Unsupervised Vector Image Segmentation by a Tree Structure—ICM Algorithm

Jong-Kae Fwu and Petar M. Djurić,* *Member, IEEE*

Abstract— **In recent years, many image segmentation approaches have been based on Markov random fields (MRF's). The main assumption of the MRF approaches is that the class parameters are known or can be obtained from training data. In this paper we propose a novel method that relaxes this assumption and allows for simultaneous parameter estimation and vector image segmentation. The method is based on a tree structure (TS) algorithm which is combined with Besag's iterated conditional modes (ICM) procedure. The TS algorithm provides a mechanism for choosing initial cluster centers needed for initialization of the ICM. Our method has been tested on various one-dimensional (1-D) and multidimensional medical images and shows excellent performance. In this paper we also address the problem of cluster validation. We propose a new maximum *a posteriori* (MAP) criterion for determination of the number of classes and compare its performance to other approaches by computer simulations.**

I. Introduction

VECTOR images are comprised of pixels that are represented by vectors. They are frequently encountered in medicine [18], [30], [32], and are inherent in color [40] and satellite imaging [8]. For example, the pixels of magnetic resonance images (MRI) are usually represented by the weighted spin-lattice relaxation time (T1), the spin-spin relaxation time (T2), and the proton density (Pd). Each image (T1, T2, or Pd) of the same slice provides information about the anatomy of the slice, and usually, the objective is to fuse the information optimally and obtain the best estimates of some important quantities. The same applies to other vector image problems, such as problems involving color and satellite images. Recall that color images are represented by three different perceptual attributes, namely, the brightness, the hue, and the saturation, and land satellite images have seven spectral measurements associated with each pixel.

An important problem in processing images is their segmentation, that is grouping the image pixels with homogeneous attributes together and assigning them an adequate label. The segmentation of vector images should yield better results than the segmentation of one-dimensional (1-D) images, but its implementation is a more a difficult task. The commonly used maximum likelihood and clustering methods for vector image segmentation tend to have an unacceptably large number of misclassified pixels since they ignore spatial dependencies. Most of the methods using nonlinear filters are based on order statistics [39], [40], but to date there is no unified theory on multivariate ordering. Other approaches typically treat each image separately and then combine the results by some user-specified rules [8]. There are also methods that project the multidimensional space onto a 1-D subspace, and proceed by applying techniques for 1-D image segmentation [21].

Markov random field (MRF) models are commonly used to model data encountered in many practical signal processing [27], image processing [15], [23], [29], [49], and computer vision [14], [33] problems. They can capture locally dependent characteristics of the images very well by tending to represent neighboring data with the same properties of signal attributes. Most of the MRF-based segmentation methods proposed recently are supervised since they rely on the assumption of known class parameters or availability of training data [4], [11], and [29]. Quite often, however, prior knowledge of class parameters may not exist, or training data are not available. Even when training data are available, intensive user interaction is needed to obtain the required class parameters. Consequently, unsupervised segmentation has received considerable attention by the image processing community [25], [31], [34], [46], [49]. Some other important references regarding MRF-based image segmentation with emphases on vector images include [26], [30], [36], and [42].

Since the underlying scenes are unobservable, we want to reconstruct them from the observed images. Geman and Geman modeled the underlying scene as an MRF [15], and they showed that the posterior distribution of the underlying image is also an MRF. They also proposed a Gibbs sampler and a simulated annealing (SA) technique to find its maximum *a posteriori* (MAP) estimate. Their approach can reach the global maximum, but it requires intensive and, typically, random amount of computation. To avoid the computational difficulties in the MAP estimation, Marroquin, Mitter, and Poggio derived the maximizer of the posterior marginals (MPM) technique, which minimizes a segmentation error [33]. Besag proposed the iterated conditional modes (ICM) algorithm which is computationally efficient but guarantees only that a local maximum solution can be reached [4]. Dubes and Jain compared the SA, ICM, and MPM and found that the ICM is the most robust of the three methods [11]. Pappas proposed a generalized K-means algorithm which is

Manuscript received April 28, 1995; revised August 11, 1996. This work was supported by the National Science Foundation under Award MIP-9506743. The Asssociate Editor responsible for coordinating the review of this paper and recommending its publication was Z.-P Liang. *Asterisk indicates corresponding author.*

J.-K. Fwu is with the Department of Electrical Engineering, State University of New York at Stony Brook, Stony Brook, NY 11794-2350 USA.

*P. M. Djurić is with the Department of Electrical Engineering, State University of New York at Stony Brook, Stony Brook, NY 11794-2350 USA (e-mail: djuric@sbee.sunysb.edu).

Publisher Item Identifier S 0278-0062(96)08707-1.

a segmentation-estimation approach that varies an estimation window size as the algorithm progresses [37]. However, it is not clear how the window sizes are changed systematically. Lakshmanan and Derin used an adaptive algorithm to estimate the edge penalty of the MRF model and to segment the images alternately [29] and [46]. Finally, Liang [32] and Zhang [49] exploited the EM algorithm [9] to estimate the parameters from incomplete data.

All these algorithms are iterative and therefore require initialization of their parameters [25]. In processing multidimensional images, it is more difficult to identify initial cluster centers and determine decision boundaries. One popular algorithm for initialization of class parameters is the K-means algorithm, but its performance strongly depends on the number of classes and the selected initial conditions [19] and [45]. If the initial conditions are chosen inappropriately, the initial classification is poor, which subsequently degrades the final segmentation. The same happens to the mixture method which is a sophisticated version of the K-means algorithm [19]. Instead of associating pixels with existing classes, it assigns only probabilities of association with these classes. On the other hand, approaches for the initialization of class parameters based on conventional histogram thresholds cannot be extended to multidimensional images in a straightforward manner [41] and [47]. Multiscale MRF models have been introduced to improve the segmentation [20]. However, initial conditions or predefined grouping techniques for initializing the coarse segmentation are still needed [6] and [7]. Banfield and Raftery applied a multidimensional model-based clustering algorithm to MRI brain images [3], but they ignored spatial information.

In this paper, we propose a novel approach for simultaneous class parameter estimation and image segmentation, which we refer to as tree structure-ICM, TS–ICM. The algorithm combines the tree structure vector quantization (TSVQ) procedures [38] and [43], with the ICM method. Vector quantization (VQ) is a powerful and popular technique in coding theory that is utilized for data compression [16]. A full search VQ can achieve optimal quantization, but is a very time consuming task. As an alternative, there is a suboptimal, low complexity VQ scheme known as TSVQ. Instead of a full search VQ, the TSVQ is composed of a sequence of binary searches which greatly reduces the computational load, and provides suboptimal, and yet good results [38]. As a TSVQ scheme, one can adopt a greedy tree structure, also used in this paper, which splits the tree one node at a time, and subsequently selects the node with the smallest overall cost. This usually gives an unbalanced tree because the node that is split can be at any depth [43]. The TSVQ is used to choose initial cluster centers, and when combined with the ICM, to provide a mechanism for unsupervised image segmentation and class parameter estimation. It should be noted that it is well known that classification can also be viewed as a form of compression and vice versa [35]. More information on VQ and TSVQ can be found in [16] or in the tutorial article [17].

