BAYESIAN ESTIMATION FOR MULTISCALE IMAGE SEGMENTATION

Srinivas Sista and R. L. Kashyap

School of Electrical & Computer Engineering
Purdue University, West Lafayette, IN 47907-1285
{sista,kashyap}@ecn.purdue.edu

ABSTRACT

We present a solution to the problem of intensity image segmentation using Bayesian estimation in a multiscale set up. Our approach regards the number of regions, the data partition and the parameter vectors that describe the probability densities of the regions as unknowns. We compute their MAP estimates jointly by maximizing their joint posterior probability density given the data. Since the estimation of the number of regions is also included into the Bayesian formulation we have a fully automatic or unsupervised method of segmenting images. An important aspect of our formulation is to consider the data partition as a variable to be estimated.

We provide a descent algorithm that starts with an arbitrary initial segmentation of the image when the number of regions is known and iteratively computes the MAP estimates of the data partition and the associated parameter vectors of the probability densities. Our method can incorporate any additional information about a region while assigning its probability density. It can also utilize any available training samples that arise from different regions.

1. INTRODUCTION

In image segmentation, the given image $Y = \{y_{i,j}, i,j = 0, \cdots, M-1\}$ has to be partitioned into mutually exclusive and totally inclusive subsets of Y namely $\boldsymbol{r} = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_s\}, \boldsymbol{r}_k \subseteq Y$ so that all the pixels belonging to a subset $\boldsymbol{r}_k$ are close to each other in some sense. We refer to each subset as a region. Further, we want to partition the image into b segments such that the segments are non-overlapping except for border pixels. In our terminology, each region contains one or more segments. The segments in a given region need not necessarily be spatially contiguous. All the pixels corresponding to the same region represent the same artifact like road, water, house, etc and are described by the same probability density function. The choice of s, the number of regions is itself a problem.

There exist several image segmentation techniques based on stochastic models [1, 2, 3]. These methods typically use Gaussian mixtures to model the regions and Gibbs random field models to model the region labels of the pixels. The associated model parameters are then estimated in an approximate MAP setting or using maximum likelihood(ML) estimation; with expectation maximization(EM) algorithm. The disadvantage of using mixture models is that even when a Gaussian mixture is used it is computationally expensive to compute a single segmentation because of the non convex nature of the cost function. Further there doesn't seem to exist a framework to systematically evaluate the obtained segmentations.

In the Bayes approach given in this paper the partition $\boldsymbol{r}$ is itself regarded as a variable to be chosen from the appropriate space. When s is known, $\boldsymbol{r} = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_s\}, \boldsymbol{r} \in \Omega_{s,s}$, the set of all partitions of set Y so that none of the sets $\boldsymbol{r}_k$ are null. When s is not specified, $s \leq s_0$, then $\boldsymbol{r} = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_{s_0}\}, \boldsymbol{r} \in \Omega_{s_0}$, the set of all partitions of Y into s_0 subsets. The pixels in the same region k are described by the probability density $p_k(y_{i,j} \mid \boldsymbol{\theta}_k)$, p_k is a known function and $\boldsymbol{\theta}_k$ is a vector parameter whose values have to be determined, $\boldsymbol{\theta}_k \in R^{n_k}$. So, the unknowns are $\{\boldsymbol{r} = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_s\}, \boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_s\}\}$, when s is known. Choosing $p_k(y_{i,j} \mid \boldsymbol{\theta}_k)$ to be Gaussian gives us a means of explicitly computing the estimates of these unknowns.

The Bayes approach allows us to estimate s, the number of classes given that $s \leq s_0$. Correspondingly the best segmentation $\boldsymbol{r}$ has to be searched in the space $\boldsymbol{r} \in \Omega_{s_0}$. It also solves the problem of comparing different segmentations. Two different segmentations $\boldsymbol{r}$ and $\boldsymbol{r}'$ involving different values of s can be compared by computing the ratio of the corresponding posterior probabilities $P(\boldsymbol{r} \mid Y)$ and $P(\boldsymbol{r}' \mid Y)$. Our method can also utilize any additional information on the classes in assigning the probability density function p_k. For example, when all the pixels $y_{i,j}$ are clustered tightly around a straight line

This work has been partially supported by National Science Foundation under contract IRI 9619812 and the office of Naval Research under contract N00014-91-J-4126.

or a convex curve or a 2-D plane.

