sites that are *all* neighbors to each other. Examples of various types of cliques are shown in Fig. 1 for the lattices there.[4]

Let $\mathbf{u} = \{u_i\}$, $i \in \mathbf{S}$, be a collection of random variables defined on **S**. Then **u** is called a MRF if: 1) all of its realizations have non-zero probabilities and 2) its conditional distribution satisfies the following Markov property:

$$p(u_i|\mathbf{u}_{\mathbf{S}-i}) = p(u_i|u_j, j \in \mathbf{N}_i), \tag{1}$$

where $u_{\mathbf{S}-i}$ denotes the field restricted on $\mathbf{S} - i$, and N_i denotes the set of neighbors of i. Let **u** be a MRF. then it is well known (e.g., see [19]) that the joint probability distribution of **u** is a Gibbs distribution, given by

$$p(\mathbf{u}) = Z^{-1} \exp[-\beta U(\mathbf{u})], \tag{2a}$$

[4] For the 2-D lattice, we have only shown the singleton and doubleton cliques which we are going to use in this paper. For more complex clique types, see, e.g., Geman and Geman [22].

where $U(\mathbf{u})$ is the energy function with

$$U(\mathbf{u}) = \sum_{c} V_c(\mathbf{u}), \tag{2b}$$

where $V_c(\mathbf{u})$'s are the clique potentials (for a given clique c, the clique potential only depends on the random variables defined on sites in that clique) and Z is the normalization factor, also called the partition function or free energy, with

$$Z = \sum_{\mathbf{u}} \exp[-\beta U(\mathbf{u})]. \tag{2c}$$

The MRF model was first studied in statistical mechanics (e.g., see [17]). Comprehensive treatments of the MRF model and its applications in signal/image processing and computer vision can be found in [12], [19]-[22].

The mean field theory concerns the following problem: How does one find the mean of the field above? More specifically, how does one find $< u_i >$ for an arbitrary $i \in \mathbf{S}$? where $< \cdot >$ represents expectation, or ensemble average. By definition,

$$\begin{aligned} < u_i > &= \sum_{\mathbf{u}} u_i p(\mathbf{u}) \\ &= Z^{-1} \sum_{\mathbf{u}} u_i \exp[-\beta U(\mathbf{u})]. \end{aligned}$$

However, it is well known that due to the interaction between the u_i's, the calculation of Z and the sum above involve all the possible realizations of the MRF. Therefore, in general, precise calculation of $< u_i >$ is exponentially complex and is not computationally feasible [17].

The mean field theory suggests an approximation to (3) based on the following assumption: the influence of u_j, $j \neq i$, in the calculation of $< u_i >$ can be approximated by the influence of $< u_j >$. For the sake of simplicity, we assume that the interactions between sites are pairwise, that is, a clique potential may be non-zero only when it contains one or two sites. In this case, the energy function can be written as

$$U(\mathbf{u}) = \sum_i \left[V_c(z_i) + \frac{1}{2} \sum_{j \in \mathbf{N}_i} V_c(z_i, z_j) \right], \tag{3}$$

where $V_c(\cdot)$ and $V_c(\cdot, \cdot)$ represent clique potentials for a single site and a pair of neighboring sites, respectively. To use mean field approximation, define a new energy function for site i as

$$\begin{aligned} U_i^{mf}(u_i) &= U(\mathbf{u})|_{u_j = <u_j>, j \neq i} \\ &= U_i^{mf'}(u_i) + R_i^{mf'}(< \mathbf{u}_{\mathbf{S}-i} >), \end{aligned} \tag{4}$$

where the first term contains all the clique potentials related to site i and is called the *mean field local energy* at i, given by

$$U_i^{mf'} = V_c(u_i) + \sum_{j \in \mathbf{N}_i} V_c(u_i, < u_j >), \tag{5a}$$

and the second term, which does not depend on u_i, is given by

$$R_i^{mf'}(< \mathbf{u}_{\mathbf{S}-i} >) = U(\mathbf{u})|_{u_j = <u_j>, j \neq i} - U_i^{mf'}(u_i). \tag{5b}$$

Similarly, define

$$Z_i^{mf} = \sum_{u_i} \exp[-\beta U_i^{mf}(u_i)] = Z_i^{mf'} \exp[-\beta R_i^{mf'}(< \mathbf{u}_{\mathbf{S}-i} >)], \tag{5c}$$

where $Z_i^{mf'}$ is called the *mean field local free energy*, given by

$$Z_i^{mf'} = \sum_{u_i} \exp[-\beta U_i^{mf'}(u_i)]. \tag{5d}$$

Then, by the mean field approximation

$$\begin{aligned} < u_i > &\simeq {Z_i^{mf}}^{-1} \sum_{u_i} u_i \exp[-\beta U_i^{mf}(u_i)] \\ &= {Z_i^{mf'}}^{-1} \sum_{u_i} u_i \exp[-\beta U_i^{mf'}(u_i)], \end{aligned} \tag{6}$$

where terms that do not depend on u_i cancel each other.

In terms of physics, the mean field theory of (4)-(6) suggests that when estimating the mean field at i, the influence of the field at other sites can be approximated by that of their mean; therefore, the fluctuations on these sites are neglected. This is a reasonable assumption when the field is in equilibrium where fluctuations at different sites cancel each other. In terms of mathematics, the mean field theory of (4)-(6) suggests that the marginal distribution of the field at i,

$$p(u_i) = Z^{-1} \sum_{u_{\mathbf{S}-i}} \exp[-\beta U(\mathbf{u})],$$

be approximated by

$$\begin{aligned} p^{mf}(u_i) &= Z_i^{mf^{-1}} \exp[-\beta U_i^{mf}(u_i)] \\ &= Z_i^{mf'^{-1}} \exp[-\beta U_i^{mf'}(u_i)]. \end{aligned}$$

In terms of computation, notice that in order to find the mean field at i, one needs the mean field at the neighbors of i.Therefore, the mean field is usually computed by iterative procedures. Since the calculation of the mean field can be decomposed into local computations, as can be seen from (5)-(6), it can be implemented in parallel. Finally, Geiger and Girosi [14] have proposed an approximation of the partition function Z by neglecting the fluctuations when the interaction of the sites are concerned, that is

$$\begin{aligned} U(\mathbf{u}) &= \sum_c V_c(\mathbf{u}) \\ &= \sum_i (V_c(u_i) + \frac{1}{2} \sum_{j \in \mathbf{N}_i} V_c(u_i, u_j)) \\ &\simeq \sum_i (V_c(u_i) + \frac{1}{2} \sum_{j \in \mathbf{N}_i} V_c(u_i, < u_j >)) \\ &= U^{mf}(\mathbf{u}) \end{aligned} \tag{7a}$$

and

$$Z \simeq Z^{mf} = \sum_{\mathbf{u}} \exp[-\beta U^{mf}(\mathbf{u})]$$

$$= \prod_{i \in \mathbf{S}} \sum_{u_i} \exp[-\beta(V_c(u_i) + \frac{1}{2} \sum_{j \in \mathbf{N}_i} V_c(u_i, < u_j >))]. \tag{7b}$$

This leads to slightly different results in calculating $< z_i >$ from those of [17] which are described in expressions (5)-(6).

3. Application to EM Procedures:

In this section, we describe how the mean field theory can be used to compute the conditional expectations in EM procedures which may otherwise be difficult to compute when the hidden variables are Markov processes. First, we briefly review the EM algorithm, starting with the following notations:

$\mathbf{y} = \{y_i\}, i \in \mathbf{S}$ – observations;

$\mathbf{z} = \{z_i\}, i \in \mathbf{S}$ – hidden states, not observable;

$p(\mathbf{z}|\Phi_\mathbf{z}) = Z^{-1} \exp[-\beta U(\mathbf{z}|\Phi_\mathbf{z})]$ – prior for $\mathbf{z}$;

$p(\mathbf{y}|\Phi_\mathbf{y}, \mathbf{z})$ – likelihood of $\mathbf{y}$;

$\Phi = (\Phi_\mathbf{y}, \Phi_\mathbf{z})$ – the set of parameters for the distribution of $\mathbf{y}$ and $\mathbf{z}$. Here, we assume that the parameters are separable. That is, $\Phi_\mathbf{y} \cap \Phi_\mathbf{z} = \phi$, where ϕ is the empty set.

A. The EM Algorithm:

The problem that the EM algorithm attempts to solve is the following maximum-likelihood estimation (MLE) problem:

$$\hat{\Phi}_{ML} = \arg\max_{\Phi} \log p(\mathbf{y}|\Phi). \tag{8}$$

Notice (8) is more general than the classical MLE problem in that part of the data is not observable.

The EM algorithm is an iterative procedure for solving (8). At each iteration, it consists of two steps:

E-Step: Find the function $Q(\Phi|\Phi^{(p)}) =< \log p(\mathbf{y}|\mathbf{z}, \Phi) + \log p(\mathbf{z}|\Phi)|\mathbf{y}, \Phi^{(p)} >$,

M-Step: Find $\Phi^{(p+1)} = \arg\max_{\Phi} Q(\Phi|\Phi^{(p)})$.

Here, p represents the pth iteration. It has been shown that under some moderate regularity conditions, the estimates converge to ML estimates, at least locally [2].

The main difficulty in using EM procedures in applications where $\mathbf{z}$ is an MRF is that the Q-function of the E-step is hard to calculate due to the interactions between the hidden variables at different sites. Now, this difficulty may be overcome by using the mean field theory. We illustrate this for both 1-D and 2-D MRF's.

B. Mean Field Approximations: 1-D Case

This is the case considered by Baum et al. [13] and is known as the hidden Markov model (HMM). The prior model is

$$\begin{aligned} p(\mathbf{z}|\Phi) &= p(z_n|z_{n-1},\Phi)p(z_{n-1}|z_{n-2},\Phi)\cdots p(z_2|z_1,\Phi)p(z_1|\Phi) \\ &= \exp[\sum_{i=1}^{n} \log p(z_i|z_{i-1},\Phi)], \end{aligned} \tag{9a}$$

where $p(z_1|z_0,\Phi) = p(z_1|\Phi)$. Suppose for any i, $1 \le i \le n$, z_i takes one of K states represented by vectors e_k, $k = 1, 2, \ldots, K$, where e_k is a binary vector with 1 at the kth component, 0 everywhere else[5]. Then the Markov process is a Markov chain. One can represent $\log p(\mathbf{z}|\Phi)$ as

$$\log p(\mathbf{z}|\Phi) = \sum_{i=1}^{n} z_{i-1}^{t} \mathbf{V}(\Phi) z_i, \tag{9b}$$

where $\mathbf{V}(\Phi)$ is a $K \times K$ matrix whose (k,l)th element is $\log p(z_i = e_l|z_{i-1} = e_k)$. Therefore, $V(\Phi)$ is the "log" of the transition matrix of the Markov chain. Here, we assumed that the Markov chain is homogeneous.

In Baum et al.'s model, the likelihood has the following decomposition

$$p(\mathbf{y}|\Phi_{\mathbf{y}},\mathbf{z}) = \prod_{i} p(y_i|\Phi_{\mathbf{y}}, z_i, z_{i-1}). \tag{10}$$

[5] Naturally, the z_i's are K-dimensional vectors such that $z_i = e_k$ for some k.