An automatic image segmentation technique has to perform segmentation without operator's assistance. A very difficult problem in automatic segmentation is the determination of the number of classes from the observed data, also known as cluster validation or multiple hypotheses testing [22] and [48]. Recently, qualitatively new methods have been proposed based on penalized maximum likelihood criteria. Among them, the most popular are the Akaike's information criterion (AIC) [1] and the Minimum description length (MDL) [6], [32], [44], and [48]. Many researchers have applied them and recognized their poor performance in many cases [28]. Modified rules have been proposed [32], however, most of them are heuristic and, in general, do not perform well. It is well known that the cluster validation is a very difficult problem and that a search for a solution is still under way [13], [28], [46]. This is more so for vector images.

In image processing, the cluster validation problem is somewhat easier because the pixels are ordered data. The ordering provides inherent spatial information that is exploited for cluster validation by imposing spatial constraints. In this paper we derive a maximum *a posteriori* (MAP) solution for cluster validation. We utilize asymptotic Bayesian theory, which has been exploited before in model selection problems [10] and [24]. The number of classes is determined along with the segmentation.

The paper is organized as follows: The problem of vector image segmentation is stated in Section II. The details of our TS–ICM algorithm are provided in Section III, where we describe the MAP criterion for choosing the best nodes in the tree structure algorithm as well as the selection of perturbation vectors for splitting the nodes. In Section IV, we propose a novel Bayesian cluster validation criterion for determining the number of classes. This criterion, when combined with our TS algorithm, provides an automatic image segmentation scheme. Some simulation results on various 1-D and multidimensional medical images are shown in Section V. Finally, a discussion on two important issues about our TS algorithm is presented in Section VI, and a brief conclusion is given in Section VII.

II. Problem Statement

Let $\mathcal{S} = \{s = (i, j) | 1 \leq i < M_1, 1 \leq j < M_2\}$ denote a two-dimensional (2-D) $M_1 \times M_2$ lattice with a neighborhood system $\mathcal{N} = \{N_s : s \in \mathcal{S}\}$, where $N_s \subset \mathcal{S}$ is a set that contains the neighboring sites of s. Let $\mathbf{Y}$ be a p-dimensional (p–D) random field defined on $\mathcal{S}$, that is $\mathbf{Y} = \{\mathbf{y}_s | s \in \mathcal{S}\}$. Y is the observation of $\mathbf{Y}$ and $\underline{y}_s = [y_s^{[1]}, y_s^{[2]}, \cdots, y_s^{[p]}]^T$ is the observed random vector $\mathbf{y}_s$. Let $\mathbf{X} = \{\mathbf{x}_s | s \in \mathcal{S}\}$ be a 1-D random field defined on $\mathcal{S}$, and let X be the realization of $\mathbf{X}$ and x_s the realization of $\mathbf{x}_s$. We assume that $\mathbf{X}$ is an MRF with a probability distribution $f(X)$, and that $\mathbf{X}$ is the set of class labels of the underlying image of $\mathbf{Y}$. $\mathbf{X}$ is comprised of pixels that belong to one of m classes. It is assumed that m and the parameters associated with each class are unknown. Let c be a clique that is a subset of $\mathcal{S}$ with one or more sites such that each site in c is a neighbor of all the remaining sites in c, and let $\mathcal{C}$ denote the set of all cliques.

The joint probability of X is a Gibbs distribution whose form is

$$f(X) = \frac{1}{Z} \exp\left[U(X)\right] \tag{1}$$

$$Z = \sum_{X} \exp[U(X)] \tag{2}$$

where Z is a normalizing constant called the partition function, and $U(X)$ is an energy function defined by

$$U(X) = \sum_{c \in \mathcal{C}} V_c(X) \tag{3}$$

with $V_c(X)$ being the potential function whose argument, X, is an element of the clique. The MRF considered here is a multilevel logic model [46] that has a second order neighborhood system (eight neighbors) with pairwise cliques [29] whose potential function is defined as

$$V_c(x_s) = \begin{cases} \beta & \text{if } x_s = x_{s'}, \quad \forall s, s' \in c \text{ and } s \neq s' \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where β is the hyperparameter of the model, which can also be interpreted as edge penalty.

The observed image $\mathbf{y}_s$ is obtained when the noise $\mathbf{w}_s$ is superimposed on the signal $g(\mathbf{x}_s)$, that is

$$\mathbf{y}_s = g(\mathbf{x}_s) + \mathbf{w}_s \tag{5}$$

where $g(\mathbf{x}_s)$ is a function that maps the underlying label $\mathbf{x}_s$ to its associated attribute vector $\underline{\mu}_{\mathbf{x}_s}$. The $\mathbf{w}_s$'s are independently distributed Gaussian random vectors with zero mean and unknown covariance matrix $\sum_{\mathbf{x}_s}$, which is class conditional. Therefore, the density of $\mathbf{Y}$, given the underlying true image $\mathbf{X} = X$, is

$$f(Y|X) = \prod_{s \in \mathcal{S}} f(\underline{y}_s|x_s) \tag{6}$$

and

$$f(\underline{y}_s|x_s) = \frac{1}{(2\pi)^{p/2}\left|\sum_{x_s}\right|^{1/2}} \cdot \exp\left[-\frac{1}{2}(\underline{y}_s - \underline{\mu}_{x_s})^T \sum_{x_s}^{-1} (\underline{y}_s - \underline{\mu}_{x_s})\right] \quad x_s \in \{1, 2, \cdots, m\} \tag{7}$$

where $\underline{y}_s$ and $\underline{\mu}_{x_s}$ are p-D vectors.

Based on the observed vector image Y, the problem is to find the number of classes m and classify the observed random vector $\underline{y}_s$ into one of the m different classes. Note that, according to the assumptions, each pixel has a Gaussian distribution whose parameters $\underline{\mu}_{x_s}$ and $\sum_{x_s}$ are not known. In addition, the edge penalty β that appears in the MRF model is also unknown.

III. The TS–ICM Algorithm

First, we address the segmentation problem when the number of classes is known. Since we plan to apply the ICM method, it is critical to resolve the problem of its initialization [6] and [25]. Our goal is to find both X and Θ which maximize the *a posteriori* (MAP) function, i.e.,

$$(\hat{X}_m, \hat{\Theta}_m) = \arg \max_{X, \Theta} f(X, \Theta|Y, m) \tag{8}$$

where $\hat{\Theta}$ includes both the class parameters $(\underline{\mu}_{x_s}, \sum_{x_s})$, $x_s \in \{1, 2, \cdots, m\}$ and the edge penalty β of the MRF and m is the assumed number of different classes. Since the maximization of (8) with respect to both X and Θ is a formidable task, the TS–ICM algorithm uses a partial optimal solution [29], [46] for $\hat{X}$ and $\hat{\Theta}$. In other words, we apply a sequential scheme based on

$$\hat{X}_k = \arg \max_{X} f(X|\hat{\Theta}_k, Y, k) \tag{9}$$

$$\hat{\Theta}_k = \arg \max_{\Theta} f(\Theta|\hat{X}_k, Y, k) \tag{10}$$

where the index $k = 1, 2, \cdots, m$, denotes both the number of classes and the current stage of the algorithm. At each stage we try to obtain intermediate solutions iteratively by using (9) and (10) that ultimately leads to the solution defined by (8). Once the partial optimal solutions are found, they are used as initial values for the next stage. Note however, the following: 1) (9) and (10) are only suboptimal solutions of $\hat{X}_k$ and $\hat{\Theta}_k$, 2) to implement (9), we need the estimates of the unknown parameters $\hat{\Theta}_k$, and 3) to apply (10), we need the segmented underlying image $\hat{X}_k$.