2. BAYESIAN ESTIMATION

Let the data set be $Y = \{y_{i,j}, i,j = 0, \cdots, M-1\}, y_{i,j}$ whose members are statistically independent. Let s be the number of distinct classes in Y, s is known to us. Let the s associated probability densities be $p_k(y_{i,j} \mid \boldsymbol{\theta}_k), \boldsymbol{\theta}_k \in R^{n_k}, k = 1, \cdots, s$. Let the set $\boldsymbol{r} = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_s\}$ be a partition of Y into s classes such that

$$\begin{array}{c} \boldsymbol{r}_k \subseteq Y,\ \forall k = 1, \cdots, s;\ \boldsymbol{r}_i \cap \boldsymbol{r}_j = \text{Null}, i \neq j \\ \bigcup_{k=1}^{s} \boldsymbol{r}_k = Y. \quad \boldsymbol{r}_k \neq \text{ Null } \forall k = 1, \cdots, s. \end{array} \tag{1}$$

Each $\boldsymbol{r}_k$ is a subset of Y whose members are described by the density p_k. Let $\Omega_{s,s}$ be the set of all possible distinct partitions of Y obeying (1). $\boldsymbol{r}$ and $\boldsymbol{\theta}_k, k = 1, \cdots, s$ are the variables to be estimated. We regard $\boldsymbol{r} \in \Omega_{s,s}, \boldsymbol{\theta}_k \in R^{n_k}, k = 1, \cdots, s$ as independent random variables. $P(\boldsymbol{r})$, the prior probability associated with $\boldsymbol{r}$ is same for all $\boldsymbol{r}$; $P(\boldsymbol{r}) = \frac{1}{\#\Omega_{s,s}}, \quad \forall \boldsymbol{r} \in \Omega_{s,s}$. Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_s\}$. Let $p(\boldsymbol{\theta}_k)$ be the prior probability density of $\boldsymbol{\theta}_k$ such that each component is uniformly distributed. Since the priors of $\boldsymbol{\theta}$ and $\boldsymbol{r}$ are uniform, the MAP estimates $(\boldsymbol{r}^*, \boldsymbol{\theta}^*)$ are given by

$$(\boldsymbol{r}^*, \boldsymbol{\theta}^*) = \text{Arg} \max_{\boldsymbol{r}, \boldsymbol{\theta}} P(Y \mid \boldsymbol{r}, \boldsymbol{\theta}). \tag{2}$$

Since the data Y is independent, the joint density of Y has the following form:

$$P(Y \mid \boldsymbol{r}, \boldsymbol{\theta}) = \prod_{k=1}^{s} \left(\prod_{y_{i,j} \in \boldsymbol{r}_k} p_k(y_{i,j} \mid \boldsymbol{\theta}_k). \right) \tag{3}$$

For a fixed s, to obtain the MAP estimates of $\boldsymbol{r}$ and $\boldsymbol{\theta}$ we have to minimize the function

$$J_s(\boldsymbol{r}, \boldsymbol{\theta}) = -2 \sum_{k=1}^{s} \sum_{y_{i,j} \in \boldsymbol{r}_k} \ln p_k(y_{i,j} \mid \boldsymbol{\theta}_k). \tag{4}$$

For a fixed $\boldsymbol{\theta}$ the value of $\boldsymbol{r}$ which minimizes $J_s(\boldsymbol{r}, \boldsymbol{\theta})$ w.r.t $\boldsymbol{r}$ can be obtained using

$$\begin{aligned} \hat{\boldsymbol{r}}_{\boldsymbol{\theta},k} \quad = \quad & \{y_{i,j} : -\ln p_k(y_{i,j} \mid \boldsymbol{\theta}_k) \leq -\ln p_u(\boldsymbol{z_i} \mid \boldsymbol{\theta}_u), \\ & \forall k \neq u, u = 1, \cdots, s\}, k = 1, \cdots, s \end{aligned} \tag{5}$$

Similarly for a fixed $\boldsymbol{r}$, the minimizing value of $\boldsymbol{\theta}$ is unique and it can be obtained using