Therefore, one can write the log likelihood as

$$\log p(\mathbf{y}|\mathbf{z},\Phi) = \sum_{i=1}^{n} z_{i-1}^{t}\mathbf{W}(y_i,\Phi)z_i, \tag{11}$$

where $\mathbf{W}(y_i,\Phi)$ is a $K \times K$ matrix with the (k,l) component being $\log p(y_i|z_{i-1}=e_k, z_i = e_l, \Phi)$. Now, the Q-function can be written as

$$\begin{aligned} Q(\Phi|\Phi^{(p)}) &= < \log p(\mathbf{y}|\mathbf{z},\Phi) + \log p(\mathbf{z}|\Phi)|\mathbf{y},\Phi^{(p)} > \\ &= \sum_{i=1}^{n} < z_{i-1}^{t}\mathbf{W}(y_i,\Phi)z_i|\mathbf{y},\Phi^{(p)} > \\ &+ \sum_{i=1}^{n} < z_{i-1}^{t}\mathbf{V}(\Phi)z_i|\mathbf{y},\Phi^{(p)} > . \end{aligned} \tag{12}$$

It is not difficult to see that the calculation of (12) lies in calculating the conditional expectation of the form $< z_{i-1,k}z_{i,l}|\mathbf{y},\Phi^{(p)} >$ (since the $\mathbf{W}$ and $\mathbf{V}$ matrices above contain only constants with respect to the conditional expectation), where $z_{i,l}$ is the lth component of vector z_i, $1 \leq l \leq K$. This may be achieved by using the mean field theory. However, a slight extension of the theory is needed, since in Section 2 the mean field theory provides only $< z_i|\mathbf{y},\Phi^{(p)} >$. To proceed along this direction, we first notice that conditioned on $\mathbf{y}$ and $\Phi^{(p)}$, $\mathbf{z}$ is a Markov random field (1-D). In other words, the posterior

$$\begin{aligned} p(\mathbf{z}|\mathbf{y},\Phi^{(p)}) &\sim p(\mathbf{y},\mathbf{z}|\Phi^{(p)}) \\ &= p(\mathbf{y}|\mathbf{z},\Phi^{(p)})p(\mathbf{z}|\Phi^{(p)}) \\ &= \exp[\sum_{i=1}^{n}(\log p(y_i|z_i,z_{i-1},\Phi^{(p)}) + \log p(z_i|z_{i-1},\Phi))] \\ &= \exp[\sum_{i=1}^{n} z_{i-1}^{t}\mathbf{W}(y_i,\Phi^{(p)})z_i + \sum_{i=1}^{n} z_{i-1}^{t}\mathbf{V}(\Phi^{(p)})z_i] \end{aligned} \tag{13a}$$

is a Gibbs distribution. Indeed, comparing (13) with (2a)-(2c), the energy function can be identified through

$$\beta U(\mathbf{z}|\mathbf{y},\Phi^{(p)}) = -\sum_{i=1}^{n} z_{i-1}^{t}\mathbf{W}(y_i,\Phi^{(p)})z_i - \sum_{i=1}^{n} z_{i-1}^{t}\mathbf{V}(\Phi^{(p)})z_i \tag{13b}$$

and clique functions can be identified through

$$\beta V_c(z_{i-1}, z_i|\mathbf{y}, \Phi^{(p)}) = -z_{i-1}^t \mathbf{W}(y_i, \Phi^{(p)}) z_i - z_{i-1}^t \mathbf{V}(\Phi^{(p)}) z_i. \tag{13c}$$

Therefore, the mean field local energy at site i can be obtained from

$$\begin{aligned}\beta U_i^{mf'}(z_i) = &- < z_{i-1}^t|\mathbf{y}, \Phi^{(p)} > \mathbf{W}(y_i, \Phi^{(p)}) z_i - z_i^t \mathbf{W}(y_{i+1}, \Phi^{(p)}) < z_{i+1}|\mathbf{y}, \Phi^{(p)} > \\ &- < z_{i-1}^t \mathbf{y}, \Phi^{(p)} > \mathbf{V}(\Phi^{(p)}) z_i - z_i^t \mathbf{V}(\Phi^{(p)}) < z_{i+1}|\mathbf{y}, \Phi^{(p)} > .\end{aligned} \tag{14}$$

However, our interest here is to calculate the mean field at a pair of sites, $< z_{i-1,k} z_{i,l}|\mathbf{y}, \Phi^{(p)} >$, or more generally, $< z_{i-1} z_i^t|\mathbf{y}, \Phi^{(p)} >$. Using the same principle that gave us the mean field local energy for z_i, we arrive at the following *mean field pairwise local energy* for z_{i-1} and z_i

$$\begin{aligned}\beta U_{i-1,i}^{mf'} = &- < z_{i-2}|\mathbf{y}, \Phi^{(p)} > \mathbf{W}(y_{i-1}, \Phi^{(p)}) z_{i-1} - z_{i-1} \mathbf{W}(y_i, \Phi^{(p)}) z_i \\ &- z_i \mathbf{W}(y_{i+1}, \Phi^{(p)}) < z_{i+1}|\mathbf{y}, \Phi^{(p)} > - < z_{i-2}|\mathbf{y}, \Phi^{(p)} > \mathbf{V}(\Phi^{(p)}) z_{i-1} \\ &- z_{i-1} \mathbf{V}(\Phi^{(p)}) z_i - z_i \mathbf{V}(\Phi^{(p)}) < z_{i+1}|\mathbf{y}, \Phi^{(p)} > .\end{aligned} \tag{15}$$

Thus, the pairwise conditional expectations can be computed using the mean field at single sites.

Notice that the mean field calculation is an approximation, unlike Baum's exact solution. However, results in statistical mechanics have shown that the mean field theory provides good approximations in solving a number of problems, such as the prediction of critical temperature in phase transition for multi-dimensional lattice gas models [17]. An advantage of the mean field approximation is its computational simplicity. The computation usually takes only a few iterations to converge and can easily be implemented in parallel. Finally, once the conditional expectations are obtained through mean field approximations, the M-step, that is, the maximization of the Q function with respect to the parameters, is straightforward, as is described in much of the literature [1]-[2], [11], [13]. In Sections 4 and 5, we will describe formulas used in the M-step for some specific observation and MRF models.

C. Mean Field Approximations: 2-D Case

This is the case in which $\mathbf{z}$ is an MRF. For simplicity, we still assume that each z_i is a binary vector of length K with one component being 1 and all others being 0. Without loss of generality, we restrict the interaction between sites to pairwise interaction between neighbors. Then we can write the prior of the MRF as

$$\begin{aligned} p(\mathbf{z}|\Phi) &= Z^{-1}\exp[-\beta U(\mathbf{z}|\Phi)] \\ &= Z^{-1}(\Phi)\exp[-\beta \sum_i (V_c(z_i|\Phi) + \frac{1}{2}\sum_{j\in \mathbf{N}_i} V_c(z_i, z_j|\Phi))]. \end{aligned} \tag{16a}$$

The log of this prior is then

$$\begin{aligned} \log p(\mathbf{z}|\Phi) &= -\beta \sum_i \left[V_c(z_i|\Phi) + \frac{1}{2}\sum_{j\in \mathbf{N}_i} V_c(z_i, z_j\Phi) \right] - \log Z(\Phi) \\ &= -\beta \sum_i \left[z_i^t \mathbf{V}_1(\Phi) + \frac{1}{2}\sum_{j\in \mathbf{N}_i} z_i^t \mathbf{V}_2(\Phi) z_j^t \right] - \log Z(\Phi), \end{aligned} \tag{16b}$$

where $\mathbf{V}_1(\Phi)$ is a K-dimensional column vector whose kth component is $V_c(z_i = e_k|\Phi)$ and $\mathbf{V}_2(\Phi)$ is a $K \times K$ matrix whose (k, l)th component is $V_c(z_i = e_k, z_j = e_l)$. Here, again, we have assumed that the field is homogeneous and isotropic.[6] For the sake of simplicity, we assume that the likelihood has the form

$$p(\mathbf{y}|\mathbf{z}, \Phi) = \prod_i^n p(y_i|z_i, \Phi). \tag{17}$$

Then the Q function can be written as

$$\begin{aligned} Q(\Phi|\Phi^{(p)}) &=< \log p(\mathbf{y}|\mathbf{z}, \Phi) + \log p(\mathbf{z}|\Phi)|\mathbf{y}, \Phi^{(p)} > \\ &= \sum_i < z_i^t|\mathbf{y}, \Phi^{(p)} > \mathbf{W}_1(y_i, \Phi) - \beta \sum_i < z_i^t|\mathbf{y}, \Phi^{(p)} > \mathbf{V}_1(\Phi) \\ &- \frac{\beta}{2}\sum_i \sum_{j\in \mathbf{N}_i} < z_i^t \mathbf{V}_2(\Phi) z_j|\mathbf{y}, \Phi^{(p)} > - \log Z(\Phi), \end{aligned} \tag{18}$$

[6] This means the clique functions are the same for neighboring pixels in different orientations. The extension to the non-isotropic case is straightforward, as will be described in Section 5.

where $\mathbf{W}_1(y_i, \Phi)$ is a K-dimensional column vector whose kth component is $\log p(y_i|z_i = e_k, \Phi)$. Again, the calculation in (18) requires the calculation of $< z_i|\mathbf{y}, \Phi^{(p)} >$ and $< z_i z_j^t|\mathbf{y}\Phi^{(p)} >$, where $j \in \mathbf{N}_i$. To solve this problem with the mean field theory, we proceed as in part B by first deriving the mean field local energy and mean field pairwise local energy from the posterior distribution

$$\begin{aligned} & p(\mathbf{z}|\mathbf{y}, \Phi^{(p)}) \sim p(\mathbf{y}|\mathbf{z}, \Phi^{(p)}) p(\mathbf{z}|\Phi^{(p)}) \\ & = Z^{-1}(\Phi^{(p)}) \exp\Big[-\beta \sum_i \Big(-\frac{1}{\beta} \log p(y_i|z_i, \Phi^{(p)}) + \\ & + z_i^t \mathbf{V}_1(\Phi^{(p)}) + \frac{1}{2} \sum_{j \in \mathbf{N}_i} z_i^t \mathbf{V}_2(\Phi^{(p)}) z_j \Big) \Big]. \end{aligned} \tag{19}$$