Suppose that we have the initial estimates of $\hat{X}_k^{(0)}$ and $\hat{\Theta}_k^{(0)}$ obtained by the TS scheme. We then apply the ICM using $\hat{\Theta}_k^{(0)}$ to get $\hat{X}_k^{(1)}$. Once we know $\hat{X}_k^{(1)}$, we estimate the class parameters $\hat{\Theta}_k^{(1)}$. One of these parameters is the edge penalty $\hat{\beta}_k^{(1)}$, which is estimated by maximizing the likelihood function $f(\hat{X}_k^{(1)})$ of the estimated underlying image $\hat{X}_k^{(1)}$. The maximum likelihood (ML) estimate of β can be written as

$$\hat{\beta}_k = \arg \max_{\beta} \{f(\hat{X}_k)\} \tag{11}$$

$$= \arg \max_{\beta} \left\{\frac{1}{Z} \exp[U(\hat{X}_k)]\right\}. \tag{12}$$

To avoid the intractable partition function Z in $f(X)$ and estimate the edge penalty β that appears in the MRF model, we use the "pseudo-likelihood" [4] and [5]. It is defined by

$$PL(X) = \prod_{s \in \mathcal{S}} f(x_s|N_s) \tag{13}$$

$$= \prod_{s \in \mathcal{S}} \frac{1}{z_s} \exp\left\{\sum_{c \in \mathcal{C}} V_c(x_s|N_s)\right\} \tag{14}$$

$$= \prod_{s \in \mathcal{S}} \frac{\exp\left\{\sum_{c \in \mathcal{C}} V_c(x_s|N_s)\right\}}{\sum_{x_{s'}} \exp\left\{\sum_{c \in \mathcal{C}} V_c(x_{s'}|N_{s'})\right\}} \tag{15}$$

where the z_s's are normalizing constants. The estimate of the edge penalty is then given by

$$\hat{\beta}_k = \arg \max_{\beta} \{PL(\hat{X}_k)\} \tag{16}$$

$$= \arg \max_{\beta} \left\{\prod_{s \in \mathcal{S}} \frac{1}{z_s} \exp \sum_{c \in \mathcal{C}} V_c(\hat{x}_s|N_s)\right\} \tag{17}$$

$$= \arg \min_{\beta} \left\{\sum_{s \in \mathcal{S}} \left[-\sum_{c \in \mathcal{C}} V_c(\hat{x}_s|N_s) + \ln\left[\sum_{\hat{x}_s} \exp\left\{\sum_{c \in \mathcal{C}} V_c(\hat{x}_s|N_s)\right\}\right]\right]\right\} \tag{18}$$

where (18) is obtained by taking the negative logarithm of (17).

Once we have $\hat{\Theta}_k^{(1)}$, we continue to estimate sequentially $\hat{X}_k^{(n)}$ and $\hat{\Theta}_k^{(n)}$, where $n = 2, 3, \cdots$, until convergence is reached.

According to the TS algorithm, we split each cluster into two clusters one at a time, where each cluster center is the estimated mean vector and denoted as a node in the tree. If there already exist $k-1$ classes, we have $k-1$ different nodes to split. The best node $\hat{\kappa}$ is the one obtained from

$$\hat{\kappa} = \arg \max_{\kappa \in \{1, 2, \cdots, k-1\}} f(X_k^{\kappa}|Y) \tag{19}$$

where κ denotes the node being split, X_k^{κ} the MRF field with k classes obtained by splitting the κth node, and $\hat{\kappa}$ is the best node to split.

From Bayes' theorem

$$f(X_k|Y) = \frac{f(Y|X_k)f(X_k)}{f(Y)} \tag{20}$$

it follows that the maximization of $f(X_k|Y)$ does not depend on $f(Y)$, and therefore $f(Y)$ can be ignored. To avoid difficulties in dealing with all the possible configurations of X as well as the intractable partition function Z, which appears in $f(X_k)$, we adopt the pseudo-likelihood again. The MAP criterion then becomes

$$\hat{\kappa} = \arg \max_{\kappa \in \{1, 2, \cdots, k-1\}} f(Y|X_k^{\kappa})f(X_k^{\kappa}) \tag{21}$$

$$= \arg \max_{\kappa \in \{1, 2, \cdots, k-1\}} \left\{ \prod_{s \in \mathcal{S}} f(\underline{y}_s|x_s) \prod_{s \in \mathcal{S}} \frac{1}{z_s} \cdot \exp \left[\sum_{c \in \mathcal{C}} V_c(x_s|N_s) \right] \right\} \tag{22}$$

$$= \arg \min_{\kappa \in \{1, 2, \cdots, k-1\}} \left\{ \sum_{s \in \mathcal{S}} \left[\tfrac{1}{2} \ln \left| \sum_{x_s} \right| + \tfrac{1}{2} (\underline{y}_s - \underline{\mu}_{x_s})^T \sum\nolimits_{x_s}^{-1} (\underline{y}_s - \underline{\mu}_{x_s}) \right] \right. \tag{23}$$

$$\left. + \sum_{s \in \mathcal{S}} \left[- \sum_{c \in \mathcal{C}} V_c(x_s|N_s) + \ln \left[\sum_{x_s} \exp \left(\sum_{c \in \mathcal{C}} V_c(x_s|N_s) \right) \right] \right] \right\} \tag{24}$$

$$= \arg \min_{\kappa \in \{1, 2, \cdots, k-1\}} \{F_d(\underline{y}_s, \hat{\Theta}_k) + F_c(x_s, \hat{\Theta}_k)\} \tag{25}$$

where

$$F_d(\cdot) = \sum_{s \in \mathcal{S}} \left\{ \tfrac{1}{2} \ln \left| \sum_{x_s} \right| + \tfrac{1}{2} (\underline{y}_s - \underline{\mu}_{x_s})^T \sum\nolimits_{x_s}^{-1} (\underline{y}_s - \underline{\mu}_{x_s}) \right\} \tag{26}$$

$$F_c(\cdot) = \sum_{s \in \mathcal{S}} \left\{ - \sum_{c \in \mathcal{C}} V_c(x_s|N_s) + \ln \left[\sum_{x_s} \exp \left\{ \sum_{c \in \mathcal{C}} V_c(x_s|N_s) \right\} \right] \right\} \tag{27}$$

Therefore, the surviving node which maximizes the MAP criterion is obtained from a function that is a penalized likelihood comprised of two terms. One is the data term, $F_d(\cdot)$, which is a measure of the fitting error, and the other, a penalty term $F_c(\cdot)$, which penalizes for spatial discontinuities.