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{r},k} = \min_{\boldsymbol{\theta}_k \in R^{n_k}} \sum_{y_{i,j} \in \boldsymbol{r}_k} -\ln p_k(y_{i,j} \mid \boldsymbol{\theta}_k),\ k = 1, \cdots, s. \tag{6}$$

When p_k are given by $p_k(y_{i,j} \mid \boldsymbol{\theta}_k) \sim \text{Gauss}(\mu_k, \rho_k)$, $\boldsymbol{\theta}_k = \{\mu_k, \rho_k\}$ an explicit expression for $\hat{\boldsymbol{\theta}}_{\boldsymbol{r},k}$ can be given because of the structure of $-\ln p_k(\cdot)$ as follows

$$-2 \ln p_k(y_{i,j} \mid \boldsymbol{\theta}_k) = ((y_{i,j} - \mu_k)^2 / \rho_k) + \ln 2\pi\rho_k \tag{7}$$

The parameter estimates with $N_{1k} = \#\boldsymbol{r}_k$ are

$$\begin{aligned} \hat{\mu}_{\boldsymbol{r},k} \quad &= \quad \frac{1}{N_{1k}} \sum_{y_{i,j} \in \boldsymbol{r}_k} y_{i,j} \\ \hat{\rho}_{\boldsymbol{r},k} \quad &= \quad \frac{1}{N_{1k}} \sum_{y_{i,j} \in \boldsymbol{r}_k} (y_{i,j} - \hat{\mu}_{\boldsymbol{r},k})^2 \end{aligned} \tag{8}$$

We use a simple descent algorithm for finding a local minimum of $J_s(\boldsymbol{r}, \boldsymbol{\theta})$. It is done by changing $\boldsymbol{\theta}$ and $\boldsymbol{r}$ alternatively using expressions (5) and (6), each time having a reduction in $J(\boldsymbol{r}, \boldsymbol{\theta})$. Note that this method yields a local minimum which need not be a global minimum, since we perturb only $\boldsymbol{r}$ or $\boldsymbol{\theta}$ at one time, not simultaneously.

Descent Algorithm

1. Let $\boldsymbol{r}^j = (\boldsymbol{r}_1^j, \cdots, \boldsymbol{r}_s^j)$ and $\boldsymbol{\theta}^j = (\boldsymbol{\theta}_1^j, \cdots, \boldsymbol{\theta}_s^j)$ be estimates at the end of j^{th} iteration. Choose $\boldsymbol{r}^1$ arbitrarily, perhaps from a solution of a clustering algorithm with random seeds.
2. Given $\boldsymbol{r}^{(j)}$, compute $\boldsymbol{\theta}^{(j)}$ using the formula in (6).
3. Given $\boldsymbol{\theta}^{(j)}$, compute $\boldsymbol{r}^{(j+1)}$ using (5).
4. Stop if $\boldsymbol{r}^{(j)} = \boldsymbol{r}^{(j+1)}$; otherwise goto 2.

End.

Note that the computational effort for finding a local minimum is very little. It involves only data comparisons in (5) apart from evaluating the expressions in (8). Since J_s decreases with each iteration and is bounded below by zero, the algorithm is assured to converge to a fixed point. This fixed point is obtained when $\boldsymbol{r}^{(j)} = \boldsymbol{r}^{(j+1)}$. Moreover this convergence happens in a finite number of steps because the size of the set $\Omega_{s,s}$ is finite.

Choice of s, the number of regions

The problem of choosing the value of s is also known as model order identification or cluster validation. In our method we obtain the estimate of s via Bayesian estimation by considering s also as a random variable. The optimal Bayes estimator of $(s, \boldsymbol{r}, \boldsymbol{\theta})$ is given by

$$(s^*, \boldsymbol{r}_{s^*}^*, \boldsymbol{\theta}^*) = \text{Arg} \min_{1 \leq s \leq s_0} \left\{ \min_{\boldsymbol{r} \in \Omega_{s,s}} \min_{\boldsymbol{\theta}_k \in R^{n_k}} H_s \right\} \tag{9}$$

where

$$H_s = -\ln \left\{ p(Y \mid s, \boldsymbol{r}, \boldsymbol{\theta}) P(\boldsymbol{r} \mid s) \left(\prod_{k=1}^{s} p(\boldsymbol{\theta}_k \mid s) \right) P(s) \right\}$$