Therefore, the mean field local energy is

$$U_i^{mf'} = -\frac{1}{\beta} \log p(y_i|z_i, \Phi^{(p)}) + z_i^t \mathbf{V}_1(\Phi^{(p)}) + \sum_{j \in \mathbf{N}_i} z_i^t \mathbf{V}_2(\Phi^{(p)}) < z_j|\mathbf{y}, \Phi^{(p)} > \tag{20}$$

and the pairwise local mean field energy is

$$\begin{aligned} U_{i,j}^{mf'} & = -\frac{1}{\beta} \log p(y_i|z_i, \Phi^{(p)}) - \frac{1}{\beta} \log p(y_j|z_j, \Phi^{(p)}) \\ & + z_i^t \mathbf{V}_1(\Phi^{(p)}) + z_j^t \mathbf{V}_1(\Phi^{(p)}) + z_i \mathbf{V}_2(\Phi^{(p)}) z_j \\ & + \sum_{i' \in \mathbf{N}_i, i' \neq j} z_i^t \mathbf{V}_2(\Phi^{(p)}) < z_{i'}|\mathbf{y}, \Phi^{(p)} > + \sum_{j' \in N_j, j' \neq i} z_j^t \mathbf{V}_2(\Phi^{(p)}) < z_{j'}|\mathbf{y}, \Phi^{(p)} > . \end{aligned} \tag{21}$$

When the conditional expectations are computed by using the local mean field energy functions and local free energy, we can proceed to the M-step. Since we have assumed that the model parameters for the observed data and hidden variables are separable, they can be estimated separately. Under the conditional independent assumption of (17), the estimation of the model parameters of the observed data, $\Phi_{\mathbf{y}}$, can be obtained in the same way as that for the case where $\mathbf{z}$ is an independent random field, a case described in previous work [1]-[2]. However, the estimation of the parameters of the hidden MRF, $\Phi_{\mathbf{z}}$, is more involved. According to the EM algorithm, at the M-step, one wants to find the estimates of the MRF parameters, $\hat{\Phi}_{\mathbf{z}}$, through

$$\begin{aligned}\hat{\Phi}_{\mathbf{z}}^{(p+1)} &= \arg\max_{\Phi_{\mathbf{z}}} < \log p(\mathbf{z}|\Phi_{\mathbf{z}}) > \\ &= \arg\max_{\Phi_{\mathbf{z}}} \Big\{ < -\beta U(\mathbf{z}|\Phi_{\mathbf{z}}) > - \log Z(\Phi_{\mathbf{z}}) \Big\}, \end{aligned} \tag{22}$$

where $< \cdot >$ is the conditional expectation conditioned on $\mathbf{y}$ and $\Phi^{(p)}$. The maximization of (22) involves the calculation of the partition function, which is difficult for general MRF models. This difficulty, however, may be avoided by using a mean field approximation of $p(\mathbf{z}|\Phi_{\mathbf{z}})$. In [17], Chandler described the following mean field approximation

$$p_{MF}(\mathbf{z}|\Phi_{\mathbf{z}}) = Z_{MF}^{-1}(\Phi_{\mathbf{z}}) \exp[-\beta U_{MF}(\mathbf{z}|\Phi_{\mathbf{z}})], \tag{23a}$$

where

$$U_{MF}(\mathbf{z}|\Phi_{\mathbf{z}}) = \sum_i \Big[z_i^t \mathbf{V}_1(\Phi_{\mathbf{z}}) + \sum_{j \in \mathbf{N}_i} z_i^t \mathbf{V}_2(\Phi_{\mathbf{z}}) < z_j > \Big] \tag{23b}$$

and

$$\begin{aligned} Z_{MF}(\Phi_{\mathbf{z}}) &= \sum_{\mathbf{z}'} \exp[-\beta U_{MF}(\mathbf{z}'|\Phi_{\mathbf{z}})] \\ &= \prod_i \sum_{z_i'} \exp\Big[-\beta \sum_i \Big(z_i^t \mathbf{V}_1(\Phi_{\mathbf{z}}) + \sum_{j \in \mathbf{N}_i} z_i^t \mathbf{V}_2(\Phi_{\mathbf{z}}) < z_j > \Big)\Big]. \end{aligned} \tag{23c}$$

Using $p_{MF}(\mathbf{z}|\Phi_{\mathbf{z}})$, the MRF parameters are estimated by

$$\hat{\Phi}_{\mathbf{z}} = \arg\max_{\Phi_{\mathbf{z}}} \Big\{ < -\beta U_{MF}(\mathbf{z}|\Phi_{\mathbf{z}}) > - \log Z_{MF}(\Phi_{\mathbf{z}}) \Big\}. \tag{23d}$$

There are several points worth noting about the mean field approximation of $p(\mathbf{z})$, $p_{MF}(\mathbf{z})$. First of all, the partition function of $p_{MF}(\mathbf{z})$, with its factorization of (23c), is easy to compute. Secondly, Chandler has shown [17] that the approximations in (23a)-(23d) is optimal with respect to the Gibbs-Bogoliubov-Feynman bound. Geiger and Girosi's approximation in expressions (7a)-(7b) is also an approximation of $p(\mathbf{z})$. In our experiments, we found that Chandler's formula provides better estimates than those obtained by Geiger and Girosi's, which seems to justify its optimality. Finally, it is interesting to notice that if $< z_j >$ is replaced by z_j, $p_{MF}(\mathbf{z})$ becomes the pseudo-likelihood of Besag [12].

4. Experiments on 1-D Data:

We have performed some experiments to observe the performance of the mean field theory approach[7] to parameter estimation and (hidden) state estimation in incomplete data problems where the hidden data is modeled by Markov models. Using notations introduced in Section 2, in parameter estimation the model parameters, Φ, are estimated based on the observations $\mathbf{y}$, and, in state estimation the hidden states, $\mathbf{z}$, are estimated based on $\mathbf{y}$ and the estimated Φ. State estimation is often referred to as signal classification or signal segmentation, since by estimating $\mathbf{z}$, we are assigning observed samples, y_i's, to different classes.

The results presented in this section are on the unsupervised classification/segmentation of 1-D mixture signals, where the observations (y_i's) are generated from different pdf's depending on the states (z_i's). While our goal is to apply the mean field theory to 2-D image segmentation (results presented in Section 5), there are two reasons for studying 1-D signals before 2-D images. First, 1-D experiments require much less programming, debugging, and computation, while their results still provide insights into corresponding 2-D problems, in this case image segmentation, since the mean field theory is the same for 1-D and 2-D. Secondly, in the 1-D case, the mean field theory results can be compared with those obtained using Baum's algorithm, which is known to be optimal for 1-D signals but does not apply to multi-dimensional problems.

A. Experiment Description:

A typical data set used in our experiments is shown in Fig. 2, where a two-class hidden state sequence (Fig. 2a) generates a two-class Gaussian mixture observation sequence

[7] Here, by the mean field theory approach, we refer to the EM procedure in which the mean field theory is used to calculate the conditional expectations in the E-step.

(Fig. 2b). The lengths of the sequences for all data sets are 1000 points. The hidden state sequence, $\mathbf{z} = \{z_i\}$, is generated from a two-state Markov chain with transition probabilities p_{kl}, where $k = 1, 2$, $l = 1, 2$. The observation sequence, $\mathbf{y} = \{y_i\}$, is generated from conditional Gaussian pdf

$$p(y_i|z_i) = p(y_i|\mathbf{a}_k), \text{ if } z_i = e_k,\ k = 1, 2, \tag{24}$$

where $\mathbf{a}_k = (m_k, v_k)$ is the parameter vector for the kth class that contains mean m_k and variance v_k. Notice, here we have assumed that the y_i's are conditionally independent given $\mathbf{z}$, which is a common assumption used in speech processing and simpler than the assumption in expression (10) of Section 3.B.

The mean field theory approach is used to estimate the model parameters from the observation sequence and to estimate the hidden state sequence. Its results are compared with those obtained by Baum's algorithm. In parameter estimation, the mean field theory is used to provide the conditional expectations for the E-step. Then the parameter estimates are updated in the M-step by [11]

$$\hat{m}_k^{(p)} = \sum_i < z_{i_k} > y_i \Big/ \sum_i < z_{i_k} > \tag{25a}$$

$$\hat{v}_k^{(p)} = \sum_i \left(< z_{i_k} > y_i - \hat{m}_k^{(p)} \right)^2 \Big/ \sum_i < z_{i_k} > \tag{25b}$$

$$\hat{p}_{kl}^{(p)} = \sum_i < z_{i,k} z_{i+1,l} > \Big/ \sum_i \sum_{l'} < z_{i,k} z_{i+1,l'} > , \tag{25c}$$

where p represents the pth iteration in the EM procedure and $< \cdot >$ represents the conditional expectation conditioned on $\mathbf{y}$ and $\Phi^{(p-1)}$. In hidden state estimation (used for both the mean field theory and Baum's algorithm), the state at time i is estimated as k, if

$$k = \arg \max_{k \in \{1,2\}} < z_{i,k} | \mathbf{y}, \hat{\Phi} > \tag{26}$$

where $\hat{\Phi}$ is the final estimate of the parameters.

In addition to Baum's algorithm, we have also compared the results of the mean field theory approach with those obtained by 1-D versions of several 2-D unsupervised statistical model-based image segmentation algorithms which are listed below:

- Ideal MAP: MAP (maximum a posteriori) classification with ideal, or true parameters;
- KMS MAP: A generalized K-means algorithm in which the hidden states in each iteration are estimated by a MAP procedure and then used to estimate the parameters for the next iteration; this idea is used in [8]. Compared with the EM algorithm, this approach is a "hard decision" scheme in which the estimates of the hidden states, rather than their probabilities, are used in each iteration for parameter estimation as if they are correct;
- EM MAP1: Parameters are estimated using the EM algorithm based on the assumption that the states at different sites are independent even if they are not; in this case, the conditional expectation can be calculated easily; estimated parameters are then used in a MAP classification procedure [10];
- EM MAP2: The same as EM-MAP1 except that during parameter estimation, the marginal probability distribution of the state at a site, $p(z_i|\mathbf{y}, \Phi_{\mathbf{z}}^{(p-1)})$, is approximated by the conditional probability distribution $p(z_i|\mathbf{y}, z_j, j \in \mathrm{N}_i, \hat{\Phi}_{\mathbf{z}}^{(p-1)})$ [10]; this approach becomes the mean field theory approach if we replace z_j by $< z_j >$.

We would like to make two remarks concerning the above 1-D segmentation algorithms. First, the MAP procedures used in the original 2-D algorithms are deterministic approximations to the simulated annealing procedure (see, e.g., [21]). In their 1-D versions, we kept this feature rather than using the Viterbi algorithm [11] which is optimal for 1-D signal segmentation in the MAP sense. The reason is that the Viterbi algorithm, which is a dynamic programming procedure, does not have a 2-D version. If it is used in the above 1-D algorithms, these algorithms will not be the *1-D versions* of the corresponding 2-D algorithms. Secondly, the MAP procedures need the transition probabilities of the Markov chain, yet none of them can estimate the transition probabilities. Therefore, a common practice is to select the parameters heuristically in a way which reflects the continuity of the states over time. In our case, we have chosen $p_{11} = p_{22} = 0.7$ and $p_{12} = p_{21} = 0.3$.[8]

[8] In fact, if $p_{11} = p_{22} = 0.95$, which are the true parameters, are chosen, they often result

Now, we briefly describe the initialization of the segmentation algorithms. All but the ideal MAP need initial values for the model parameters. While other algorithms only need the initial values for the means and variances of the two Gaussian pdf's, the Baum algorithm and the mean field approach also need initial values for the transitional probabilities. In our experiments, the initial values for the means and variances are obtained by a few (usually three) iterations of a K-means clustering procedure [24]. For Baum's algorithm and the mean field approach, the initial value for all the transition probabilities is set to 0.5. Finally, the mean field approach also needs initial values for the mean field at all sites, i.e., $< z_i|\mathbf{y}, \Phi^{(0)} >$, for $i = 1, 2, \ldots, 1000$. We have simply chosen $(0.5, 0.5)^t$ for all $< z_i|\mathbf{y}, \Phi^{(0)} >$.