Before we outline the algorithm, two important issues are discussed. a) We add a small value of perturbation $\underline{\epsilon}$ to each cluster center in opposite directions to stimulate breaks of clusters. In the conventional TSVQ scheme, the perturbation $\underline{\epsilon}$ can be any small arbitrary fixed vector, and it is usually not in the direction of the cluster orientation. To avoid unwanted effects caused by using arbitrary $\underline{\epsilon}$ and to improve the performance, $\underline{\epsilon}$ is chosen in the direction of the largest variability of the cluster [3] and [12]. This direction is given by the eigenvector $\underline{v}_{1x_s}$ associated with the largest eigenvalue of the cluster covariance matrix $\hat{\Sigma}_{x_s}$. b) The initial conditions provided by the previous cluster splittings lead to a local minimum and many misclassified pixels. The ICM algorithm, however, will correct the classification of most of them by exploiting their spatial interdependence, thus precluding propagation of the misclassified pixels to the next stages. Since a big portion of the initially misclassified pixels is now correctly classified, this not only improves the segmentation results, but also the partial optimal parameter estimates. We emphasize that the algorithm does not require *a priori* information, and therefore is completely data-driven.

The algorithm is implemented along the steps outlined below. Note that Steps 3) and 4) are identical to the ISODATA algorithm [2], and Steps 5) and 6) are from [46].

Step 1) Choose the initial number of classes equal to one ($k = 1$), i.e., $x_s = 1$, $\forall s \in \mathcal{S}$. Estimate the cluster center $\hat{\underline{\mu}}_1$, which is simply the sample mean vector of all the pixels, and $\hat{\Sigma}_1$, where $\hat{\Sigma}_1 = 1/(M_1 \times M_2) \sum_{s \in \mathcal{S}} (\underline{y}_s - \underline{\mu}_1)(\underline{y}_s - \underline{\mu}_1)^T$.

Step 2) Increase k by one, ($k = 2$). Split $\hat{\underline{\mu}}_1$ to form two initial cluster centers by using $\hat{\underline{\mu}}_1 + \underline{\epsilon}$ and $\hat{\underline{\mu}}_1 - \underline{\epsilon}$. The perturbation $\underline{\epsilon}$ is chosen in the direction of the eigenvector associated with the largest eigenvalue of $\hat{\Sigma}_1$.

Step 3) Classify all the pixels to one of the two classes by

$$x_s = l \quad \text{if} \quad d(\underline{y}_s, \hat{\underline{\mu}}_l) \leq d(\underline{y}_s, \hat{\underline{\mu}}_{l'}), \qquad l \neq l' \tag{28}$$

where we choose $d(\cdot)$ to be the Euclidean distance in the p-D space. Then update $\hat{\underline{\mu}}_1$, $\hat{\underline{\mu}}_2$, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ by

$$\hat{\underline{\mu}}_l = \frac{1}{r_l} \sum_{s \in \mathcal{S}} \underline{y}_s, \qquad x_s = l \tag{29}$$

$$\hat{\Sigma}_l = \frac{1}{r_l} \sum_{s \in \mathcal{S}} (\underline{y}_s - \underline{\mu}_l)(\underline{y}_s - \underline{\mu}_l)^T, \quad x_s = l \tag{30}$$

where $l = 1, 2$, and r_l is the number of pixels in class l.

Step 4) Repeat Step 3) until convergence is reached.

Step 5) Use $\hat{\underline{\mu}}_1$, $\hat{\underline{\mu}}_2$, $\hat{\Sigma}_1$, and $\hat{\Sigma}_2$ as well as the ICM to segment the image, i.e., find x_s. Update $\hat{\underline{\mu}}_1$, $\hat{\underline{\mu}}_2$, $\hat{\Sigma}_1$,

and $\hat{\Sigma}_2$ by (29) and (30). Update $\hat{\beta}_2$ by maximizing the pseudo-likelihood (18).

Step 6) Repeat Step 5) until convergence is reached.

Step 7) Increase the number of classes by one. In a similar fashion as in Steps 2)–6), split the existing nodes one at a time to form additional cluster centers. Choose the best set of cluster centers from the set of $k-1$ candidates for the next stage according to the MAP criterion (25).

Step 8) Repeat Step 7), forming one node at each stage until $k = m$.

IV. Cluster Validation

From Bayes' theorem, the joint *a posteriori* estimate of X_m and m is given by

$$f(X_m, m|Y) = \frac{f(Y|X_m, m)f(X_m|m)f(m)}{f(Y)} \tag{31}$$

where $f(Y|X_m, m)$ is the likelihood function, $f(X_m|m)$ is the Gibbs distribution of the underlying image with m classes, and $f(m)$ is the *a priori* probability of the model with m classes. If we assume uniform $f(m)$, the MAP solution of (31) becomes

$$(\hat{X}_m, \hat{m})_{\text{MAP}} = \arg \max_{(X,m)} \{f(Y|X_m, m)f(X_m|m)\}. \tag{32}$$

Given the number of classes m, we can find the underlying image X_m by the TS–ICM technique.

Once we obtain $\hat{X}_m$ for $m = 1, 2, \cdots, m_{\max}$, the number of classes is selected according to

$$\hat{m}_{\text{MAP}} = \arg \max_{m} \{f(Y|\hat{X}_m)f(\hat{X}_m)\}. \tag{33}$$

Note that $f(Y|\hat{X}_m) = \int_{\Theta_m} f(Y|\hat{X}_m, \Theta_m)f(\Theta_m)d\Theta_m$, where Θ_m is the parameter vector of all the classes. If we Taylor expand $f(Y|\hat{X}_m, \Theta_m)$ around the ML estimate $\hat{\Theta}_m$ and use asymptotic approximation, we find that (33) simplifies to

$$\hat{m}_{\text{MAP}} = \arg \min_{m} \left\{ -\ln f(Y|\hat{X}_m, \hat{\Theta}_m) + \sum_{i=1}^{m} p \ln n_i - \ln PL(\hat{X}_m) \right\} \tag{34}$$

where p is the dimension of the vector images, $\hat{\Theta}_m$ is the ML estimate of Θ_m, n_i is the number of pixels that belong to the ith class, and $PL(\hat{X}_m)$ is the pseudo-likelihood. It is a penalized maximum likelihood criterion with a simple interpretation. The first term is a data term which corresponds to the fitting error of the applied model. The second term is a penalty for overparameterization due to unnecessary classes. The third term is also a penalty, which penalizes for additional spatial discontinuities as quantified by the MRF model. For the purpose of comparison, we list the following criteria for cluster validation that have been reported in the literature [28], [32], [46], and [48]

$$\hat{m}_{\text{AIC}} = \arg \min_{m} \{-\ln f(Y|\hat{X}_m) + d_f\} \tag{35}$$

$$\hat{m}_{\text{MDL}} = \arg \min_{m} \{-\ln f(Y|\hat{X}_m) + \tfrac{1}{2} d_f \ln N\} \tag{36}$$

TABLE I
The Initial and Final Estimated Parameters of ICM Initialized by the TS–ICM and the K-Means Algorithm

	True Parameters	20.0, 80.0, 100.0, 120.0, 160.0
TS-ICM	Initial parameters	20.06, 70.79, 96.66, 122.18, 158.78
TS-ICM	Final estimates	20.00, 79.90, 99.58, 118.97, 158.90
K-means	Initial parameters	20.15, 80.03, 109.99, 137.40, 172.02
K-means	Final estimates	20.00, 88.59, 107.32, 130.65, 161.07

$$\hat{m}_{\text{CL}} = \arg \min_{m} \{-\ln f(Y|\hat{X}_m) - \ln PL(\hat{X}_m) + N^c d_f\}. \tag{37}$$

where d_f is the number of free parameters in the model, c is a prespecified constant, and N is the data size. Note that all the criteria have identical data terms; they differ in their penalty functions only. The AIC's penalty is not a function of the data lengths, while the MDL's, unlike the MAP's, depends only on the total size of the image. Also the AIC and MDL do not have terms that arise from spatial information. We must point out, however, that an MDL segmentation with spatial penalties has been proposed in [26]. The compensated likelihood (CL) [46] includes the spatial information via the MRF as in our rule, but it has a third term, different from ours, which has been determined empirically by experimentation.