B. Experimental Results:

The experiments are performed on four data sets which have the same 2-class hidden state sequences, generated with transition probabilities $p_{11} = p_{22} = 0.95$, $p_{12} = p_{21} = 0.05$. The observation sequence for each data set is generated based on the hidden state sequence and two Gaussian pdf's whose means, m_1, m_2, and variances, v_1, v_2, are shown in Table 1 (under "true parameters") and whose plots are shown in Fig. 3. One can see from Fig. 3 that the separation between the two conditional pdf's that generate the observations degraded consistently from data set 1 to data set 4. This suggests that the degree of difficulty in segmenting the observation sequences increases from data set 1 to data set 4.

The results of parameter estimation and state classification/segmentation are shown in Tables 1 and 2, respectively. Here, the results obtained by the mean field theory approach is indicated by MFT. While the quality of the parameter estimation can be readily seen by comparing estimates with true parameters in Table 1, the quality of segmentation is characterized by the *probability of classification error*, p_e, and the *quality of match* (QOM). The probability of classification error is estimated as the percentage of the number of correctly classified data points while the QOM indicates how well the "regions" of points

in segmentations where all the observations are classified into the same class due to error propagation in these iterative MAP procedures. This is especially true for data sets where the separation between the observations from different classes is small, e.g., in data sets 3-4.

(consecutive points from the same class) obtained in segmentation match regions in the true hidden state sequence. The first number in a QOM indicates the number of "good" regions in a segmentation, where a good region, as shown in Fig. 4, is one that agrees (has the same class) with a region in the true state sequence (perfect segmentation) by more than M points (in our case $M = 5$). The second number indicates the total number of correctly classified points that are in good regions. The smaller the classification error probability and the larger the second number in the QOM, the better the segmentation.

The experimental results are summarized as follows. For parameter estimation, all the techniques provide reasonably good estimations of the means and variances of the Gaussian distributions except for KMS-MAP which often underestimates the variances. For the parameters of the Markov chain, only Baum's algorithm and the mean field theory can provide estimates. Their results are quite close to each other. In fact, as shown in Table 1, the mean field theory seems to provide better estimates of the transition probabilities in data sets 2-4. This is due to the similar stopping criteria for the mean field approach and Baum's algorithm. For both algorithms, the computation will stop if the Euclidean distance between parameters obtained in consecutive iterations is less than a small number ϵ ($\epsilon = 0.1$ in our experiments). Indeed, in the experiments we have observed that if we decrease the ϵ for Baum's algorithm, that is, increase the number of iterations, its results did "catch up" with those of the mean field. Therefore, our observations suggest that the mean field theory provides estimates that are comparable to those from Baum's algorithm.

For sample classification or segmentation, Baum's algorithm and the mean field theory perform best among all techniques for data sets 1-3. In the case of data set 4, where the observations from the two classes are heavily overlapped, the EM MAP1 seems to be the most robust. Again, the performance of the mean field theory is comparable to that of Baum's algorithm. Finally, we notice that the KMS MAP is the worst.

5. Experiments on Image Segmentation:

As a natural extension of the 1-D experiments described in Section 4, we applied the mean field theory approach to MRF model-based image segmentation. As described in Section 1, an image is segmented by assigning its pixels to a finite number of classes which have different statistical properties. The images are modeled on two-levels. On the first level, the observed pixel intensities are modeled as random fields conditioned on their class status, or states. On the second level, the spatial distribution of the classes are modeled by a MRF which reflects the physical constraints on the segmented regions (e.g., continuity of states for interior points). The mean field theory approach (i.e., the EM algorithm which uses the mean field theory to compute conditional expectations) is used to estimate the model parameters and perform the segmentation. This is straightforward by using the results of Section 3.C and letting the observed image be $\mathbf{y}$, the segmentation be $\mathbf{z}$, and the model parameters be Φ.

A. Experiment Description:

Image segmentation experiments are performed on both synthetic and real-world images using the mean field theory approach. The synthetic images are used to study its performance in the ideal case, where the hidden states are MRF's and the pdf's of the observed image data satisfy the conditional independence assumption made in Section 3, while the real-world images are used to study its practical applicability.

The synthetic images are generated as follows. First, a realization of a prespecified MRF (specifications include the number of classes, model type, and model parameters) is generated. This realization, also referred to as a *region map*, contains disjoint regions of different classes. The regions are then "colored" by realizations of different independent Gaussian random fields according to their classes. The MRF model used here is the model used by Geman and Geman [22] and Lakshmanan and Derin [23]. More specifically, the clique functions are defined as follows:

$$\text{Singleton: } V_c(z_i) = \alpha_k, \text{ if } z_i = e_k,\ k = 1, 2, \ldots, K, \tag{27a}$$

$$\text{Doubleton: } V_c^{(m)}(z_i, z_j) = \gamma_m(2\delta(z_i, z_j) - 1), \tag{27b}$$

where

$$\delta(z_i, z_j) = 0, \text{ if } z_i = z_j,$$

$$\delta(z_i, z_j) = 1, \text{ if } z_i \neq z_j,$$

and m indicates the type of doubleton clique, with $m = 1, 2, 3, 4$ for horizontal, vertical, $3\pi/4$ diagonal and $\pi/4$ diagonal cliques, respectively. While the mean field theory approach described in Section 3.C was for the isotropic MRF model, which is a special case ($\gamma_m = \gamma$, $m = 1, 2, 3, 4$) of this model, it can be extended to this model straightforwardly. Two typical synthetic images (labeled as synthetic image 1 and synthetic image 2) of size 128 × 128 generated this way with two and three classes are shown in Figs. 5 and 6 along with their region maps. The real-world images are digitized aerial photographs of size 256 and a typical image used in our experiments is shown in Fig. 8.

The (conditionally) independent Gaussian model and the second-order non-isotropic MRF are used to model both synthetic and real-world images. The EM algorithm is used to perform both parameter estimation and segmentation. In each iteration, the conditional expectations of the hidden MRF is computed by the mean field theory approach in the E-step and the parameters are updated in the M-step. The update formulas for the Gaussian parameters are the same as those for the 1-D experiments, i.e., expressions (25a)-(25c), where the summation now is over all the pixels. The MRF parameters are estimated by using expression (23d). In the experiments, we have performed the maximization by using a conjugate gradient subroutine in the IMSL Math Library [25]. Lastly, the final segmentation is obtained by assigning a pixel at site i to the kth class if

$$k = \arg \max_{1 \leq k \leq K} < z_{i,k} | \mathbf{y}, \hat{\Phi} >, \tag{28}$$

where $\hat{\Phi}$ is the final estimate of Φ from the EM procedure.

A problem of practical interest is whether one really needs to estimate $\Phi_{\mathbf{z}}$, the parameters of the MRF. The estimation, which is performed in each iteration of the EM algorithm, requires a lot of computation time (usually 30-50 iterations of the conjugate gradient procedure where each iteration requires the gradient be updated at all the pixels in the images). Previously, it has been shown that these parameters can be set heuristically to provide good segmentation results [22]. In fact, when the parameters are chosen in such a way that they encourage/discourage certain types of interactions of the hidden variables, the segmentation results are relatively insensitive to moderate variations of the numerical values in these parameters. Perhaps these two approaches (to estimate or not to estimate) reflect the difference between Bayesian and ML approaches (which we will call the total ML). In the Bayesian approach, the parameters for the MRF are considered as prior constraints which are set rather than estimated, while in the total ML approach, the parameters are considered a characteristic of the data and therefore are estimated. We have performed segmentation experiments using both approaches, in the hope that if the Bayesian approach provides comparable results to those of the total ML approach, we can use only the Bayesian approach in the future, since it requires much less computation. The maximum number of iterations for the EM algorithm is set to be 20 for the Bayesian approach and 40 for the total ML approach, since the latter needs more time to arrive at a reasonable set of MRF parameters.

Another problem of practical concern is how the quality of segmentation can be assessed. For real-world images, the probability of error and QOM used in Section 4 can not be calculated directly since the ideal segmentation is not available. Even for a synthetic image, the QOM requires much more computation to "single out" the segmented regions since they are irregularly shaped and there may be many of them. Therefore, the approach we have taken is mainly qualitative rather than quantitative. More specifically, for a real-world image, we will see if the segmentation successfully separates major objects in an image (e.g., objects that are large in size). For a synthetic image, we will see if the segmented regions resemble those in the true region map.

B. Typical Results: Synthetic Images

First, we describe results obtained by the total ML approach which estimates the MRF parameters during segmentation. The results of parameter estimation and segmentation for the two synthetic images are shown in Table 3 and Figs. 5 and 6, respectively. For comparison purposes, in Table 3, we have listed the model parameters used to generate these synthetic images as well as the MRF parameters estimated from their region maps.[9] As can be seen from these results, the EM algorithm provides good estimates for both Gaussian parameters and the MRF parameters. One practical problem that we have encountered in using the IMSL routine for estimating the MRF parameters is that trial-and-error is needed to set the stopping parameters for the routine. This is done by experimenting on one image and then using the same setting for all other images. It seems that the results are relatively insensitive to the settings when they are within a certain range.

To compare the Bayesian approach with the total ML approach, we have also segmented the synthetic images by setting the MRF parameters to be $\alpha_k = 0$ for all k and $\gamma = 0.5$. The results, shown in Figs. 5d and 6d, respectively, are good and comparable to those obtained with estimated MRF parameters. These, and our other results, indicates that it may not be necessary to estimate the MRF parameters.

Finally, we have applied the mean field approach to a synthetic image (labeled as synthetic image 3), shown in Fig. 7, studied by Shen [10] who has used algorithms EM MAP1 and EM MAP2 described in Section 4 to estimate the Gaussian and MRF parameters. The results of parameter estimation in comparison with those of Shen's are shown in Table 3. One can see that the mean field theory produces, overall, better variance estimates than those by EM-MAP1 and EM-MAP2.

C. Typical Results: A Real-World Image

Typical results of the segmentation of real-world images are shown in Fig. 8 for a digitized aerial photograph of size 256 × 256. From the results on synthetic images, we are reasonably confident about the Bayesian approach which presets MRF parameters. Therefore, the image is segmented by the mean field theory approach with the MRF parameters preset as $\alpha_k = 0.0$ for all k (classes are equally likely), $\gamma_m = \gamma$ for $m = 1, 2, 3, 4$ (classes continue equally likely in all directions inside a region).

The number of image classes are estimated to be four by using a cluster validation scheme [26]. In Fig. 4b-d, we have shown segmentation results of the original image of

[9] This will provide a fairer comparison between the estimated and the "true" MRF parameters.

Fig. 8a. The segmentations are obtained with different choices of MRF parameters and β, as summarized in the captions of Figs. 8b-8d. In Figs. 8b and 8c, β is fixed to be 1.0 while γ is varied from 0.2 to 0.5. As a result, Fig. 8b preserves more detail but is more noisy; Fig. 8c is a much smoother segmentation but some of the details of the image are lost. Finally, in Fig. 8d, γ is fixed to be 0.2, and β varies with the number of iterations by

$$\beta^{(p)} = [\beta^{(0)}]^p, \tag{29}$$

where p is the current number of iterations and $\beta^{(0)} = 1.12$. The role of β here is similar to that of the temperature in simulated annealing. As can be seen from Fig. 8d, the result is a nice compromise of those of Fig. 8b and 8c. In all segmentations, major regions that correspond to road, vegetation areas, and lawn are well separated.

As a comparison, in Fig. 9, we have shown a segmentation of the image by the total ML approach where the MRF parameters are estimated during segmentation. In this case, the segmentation produces similar regions to those of Fig. 8. However, the segmentation without estimating MRF parameters produces smoother region boundaries. The rough region boundaries produced by the total ML approach seem to be caused by the negative values of some of the MRF parameters, shown in Table 4, indicating that the total ML approach may not be very robust (since the distribution of the regions in real images, unlike those in the synthetic images, may not be close to a non-isotropic second-order MRF). Therefore, for real images, preset MRF parameters may be preferable.

6. Summary:

In this paper, we have described how the mean field theory can be used to efficiently calculate the conditional expectations in EM procedures where the hidden random variables are modeled as Markov processes. The key idea of the mean field theory is: When calculating the mean field (in our applications, the conditional expectation) at a given pixel position, the influence of the variables at other pixel positions can be approximated by that of the mean field at those positions. This results in a simple iterative procedure for computing the conditional expectations in the E-step for MRF models. In this paper, the efficacy of the mean field theory approach is demonstrated on 1-D and 2-D signal/image segmentation and parameter estimation experiments. In 1-D experiments, the performance of the mean field theory approach is observed to be comparable to that of Baum's algorithm, which is known to be optimal. Furthermore, compared to the performance of the 1-D versions of some previously proposed unsupervised image segmentation algorithms, the mean field theory approach seems consistently better. This gives us reason to believe that a 2-D mean field theory approach will do better than these algorithms in image segmentation. In 2-D experiments, the mean field theory approach is observed to provide good parameter estimates on synthetic test data and good segmentations for synthetic and real-world images.

Acknowledgements:

I would like to thank Prof. Jerry Feldman, Dr. Nelson Morgan of the International Computer Science Institute (ICSI), and Dr. Mike Foster of the National Science Foundation (NSF), for making it possible for me to do this work at the ICSI in the summer of 1990. I would also like to thank Dr. Wolfgang Kuepper and Dr. Nelson Morgan of ICSI for many discussions and suggestions. Finally, I would like to thank the NSF for its support through a ROA award as an extension of grant MIP-8922354.

References

[1] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Roy. Soc. Statist., Series B., No. 1, pp. 1-38, 1977.

[2] R. A. Redner and H. F. Walker, "Mixture densities, maximum-likelihood and the EM algorithm," SIAM Rev., Vol. 26, pp. 195-239, 1984.

[3] C. W. Therrien, T. F. Quatieri and D. E. Dudgeon, "Statistical model-based algorithms for image analysis," Proc. IEEE, Vol. 74, April, 1986.

[4] J. Zhang and J. W. Modestino, "Markov random fields with applications to texture classification and discrimination," Proc. The 20th Annual Conf. on Information Science and Systems, Princeton University, NJ, March, 1986.

[5] H. Derin and H. Elliot, "Modeling and segmentation of noisy and textured images using Gibbs random fields," IEEE Trans. Pattern Anal. Machine Intel., Vol. PAMI-9, pp. 39-55, Jan., 1987.

[6] F. S. Cohen and D. B. Cooper, "Simple, parallel, hierarchical and relaxation algorithms for segmenting non-casual Markovian random field models," Proc. IEEE Pattern Anal. Machine Intel., Vol. PAMI-9, pp. 195-219, March, 1987.

[7] P. A. Kelly, H. Derin, and K. D. Hart, "Adaptive segmentation of speckled images using a hierarchical random field model," IEEE Trans. ASSP, Vol. 36, pp. 1628-1641, Oct., 1988.

[8] J. Zhang, "Two-dimensional stochastic model-based image analysis," Ph.D. Thesis, Rensselaer Polytechnic Institute, Troy, New York, Aug., 1988.

[9] J. F. Silverman and D. B. Cooper, "Bayesian clustering for unsupervised estimation of surface and texture models," IEEE Trans. Pattern Anal. Machine Intel., Vol. 10, pp. 482-485.

[10] Y. Shen, "A statistical model-based approach for cluster validation with applications to image segmentation," MS Thesis, Sept. 1990.

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, Vol. 77, pp. 257-286, Feb., 1989.

[12] J. Marroquin, S. Mitter, and T. Poggio, "Computer vision," J. Amer. Stat. Associa-

tion, Vol. 82, pp. 76-89, March, 1987.

[13] L. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," Ann. Math. Stat., Vol. 41, No.1, pp164-171, 1970.

[14] D. Geiger and Federico Girosi, "Parallel and deterministic algorithms for MRFs: surface reconstruction and integration," AI Memo No. 1114, MIT, June, 1989.

[15] J. Zerubia and R. Chellappa, "Mean field approximation using compound Gauss-Markov random field for edge detection and image restoration," ICASSP'90, pp. 2193-2196.

[16] A. L. Yuille, "Generalized deformable models, statistical physics, and matching problems," *Neural Computation*, Vol. 2, pp. 1-24, 1990.

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

[18] W. Kuepper and J. Zhang, "Image interpretation using Markov random fields," Memo-IT-1, ICSI, June, 1990.

[19] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," J. Roy. Statist. Soc., Series B, Vol. 36, pp. 192-226, 1974.

[20] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications*, Providence, RI: Amer. Math. Soc., 1980.

[21] J. Besag, "On the statistical analysis of dirty pictures," J. Royal Stat. Soc., Series B, Vol. 48, pp. 259-302, 1986.

[22] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," IEEE Trans. PAMI, Vol.6, pp.721-741, Nov., 1984.

[23] S. Lakshmanan and H. Derin, "Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing," IEEE Trans. Pattern Anal. Machine Intel., Vol. 11, pp. 799-813, Aug., 1989.

[24] J. Tou and R. Gonzalez, *Pattern Recognition Principles*, Addison-Wesley, Reading, Mass, 1979.

[25] $Math/Library^{T.M.}$, IMSL Inc., 1989.

[26] J. Zhang and J. W. Modestino, "A statistical model-fitting approach to cluster valida-

tion with applications to image segmentation," IEEE Trans. Pattern Anal. Machine Intel., pp. 1009-1016, Oct., 1990.

Table 1. Results of Parameter Estimation

a. Data set 1

est.	m_1	m_2	v_1	v_2
True	10	5	2	4
KMS MAP	10.06	4.78	1.67	3.45
EM MAP1	10.21	5.24	1.69	4.97
EM MAP2	9.96	4.80	1.99	3.75
Baum	10.02	4.92	1.94	4.06
MFT	10.02	4.90	1.93	3.98

est.	p_{11}	p_{12}	p_{21}	p_{22}
True	0.95	0.05	0.05	0.95
Baum	0.94	0.06	0.05	0.95
MFT	0.95	0.05	0.044	0.956

b. Data set 2

est.	m_1	m_2	v_1	v_2
True	10	8	4	1
KMS MAP	11.05	7.87	1.68	1.03
EM MAP1	10.02	7.96	3.91	1.18
EM MAP2	10.11	7.91	3.78	1.05
Baum	10.21	7.94	3.79	1.06
MFT	10.28	7.95	3.76	1.07

est.	p_{11}	p_{12}	p_{21}	p_{22}
True	0.95	0.05	0.05	0.95
Baum	0.88	0.12	0.09	0.90
MFT	0.95	0.05	0.04	0.96

c. Data set 3

est.	m_1	m_2	v_1	v_2
True	10	8	4	3
KMS MAP	10.63	7.21	1.74	1.63
EM MAP1	10.04	7.96	3.70	3.43
EM MAP2	10.06	7.77	3.58	3.05
Baum	10.21	7.73	3.26	2.92
MFT	10.68	8.02	2.90	3.17

est.	p_{11}	p_{12}	p_{21}	p_{22}
True	0.95	0.05	0.05	0.95
Baum	0.76	0.24	0.20	0.80
MFT	0.91	0.09	0.04	0.96

d. Data set 4

est.	m_1	m_2	v_1	v_2
True	10	8	8	6
KMS MAP	10.88	0.69	7.95	8.88
EM MAP1	10.02	7.51	8.18	7.37
EM MAP2	9.31	7.94	8.21	6.94
Baum	9.74	7.95	7.91	6.89
MFT	10.90	7.20	4.66	4.95

est.	p_{11}	p_{12}	p_{21}	p_{22}
True	0.95	0.05	0.05	0.95
Baum	0.57	0.43	0.42	0.58
MFT	0.71	0.29	0.25	0.75

Table 2. Classification Performance

a. data set 1

Method	P_e	QOM
Ideal MAP	5.0×10^{-2}	(62,928)
KMS MAP	5.2×10^{-2}	(66,892)
EM MAP1	4.8×10^{-2}	(55,900)
EM MAP2	4.7×10^{-2}	(61,907)
Baum	1.2×10^{-2}	(41,973)
MFT	1.3×10^{-2}	(41,973)

b. data set 2

Method	P_e	QOM
Ideal MAP	0.13	(53,742)
KMS MAP	0.17	(50,672)
EM MAP1	0.13	(54,754)
EM MAP2	0.14	(55,734)
Baum	7.4×10^{-2}	(43,899)
MFT	8.5×10^{-2}	(40,891)

c. data set 3

Method	P_e	QOM
Ideal MAP	0.19	(58,708)
KMS MAP	0.26	(59,439)
EM MAP1	0.19	(51,709)
EM MAP2	0.29	(50,363)
Baum	0.17	(62,761)
MFT	0.16	(41,802)

d. data set 4

Method	P_e	QOM
Ideal MAP	0.26	(48,664)
KMS MAP	0.39	(44,323)
EM MAP1	0.25	(40,697)
EM MAP2	0.55	(19,463)
Baum	0.31	(49,357)
MFT	0.27	(64,562)