V. Simulation Results

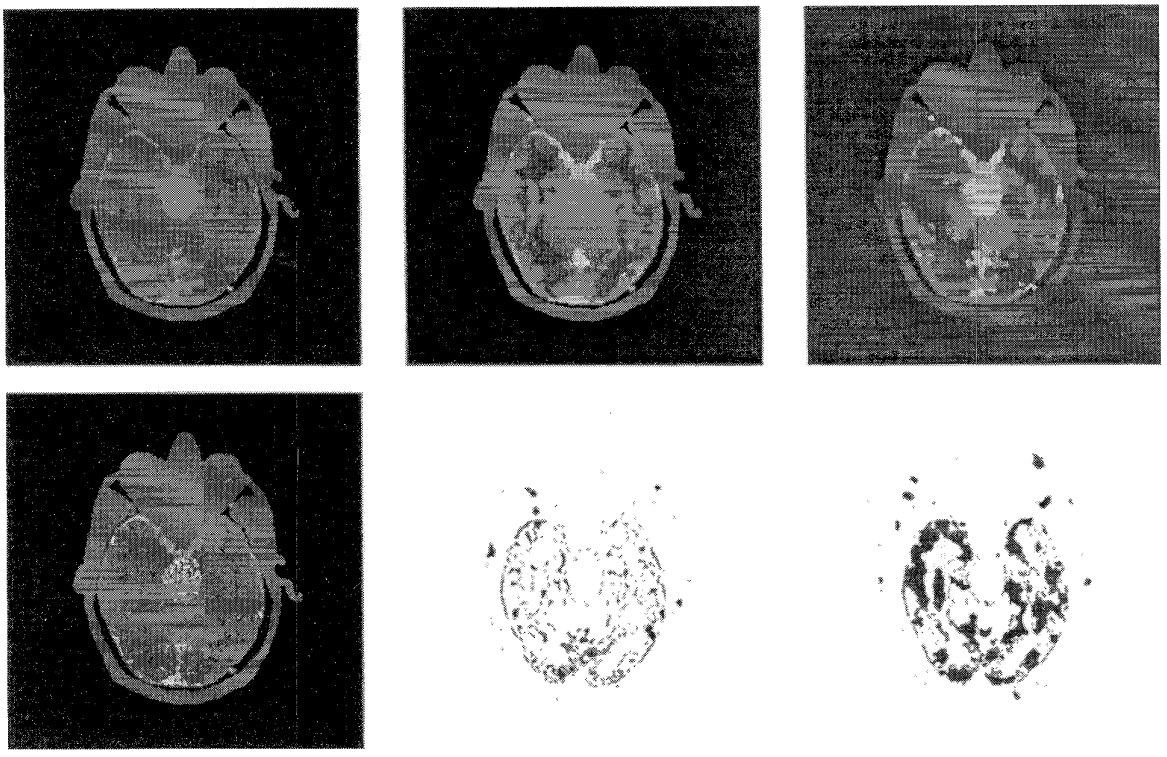
A. TS–ICM

In this section, we present the results of three experiments. In the first, we applied the TS–ICM to 1-D ($p = 1$) synthesized MR brain images, where the shapes of the various tissues were obtained from a hand segmented MR image. We compare its performance to that of the K-means algorithm. The size of the image was 256×256. The contrast-to-noise ratio (CNR) defined by

$$\text{CNR} = \min_{l,k} \left\{ \frac{|\mu_l - \mu_k|}{\sigma_l}, \frac{|\mu_l - \mu_k|}{\sigma_k} \right\} \quad l \neq k \text{ and } l, k \in \{1, 2, \cdots, m\} \tag{38}$$

was equal to 20/20. In (38) μ_l and μ_k are the intensity levels of the pixels in two *adjacent* regions respectively, and σ_l, σ_k are noise standard deviations. There were five classes of objects (tissues) with true mean values $\mu_1 = 20$, $\mu_2 = 80$, $\mu_3 = 100$, $\mu_4 = 120$, and $\mu_5 = 160$. The noise deviations in the background and bone was 3.5 and for the remaining region, it was 20, that is $\sigma_1 = 3.5$, $\sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 20$.

In Fig. 1, the left two images are a noiseless (top) and noisy (bottom) MR images, and the middle two images are the segmented image by the TS–ICM and the corresponding error-map. The two images on the right are the segmented image obtained by the ICM initialized by the K-means algorithm and its error-map. When we simulated the K-means algorithm, we updated the parameters $\{\mu_l, \sigma_l\}$ after each iteration [46]. The results clearly show improved performance of our procedure over the K-means procedure.



(a) (b) (c)

Fig. 1. (a) A noiseless (top) and noisy (bottom) MR images. (b) The segmented labels by the TS–ICM and the corresponding error-map. (c) The segmented labels obtained by the ICM initialized by the K-means algorithm and its error-map.

TABLE II
THE ROBUSTNESS AGAINST NOISE WITH VARIOUS CNR's. THE TABLE INCLUDES THE ICM INITIALIZED BY THE TRUE PARAMETERS, THE TS–ICM, THE MODIFIED TS–ICM (ALLOWS BOTH CLUSTER SPLITTING AND MERGING), AND THE K-MEANS ALGORITHM

CNR	20/10	20/15	20/20	20/25
ICM (True parameters)	99.2 %	96.4 %	91.2%	74.6%
TS-ICM	99.2 %	96.2 %	91.2 %	72.4%
Modified TS-ICM	99.2 %	96.2 %	91.2 %	72.6%
K-means algorithm	99.2 %	95.8 %	79.5 %	67.0%

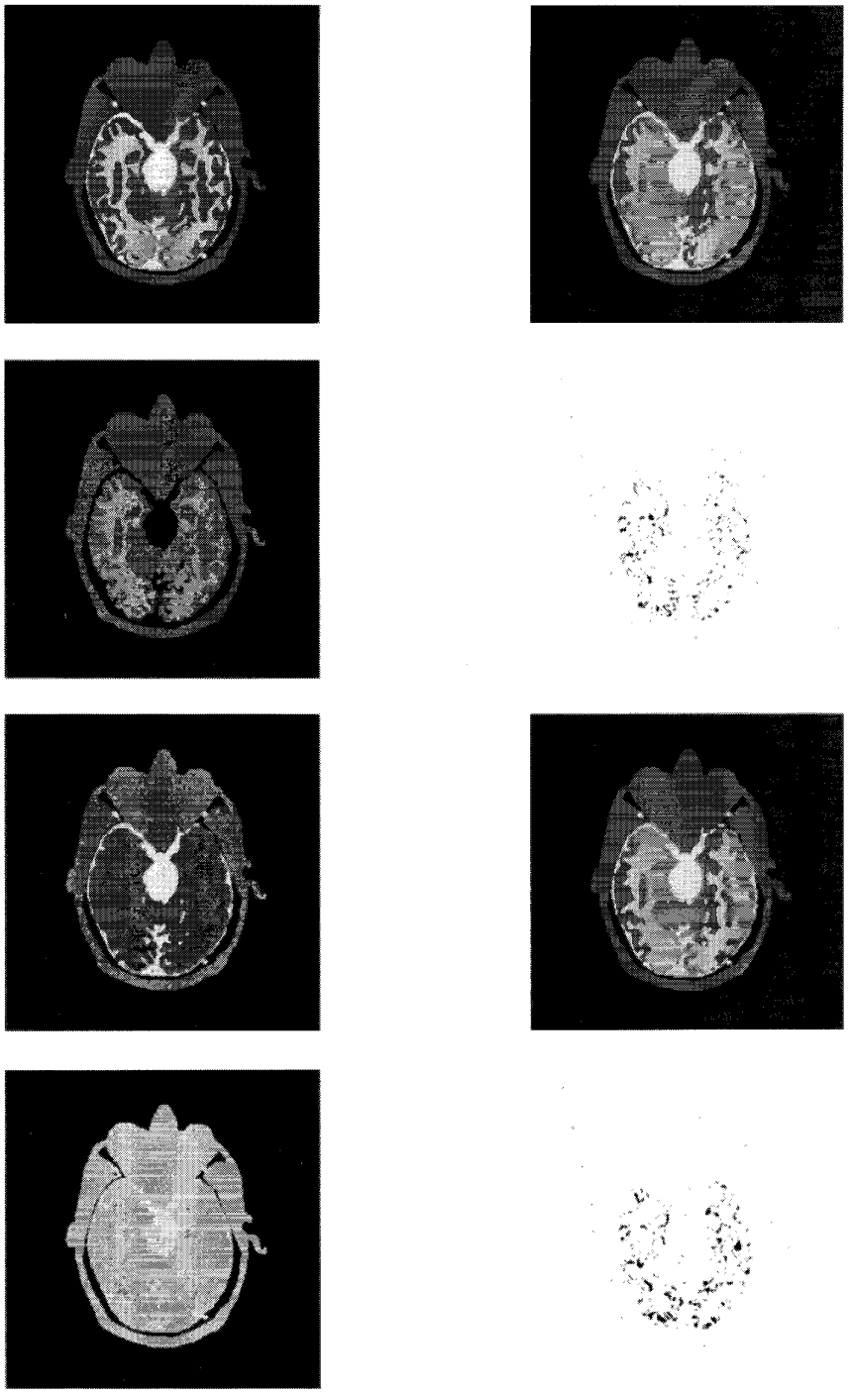
This improvement is due to the complete ICM segmentation after every split, which improves the initial parameters significantly. To compare these schemes, in Table I we show their initial and final estimated parameters obtained in the previous experiment. Our algorithm outperforms the K-means approach, and therefore yields better segmentation results.

To test the robustness against noise, we performed an experiment with various CNR's on the first image of Fig. 1. The simulation results that show the Percentage of Correct Classifications for the different CNR's is shown in Table II. The table includes the results of the ICM initialized by the true parameters, the TS–ICM, the modified TS–ICM (allows both cluster splitting and merging), and the K-means algorithm. When the CNR is high, all the approaches perform similarly. As the CNR decreases, the TS–ICM outperforms the K-means algorithm significantly. Although, the modified TS–ICM has larger computational requirement than the TS–ICM, their performances are comparable.

TABLE III
PARAMETERS OF THE SYNTHESIZED IMAGE IN EXPERIMENT 2

	White Matter	Gray Matter	CSF	Air/Bone	Other
T1 (μ)	168	126	8.0	5	117
T1 (σ)	21	25	2.7	1.5	17
T2 (μ)	106	119	247	5	140
T2 (σ)	15	18	3.0	1.5	18
Pd (μ)	182	200	230	5	200
Pd (σ)	18	19	11	1.5	19

In the next experiment, we applied our algorithm to a three-dimensional (3-D) synthesized MR image ($p = 3$) whose size was also 256×256. The parameters of the different tissues are shown in Table III, and the noiseless and three noisy images (T1, T2, Pd) are displayed on the left side of Fig. 2. The parameters of the different tissues are chosen according to [30] except that we added more noise. Note that there are only four different intensity levels shown in the third image (Pd), since two classes in the third image



(a) (b)

Fig. 2. (a) The class labels of the underlying image, and the three bands of the noisy synthesized MR images. (b) The top two images display the segmented labels and the error-map of the TS–ICM. The bottom two images are the segmented labels and the error-map of the ICM initialized by the K-means algorithm.

have the same intensity level. In addition, the white matter and gray matter are two adjacent regions with small CNR, which makes them difficult for discrimination.

In Fig. 2, the top two images on the right side display the segmentation result and the error-map of the TS–ICM. The bottom two images on the right display the segmentation result and the error-map of the ICM initialized by the K-means algorithm. Table IV shows the numbers of misclassified pixels in various regions obtained by the ICM initialized by the TS and the K-means algorithm, respectively. Clearly, the TS algorithm had better performance, and it is more so as we keep increasing the noise. In our experiments the algorithm created a few 1-pixel regions, which can be prevented by additional penalization.

TABLE IV
THE NUMBER OF MISCLASSIFIED PIXELS IN VARIOUS REGIONS FOR ICM INITIALIZED BY THE TS–ICM AND THE K-MEANS ALGORITHM

	White Matter	Gray Matter	CSF	Air/Bone	Other
TS-ICM	181	311	3	0	11
ICM initialized by the K-means algorithm	248	507	3	0	8

TABLE V
COMPARISON OF THE MAP, AIC, AND MDL RULES FOR CLUSTER VALIDATION. THE TRUE NUMBER OF CLASSED IS FIVE

rule\segm.	2	3	4	5	6	7
MAP	0	0	0	49	1	0
AIC	0	0	0	0	7	43
MDL	0	0	0	0	14	36

TABLE VI
SENSITIVITY ANALYSIS OF THE TS–ICM ALGORITHM

Image Size/CNR	20/2	20/4	20/6	20/8	20/10	20/12
32 x 32	0	0	0	1.1	4.0	F
64 x 64	0	0	0.5	15	F	F
128x128	0	F	F	F	F	F

In the last experiment, we applied our algorithm to a real 3-D MR image (T1, T2, and Pd) acquired by a 1.5-Tesla GE scanner, whose size was also 256×256. In Fig. 3, the images on the left side from top to bottom, are the T1, T2, and Pd images. The T1 image is acquired with $T_R = 1000$ ms and $T_E = 20$ ms, the T2 image with $T_R = 2000$ and $T_E = 75$ ms, and the Pd image with $T_R = 3000$ and $T_E = 20$ ms. The images on the right hand side from top to bottom, are the segmentation results for $m = 4, 5,$ and 6, respectively.

B. Cluster Validation

To verify the cluster validation performance of the proposed MAP criterion, we applied it to the synthesized MR brain images as shown in Fig. 2. Table V displays the simulation results from 50 independent trials of the MAP, AIC, and MDL rules. As can be seen, the AIC and MDL showed strong tendencies to overestimate the number of classes. In particular, the AIC exhibited very poor performance and can be considered unreliable. By contrast, our criterion yielded excellent results.