Table 3. Parameter Estimates for Synthetic Images 1-3

a. synthetic image 1

estimates of Gaussian parameters

		with estimated	with preset
	true parameters	MRF parameters	MRF parameters
class	(mean,var)	(mean,var)	(mean,var)
1	100, 400	99.74, 395.93	99.70, 394.15
2	150, 400	149.78, 408.40	149.80, 407.11

estimates of MRF parameters

	α_1	α_2	α_3	β_1	β_2	β_3
true parameters	0.00	0.00	1.00	1.00	1.00	1.00
estimated in EM procedure	-0.03	0.03	0.88	0.80	0.69	0.56
estimated from the region map	-0.06	0.06	0.87	0.98	0.86	0.80

b. synthetic image 2

estimates of Gaussian parameters

		with estimated	with preset
	true parameters	MRF parameters	MRF parameters
class	(mean,var)	(mean,var)	(mean,var)
1	100, 400	100.06, 409.56	100.33, 417.38
2	150, 400	149.22, 415.35	149.27, 405.61
3	200, 400	199.81, 409.02	199.78, 409.15

estimates of MRF parameters

	α_1	α_2	α_3	β_1	β_2	β_3	β_4
true parameters	0.00	0.00	0.00	1.00	1.00	1.00	1.00
estimated in EM procedure	0.06	0.09	-0.15	0.76	1.00	0.68	0.69
estimated from the region map	-0.04	-0.03	0.08	0.95	0.99	0.69	0.79

c. synthetic image 3

		estimated by	estimated by	estimated by
class	true parameters	EM-MFT	EM-ML	EM-MAP
	(mean, variance)	(mean, variance)	(mean, variance)	(mean, variance)
1	(50, 400)	(49.69, 404.04)	(51.66, 403.70)	(49.86, 361.62)
2	(100, 400)	(99.91, 390.13)	(103.50, 375.61)	(100.07, 330.52)
3	(150, 400)	(149.44, 390.90)	(149.39, 262.00)	(149.85, 320.30)
4	(200, 400)	(199.75, 395.66)	(196.80, 447.76)	(199.77, 382.98)

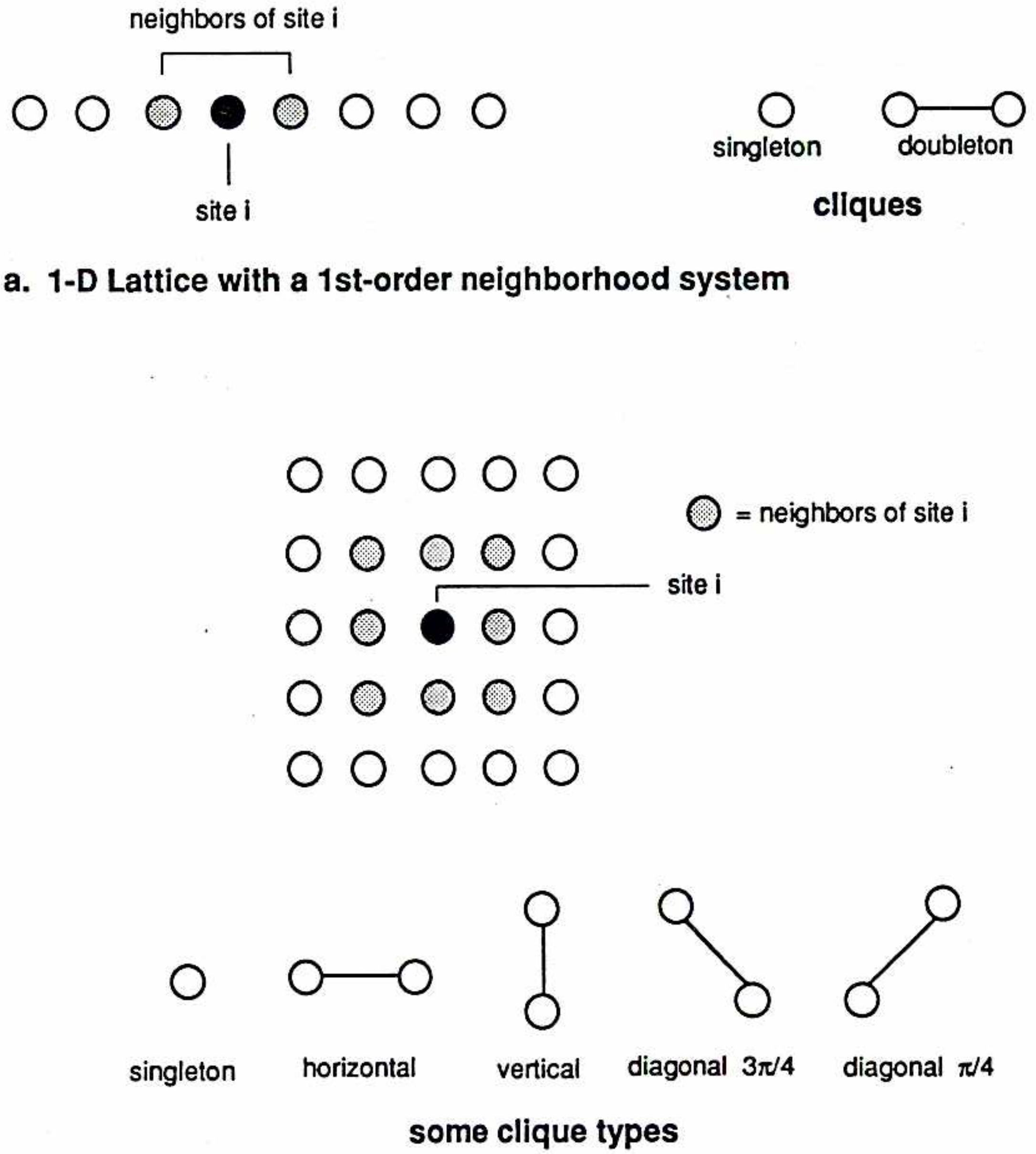
Table 4. Estimated Parameters for the Aerial Photograph

estimates of Gaussian parameters

class	with estimated MRF parameters (mean,var)	with preset MRF parameters (mean,var)
1	210.15, 377.54	215.54, 320.40
2	156.50, 92.78	155.89, 98.90
3	112.95, 342.78	114.25, 183.89
4	185.30, 92.60	183.89, 132.25

estimates of MRF parameters

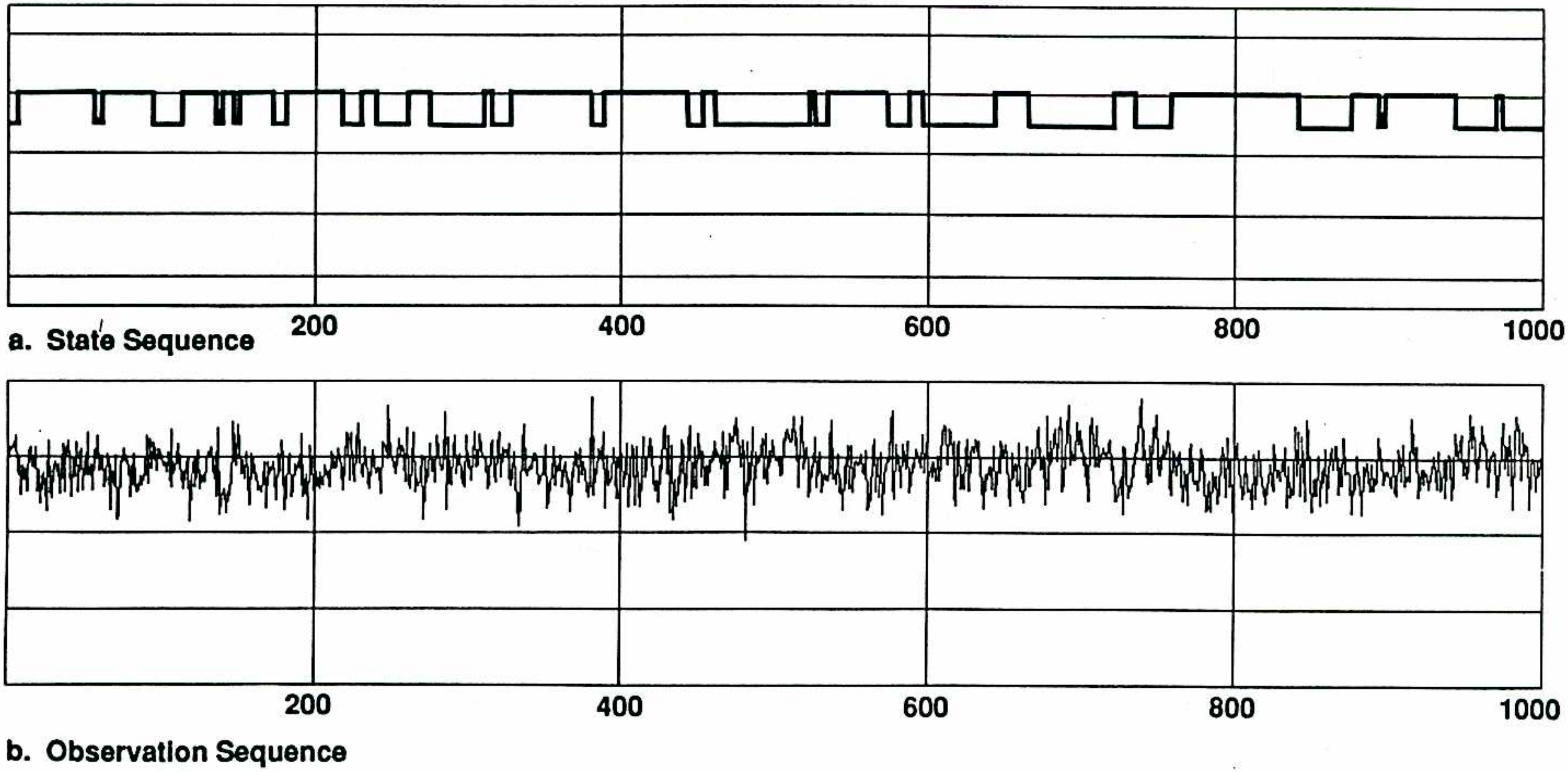
	α_1	α_2	α_3	α_4	β_1	β_2	β_3	β_4
estimated	-0.07	-0.07	0.09	0.05	1.48	1.85	-0.41	-0.54
preset	0.00	0.00	0.00	0.00	0.5	0.5	0.5	0.5



a. 1-D Lattice with a 1st-order neighborhood system

b. 2-D Lattice with a 2nd-order neighborhood system

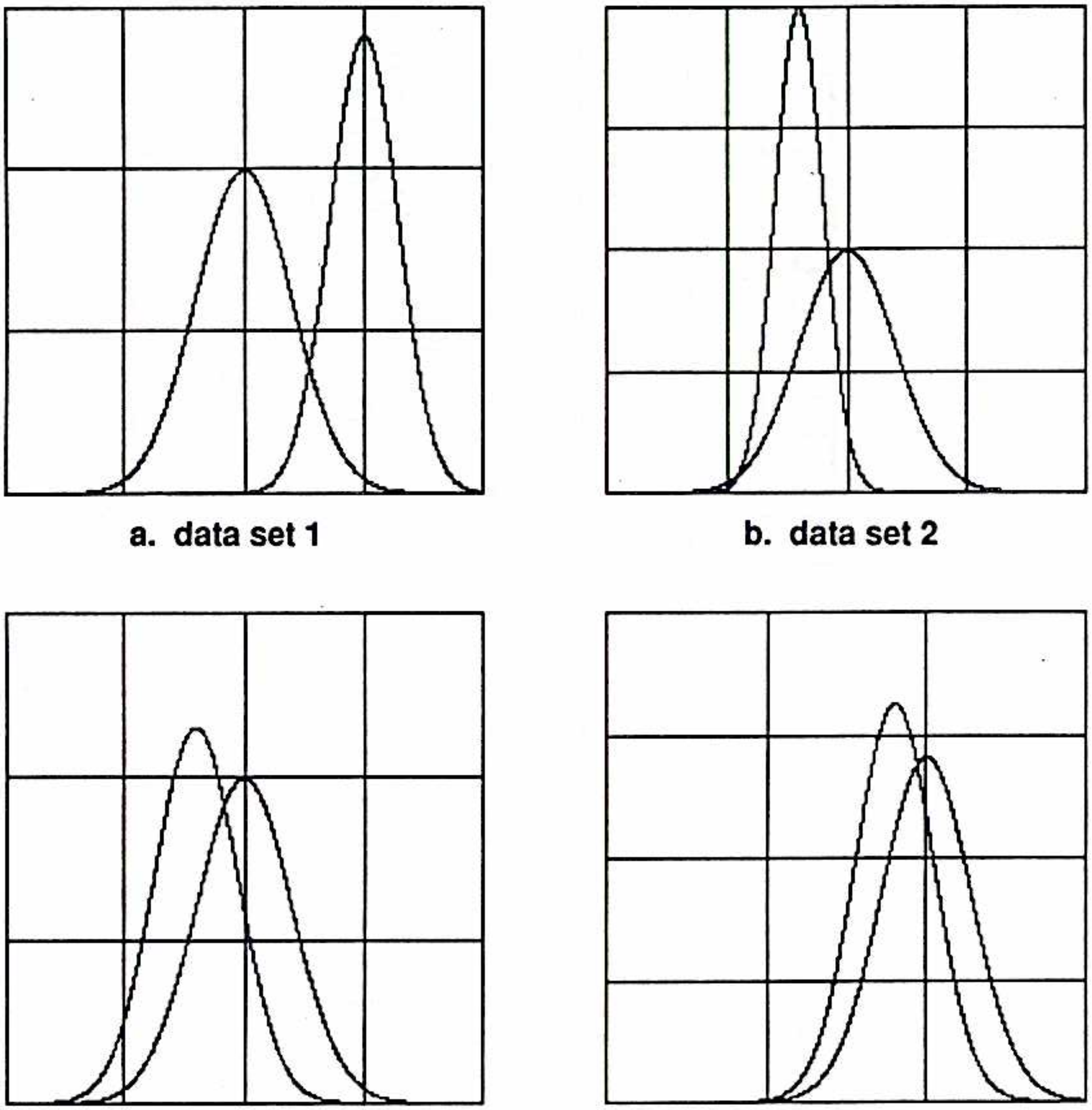
Figure 1. Example of Sets of Sites and Neighborhood Systems



a. State Sequence

b. Observation Sequence

Figure 2 State Sequence and Observation Sequence for Data Set 1



a. data set 1

b. data set 2

c. data set 3

d. data set 4

Figure 3 Gaussian PDFs for the Four Data Sets

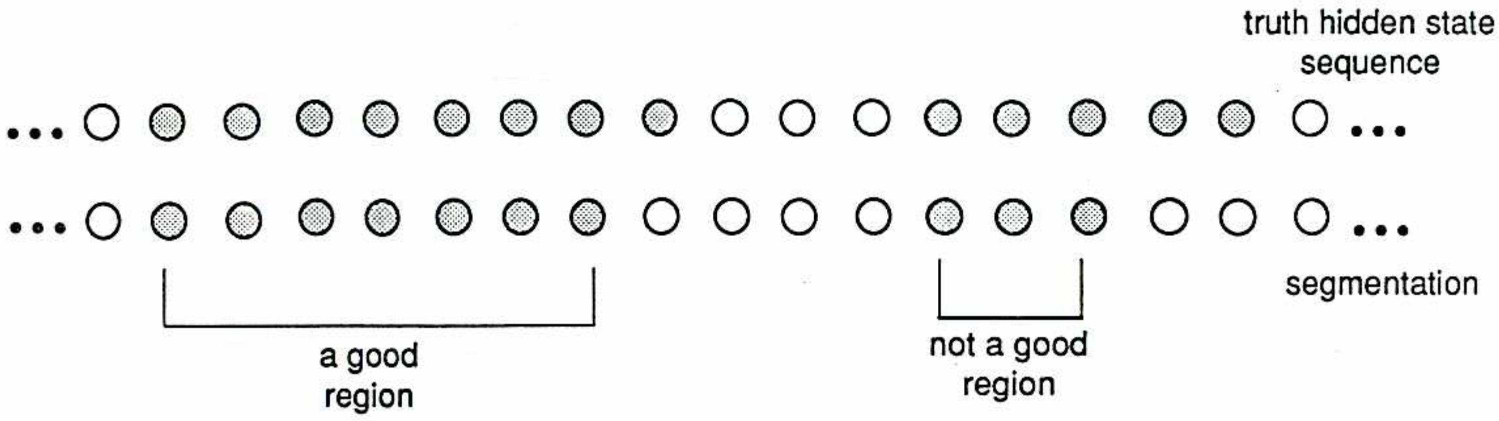
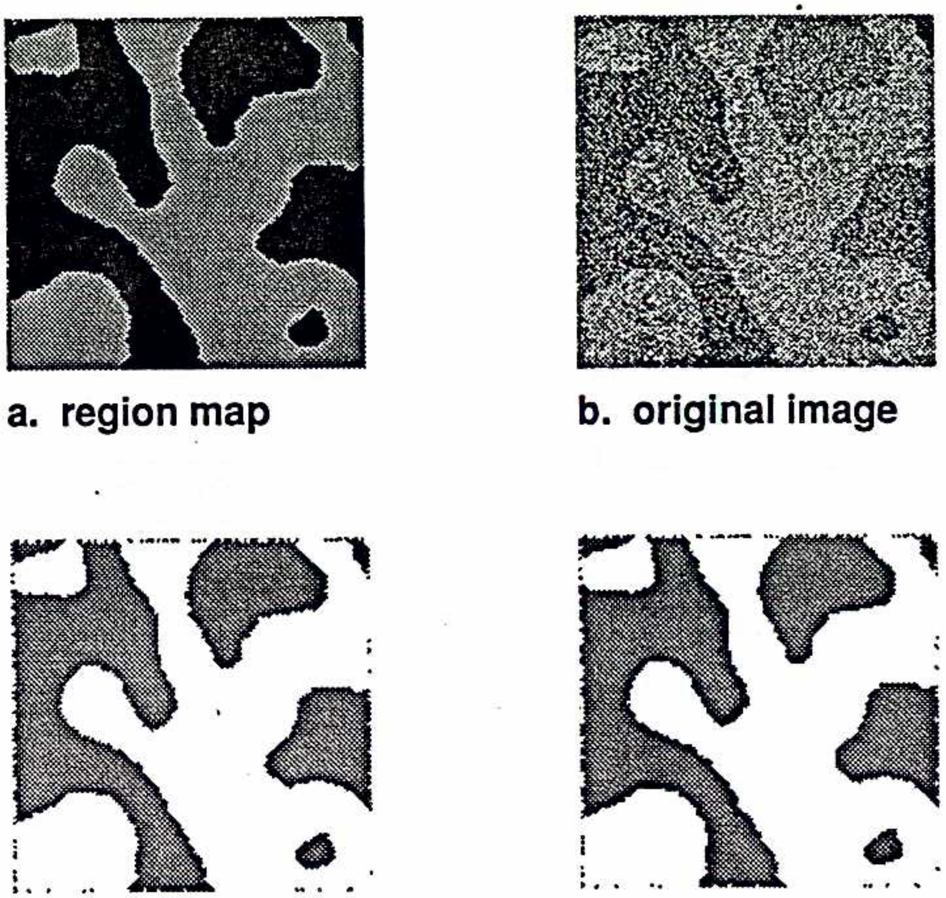