VI. DISCUSSION

A. Sensitivity Evaluation

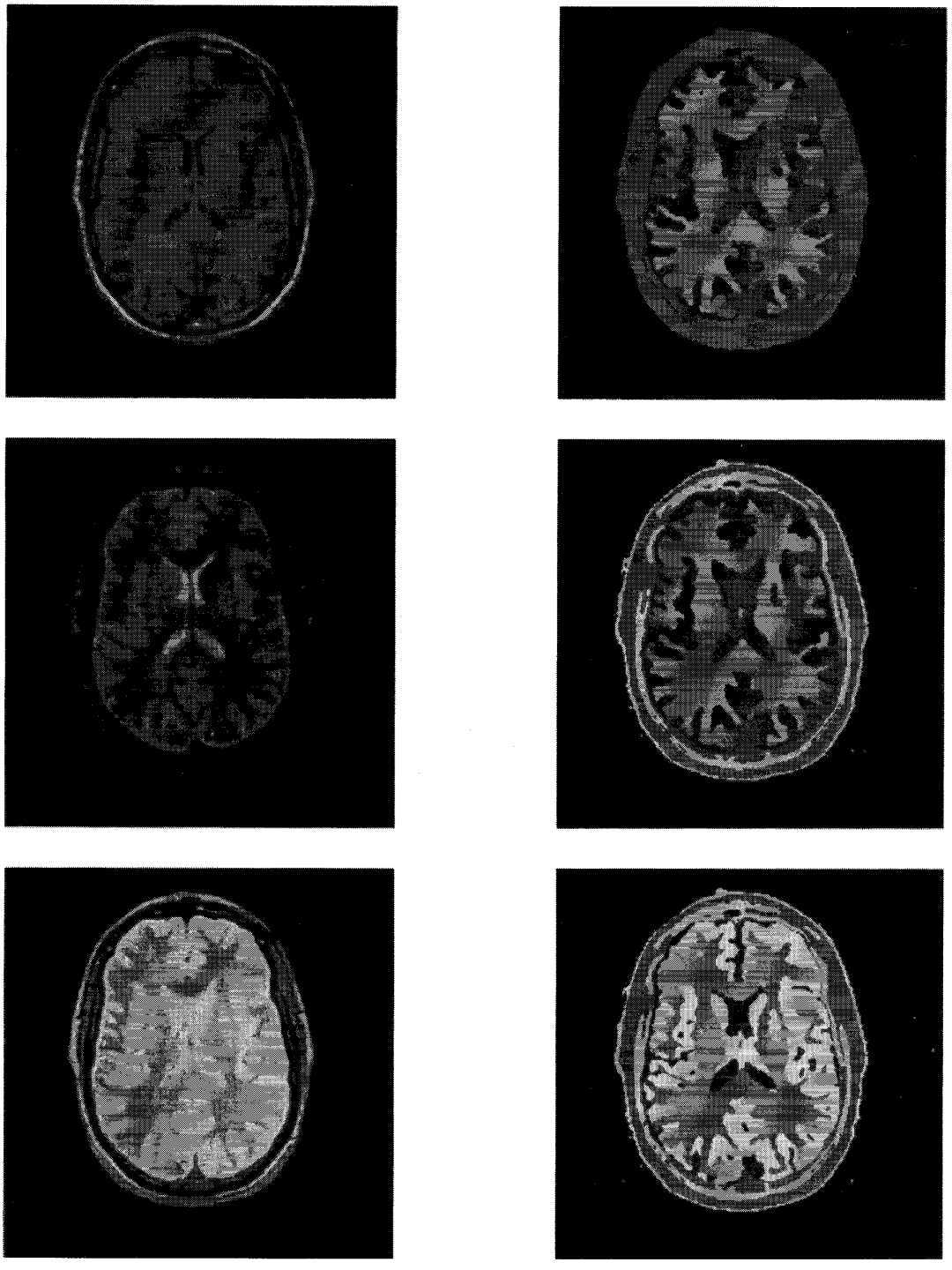
Cluster splitting can only perform well if the clusters contain large enough number of pixels, and there seems no obvious way to circumvent this problem. In order to examine this issue, we generated a small region (10×10 pixels) embedded in a larger region with different CNR's. The experiment was performed with various vector images who sizes were 32×32, 64×64, and 128×128 pixels. The results are listed in Table VI, where the entries represent the misclassified pixels (average value out of 50 trials) and "F" denotes more than 35 misclassified pixels or failure of discrimination.

As the image size gets larger, the required CNR for successful separation of the small object from the background increases. This shows that a subcluster can only be separated from the other cluster if its size is not too small compared to the other one. Of course, this also depends on the CNR. For example, at CNR = 2/1, the small object can be separated from the larger one, if their size ratio is larger than 1/10.

B. Computational Efficiency

It is known that the ICM is an efficient algorithm, when the class parameters are known. If the initial conditions are unknown, the TS algorithm provides an efficient way for obtaining them. The TS algorithm when applied to a 3-D (T1, T2, and Pd) 256×256 brain image with five classes, takes about 12 min on a Pentium 90 machine with 16 MB RAM and 32-MB swap memory. Although Dubes *et al.* [11] show that the ICM can converge in five or six raster scans of an image, we performed ten raster scans for the segmentations in our simulations.

The computational requirements mainly depend on the image size and the number of classes contained in the image. It does not vary dramatically when it is applied to different types of images. When the image has m different classes, the total computation requirement is $\sum_{i=2}^{m-1} i = m^2 - m - 1$ segmentations. For a medium size image (e.g., 256×256) with a medium number of classes (e.g., 10), the computational requirements are modest. When the image size and the number of classes are large, as can occur, for example, in some satellite image applications, the direct application of the TS algorithm may become computationally infeasible. Since a significant portion of the computation is used for initialization, to reduce the computational requirements, coarse segmentation can be applied to feature vectors extracted from $b \times b$ windows. Once the initial parameters are obtained, the segmentation can proceed as before. In some practical applications, this may decrease the computation by a factor of about $1/b^2$. In certain cases, however, a technique based on this approach may miss some extremely narrow regions. Another way to reduce the computational requirement in applications with larger number of classes is to start the TS–ICM from certain number of classes rather than always starting from one class. In addition, parallel computation can be straightforwardly applied which will further reduce the computation time.



(a) (b)

Fig. 3. (a) Three bands of the noisy real MR images (T1, T2, and Pd weighted). (b) From top to bottom, the segmented labels for $m = 4$, 5, and 6, respectively.

VII. Conclusions

We have presented a novel algorithm for simultaneous parameter estimation and vector image segmentation. The algorithm is implemented sequentially in stages, where at each stage a specific number of classes is assumed. It increases the number of classes by one at each stage by using a tree structure technique that provides initial parameter estimates which are then improved by the ICM procedure. The estimation results of every stage are used as initial estimates for the next stage. As the number of stages increases, the number of classes also increases. We have also proposed a MAP criterion for cluster validation. This criterion has the

form of a penalized likelihood function that is composed of two terms, a data term and a penalty term. The data term represents the fitness of the segmented image to the original image, and the penalty term penalizes for the spatial discontinuity and the overparameterization. The performance of the algorithm and the MAP criterion were examined by computer simulations on synthesized images, and they show excellent results. Segmentation results on real brain images are also included.

ACKNOWLEDGMENT

The authors are grateful to Prof. N. Phamdo for many fruitful suggestions and discussions, as well as to the reviewers for their constructive comments and suggestions, which have greatly improved this manuscript.

REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.,* vol. AC-19, pp. 716–723, 1974.
[2] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," *Behavioral Sci.,* vol. 12, pp. 153–155, 1967.
[3] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics,* pp. 803–821, 1993.
[4] J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Stat. Soc.,* ser. B, vol. 48, no. 3, pp. 259–302, 1986.
[5] ———, "Spatial interaction and statistical analysis of lattice systems," *J. Roy. Stat. Soc.,* ser. B, vol. 36, pp. 192–236, 1974.
[6] C. Bouman and B. Liu, "Multiple resolution segmentation of texture images," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 13, pp. 99–113, 1991.
[7] C. Bouman and M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. Image Processing,* vol. 3, pp. 162–177, 1994.
[8] C. C. Chu and J. K. Aggarwal, "The integration of image segmentation maps using region and edge information," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 15, pp. 1241–1252, 1993.
[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, " Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.,* ser. B, vol. 39, no. 1, pp. 1–38, 1977.
[10] P. M. Djurić, "Model selection based on asymptotic Bayes theory," in *Proc. 7th SP Workshop on Statistical Signal & Array Processing,* 1994, pp. 7–10.
[11] R. C. Dubes, A. K. Jain, S. G. Nadabar, and C. C. Chen, "MRF model-based algorithms for image segmentation," in *IEEE Conf. Pattern Recog..,* 1990, pp. 808–814.
[12] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973, pp. 332–335.
[13] B. S. Everitt, *Finite Mixture Distributions.* Chapman-Hall, 1981.
[14] D. Geiger and F. Girosi, "Parallel and deterministic algorithms from MRF's: Surface reconstruction," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 13, pp. 401–412, 1991.
[15] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. PAMI-6, pp. 721–741, 1984.
[16] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Boston: Kluwer, 1991, pp. 309–423.
[17] R. M. Gray, "Vector quantization," *IEEE Acout. Speech, Signal Processing Mag.,* vol. ASSP-1, pp. 4–29, 1984.
[18] Y. S. Han and W. E. Snyder, "Discontinuity-preserving vector smoothing on multivariate MR images via vector mean field annealing," *Proc. SPIE Math. Methods Med. Imag.,* pp. 69–80, 1992.
[19] J. A Hartigan, *Clustering Algorithms.* New York: Wiley , 1975, pp. 84–125.
[20] F. Heitz, P. Perez, and P. Bouthemy, "Multiscale minimization of global energy functions in some visual recovery problems," *CVGIP: Image Understanding,* vol. 59, pp. 125–134, 1994.
[21] C. L. Huang, T. Y. Cheng, and C. C. Chen, "Color images' segmentation using scale space filter and Markov random field," *Pattern Recog.,* vol. 25, pp. 1217–1229, 1992.
[22] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data.* Englewood Cliffs, NJ: Prentice Hall, 1988.
[23] F. C. Jeng and J. W. Wood, "Compound Gauss–Markov random fields for image estimation," *IEEE Trans. Signal Processing,* vol. 39, pp. 683–697, 1991.
[24] R. E. Kass and A. E. Raftery, "Bayes factors," *J. Amer. Stat. Assoc.,* vol. 90, pp. 737–795, 1995.
[25] Z. Kato, J. Zerubia, and M. Berthod, "Unsupervised adaptive image segmentation," in *Proc. ICASSP,* 1995, pp. 2399–2402.
[26] I. B. Kerfoot and Y. Bresler, "Design and theoretical analysis of a vector field segmentation algorithm," in *Proc. ICASSP,* 1993, vol. 5, pp. 5–8.
[27] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications.* Providence, RI: American Mathematical Society, 1980, vol. 1.
[28] D. A. Langan, J. W. Modestino, and J. Zhang, "Cluster validation for unsupervised stochastic model-based image segmentation," in *Proc. ICIP,* 1994, vol. 2 , pp. 197–201.
[29] S. Lakshmanan and H. Derin, "Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 11, pp. 799–813, 1989.
[30] R. Leahy, T. Hebert, and R. Lee, "Applications of Markov random fields in medical imaging," in *Proc. 11th Int. Conf. Inform. Processing Med. Imag.,* 1989, pp. 1–14.
[31] Y. G. Leclerc, "Constructing simple stable descriptions for image partitioning," *Int. J. Comput. Vision,* vol. 3, pp. 73–102, 1989.
[32] Z. Liang, R. J. Jaszczak, and R. E. Coleman, "Parameter estimation of finite mixtures using EM algorithm and information criteria with application to medical image processing," *IEEE Trans. Nucl. Sci.,* vol. 39, pp. 1126–1133, 1992.
[33] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Amer. Stat. Assoc.,* vol. 82, pp. 76–89, 1987.
[34] P. Masson and W. Pieczynski, "SEM algorithm and unsupervised statistical segmentation of satellite images," *IEEE Trans. Geosci. Remote Sensing,* vol. 31, pp. 618–633, 1993.
[35] K. L. Oehler and R. M. Gray, "Combining image compression and classification using vector quantization," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 17, pp. 461–473, 1995.
[36] D. K. Panjwani and G. Healey, "Unsupervised segmentation of textured color images using Markov random field models," in *IEEE Conf. Comput. Vision and Pattern Recog.,* 1993, pp. 776–777.
[37] T. N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Signal Processing,* vol. 40, pp. 901–914, 1992.
[38] N. Phamdo, N. Farvardin, and T. Moriya, "A unified approach to tree-structure and multistage vector quantization for noisy channels," *IEEE Trans. Inform. Theory,* vol. 39, pp. 835–850, 1993.
[39] I. Pitas and A. N. Venetsanopoulos, "Order statistics in digital image processing," *Proc. IEEE,* pp. 1893–1921, 1992.
[40] I. Pitas and P. Tsakalides, "Multivariate ordering in color image filtering," *IEEE Trans. Circ. Syst. Video Tech.,* vol. 1, pp. 247–259, 1991.
[41] J. G. Postaire and C. P. A. Vasseur, "An approximate solution to normal mixture identification with application to unsupervised pattern classification," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. PAMI-3, pp. 163–179, 1981.
[42] E. Rignot and R. Chellappa, "Segmentation of polarimetric synthetic aperture radar data," *IEEE Trans. Image Processing,* vol. 1, pp. 281–300, 1992.
[43] E. A. Riskin and R. M. Gray, "A greedy tree growing algorithm for the design of variable rate vector quantizers," *IEEE Trans. Signal Processing,* vol. 39, pp. 2500–2507, 1991.
[44] J. Rissanen, "Modeling by shortest data description," *Automatica,* vol. 14, pp. 465–478, 1978.
[45] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles.* Reading, MA: Addison-Wesley, 1974.
[46] C. Won and H. Derin, "Unsupervised segmentation of noisy and textured images using Markov random fields," *CVGIP: Graphical Models and Image Processing,* vol. 54, pp. 308–328, 1992.
[47] J. C. Yen, F. J. Chang, and S Chang, "A new criterion for automatic multilevel thresholding," *IEEE Trans. Image Processing,* vol. 4, pp. 370–378, 1995.
[48] J. Zhang and J. W. Modestino, "A model-fitting approach to cluster validation with application to stochastic model-base image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 12, pp. 1009–1017, 1990.
[49] J. Zhang, "The mean fields theory in EM procedures for Markov random fields," *IEEE Trans. Signal Processing,* vol. 40, pp. 2570–2583, 1992.