Figure 4. Illustration of QOM for M = 5



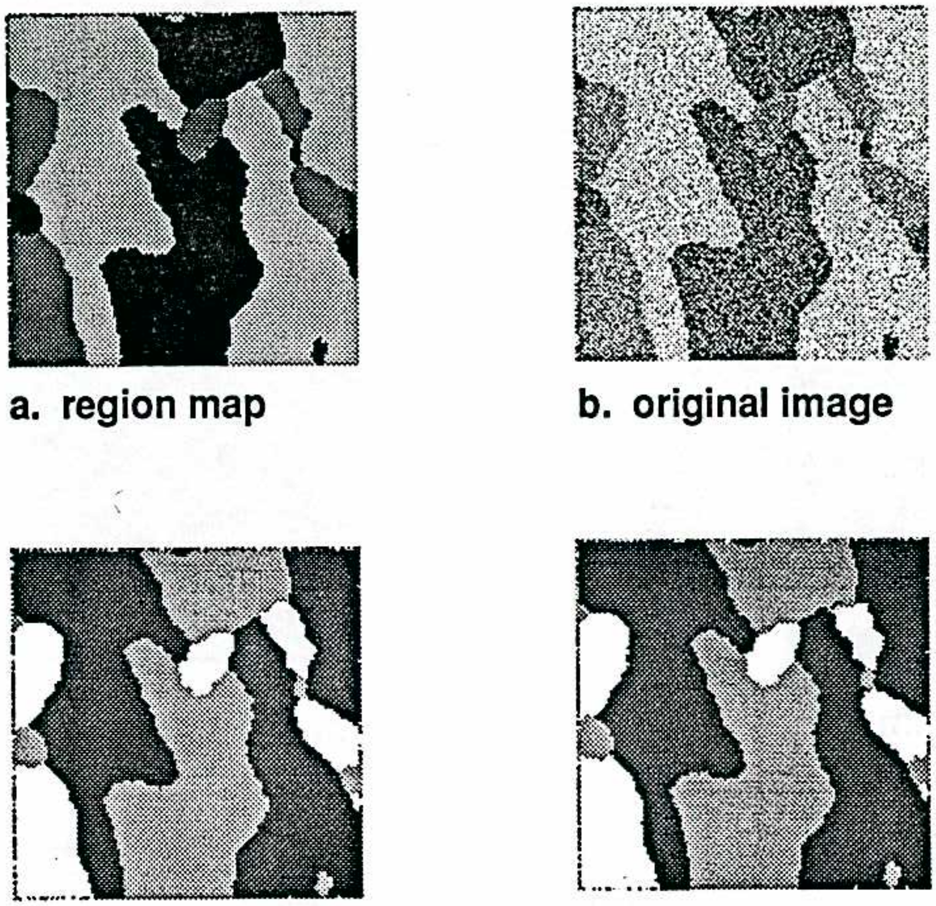
a. region map

b. original image

c. segmentation by
total ML approach

d. segmentation by
Bayesian approach

Figure 5. Segmentation of Synthetic Image 1
with Mean Field Theory



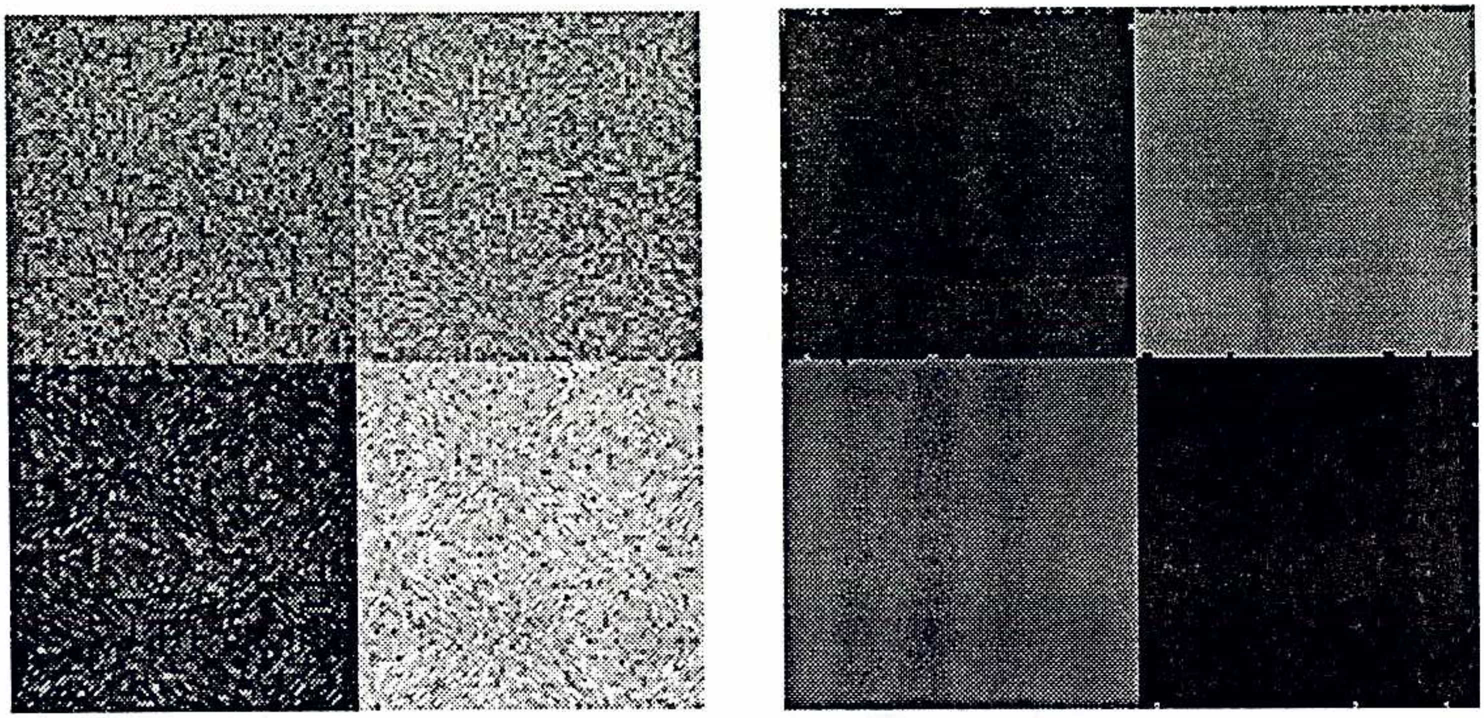
a. region map

b. original image

c. segmentation by total ML approach

d. segmentation by Bayesian approach

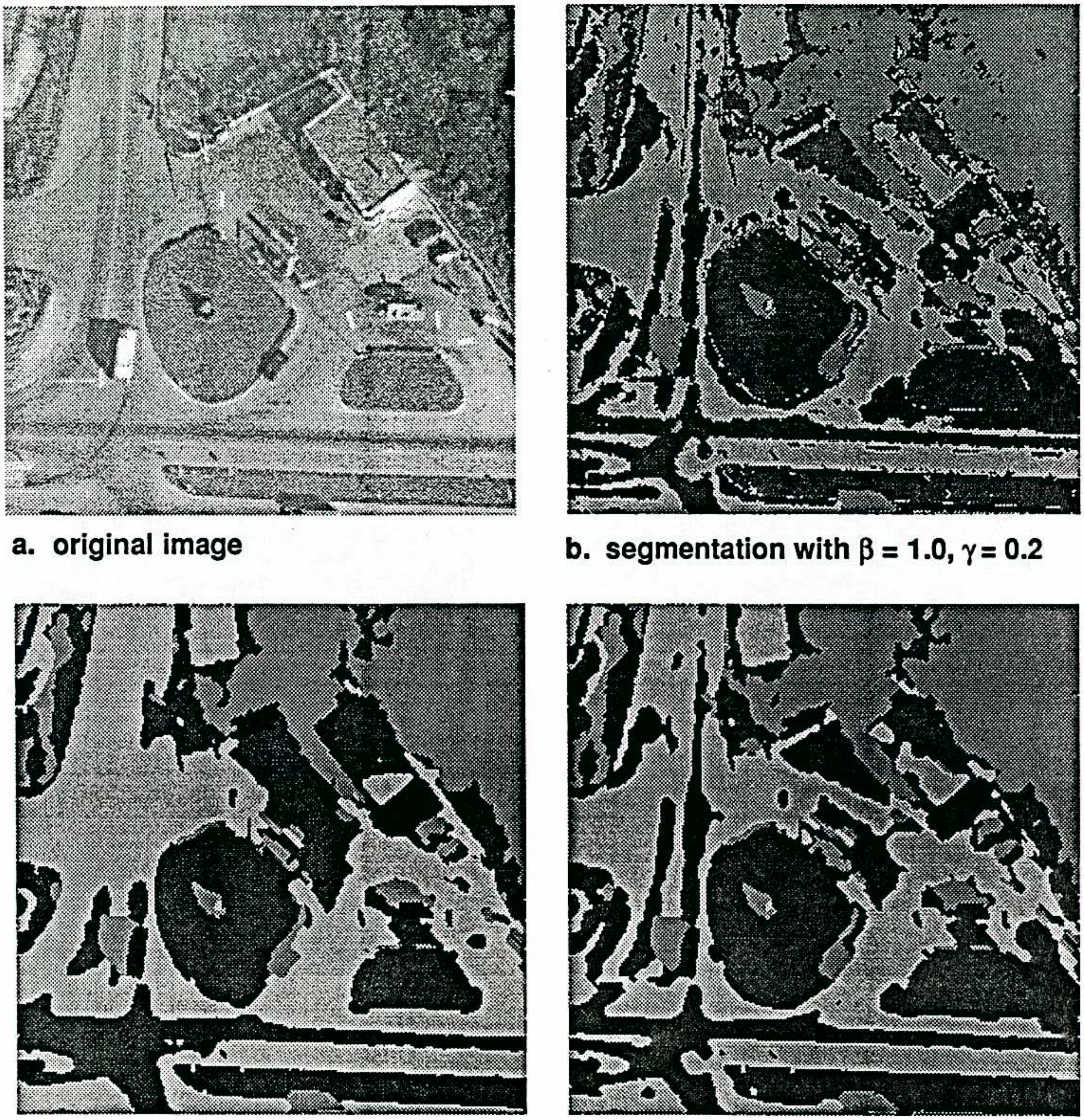
Figure 6. Segmentation of Synthetic Image 2 with Mean Field Theory



a. original image

b. segmentation with $\beta = 1.0, \gamma = 0.5$

Figure 7 Segmentation of Synthetic Image 3 with Mean Field Theory



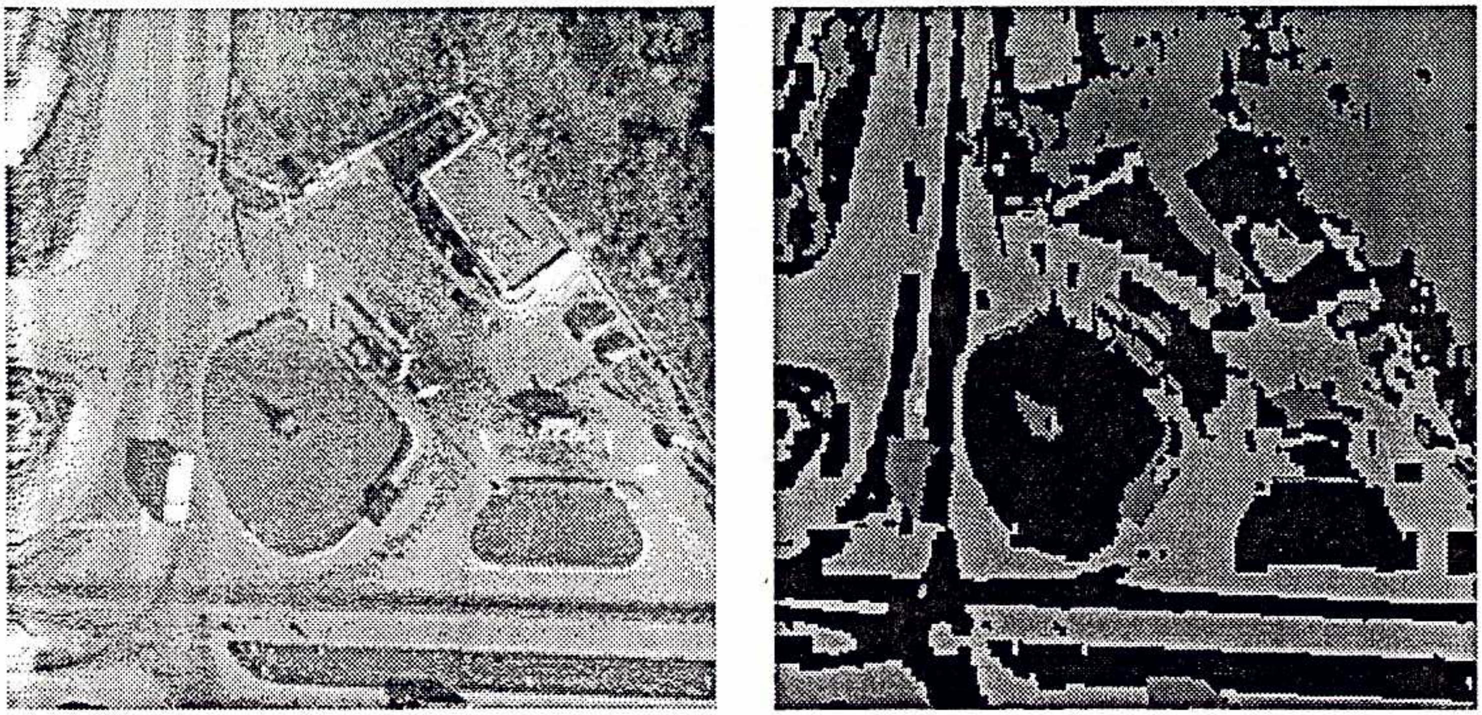
a. original image

b. segmentation with $\beta = 1.0, \gamma = 0.2$

c. segmentation with $\beta = 1.0$, ~~α~~ γ $= 0.5$

d. segmentation with $\beta = 1.12, \gamma = 0.2$

Figure 8 Segmentation of the Aerial Photograph with Mean Field Theory



a. original image

b. segmentation

Figure 9. Segmentation of the Aerial Photograph by Total ML Approach