The prior probabilities of s and $\boldsymbol{r}$ given s are chosen as follows:

$$P(s) \quad = \quad 1/s_0,\ s = 1, \cdots, s_0 \tag{10}$$

$$P(\boldsymbol{r} \mid s) \quad = \quad \frac{1}{\#\Omega_{s,s}}, \quad \sum_{\boldsymbol{r} \in \Omega_{s,s}} P(\boldsymbol{r} \mid s) = 1 \tag{11}$$

the prior probability of each component in $\boldsymbol{\theta}_k$ is uniform and equals $1/L_k$. Since L_k is the prior density of $\boldsymbol{\theta}_k$, it should cover the total range of all the components of $\boldsymbol{\theta}_k$.

3. MULTISCALE IMAGE SEGMENTATION

We assume that the $y_{i,j}$ are clustered around polynomials specifiable through facet models.

$$y_{i,j} = \alpha_0 + \alpha_1 i + \alpha_2 j + \alpha_3 ij + \eta_{i,j} \tag{12}$$

$\eta_{i,j} \sim \text{Gauss}(0, \rho_k)$ is the white noise with variance ρ_k. Then the density of $y_{i,j}$ belonging to the k^{th} segment is given by

$$p_k(y_{i,j} \mid \boldsymbol{\theta}_k) = \text{Gauss}\left(\alpha_0 + \alpha_1 i + \alpha_2 j + \alpha_3 ij,\ \rho_k\right) \tag{13}$$

Note all the pixels belonging to region k have the same variance but not the same mean owing to the dependence on i and j. Let $f_k(y_{i,j}; \boldsymbol{\theta}_k) = -2 \ln p_k(y_{i,j} \mid \boldsymbol{\theta}_k)$. For a fixed s, to obtain the MAP estimates of $\boldsymbol{r}$ and $\boldsymbol{\theta}$ we have to minimize the function

$$J_s(\boldsymbol{r}, \boldsymbol{\theta}) = \sum_{k=1}^{s} \sum_{y_{ij} \in \boldsymbol{r}_k} f_k(y_{i,j}; \boldsymbol{\theta}_k) \tag{14}$$

Segmentation with multiple scales: Since the number of pixels N is large we have to carry out the partition at several scales. In the beginning let us deal with blocks of pixels say 4×4. Let the block be denoted by the leading pixel. For instance the block $\{(i+k, j+l), k, l = 0, 1, 2, 3\}$ will be denoted by $b_{i,j}$. We assign the entire pixel block to one region $\boldsymbol{r}_k$ in the partition. Note we are not averaging the intensities in the block. Each pixel retains its identity. Thus we have $(N/16) = N_1$ blocks. The region assignment of the block is given by

Assign all pixels $\in b_{i,j}$ to $\boldsymbol{r}_k$ if

$$\left\{ \sum_{i+u} \sum_{j+v} f_k(y_{i,j}; \boldsymbol{\theta}_k) \leq \sum_{i+u} \sum_{j+v} f_u(y_{i,j}; \boldsymbol{\theta}_k) \quad \forall u \right\} \tag{15}$$

Partition at the coarse level: Since in the example we deal with 80×80 image, $N = 6400$. We carry out segmentation at 3 levels: 4×4, 2×2 and the finest level. Consider the coarsest level. Let

$$Y_2 = \{b_2^{2i,2j}, 0 \leq i, j \leq 39\} \tag{16}$$

$$Y_4 = \{b_4^{4i,4j}, 0 \leq i, j \leq 19\} \tag{17}$$

where $b_k^{u,v} = \{y_{u+i,v+j}, 0 \leq i, j \leq k-1\}$. Let the corresponding partition be $\boldsymbol{r}_4 = \{\boldsymbol{r}_{4,1}, \cdots, \boldsymbol{r}_{4,s}\}$ where $\boldsymbol{r}_{4,s} \subseteq Y_4$. All the pixels in the same block $b_4^{u,v}$ will have the same region assignment, i.e they are assigned the same density $p_k(\cdot \mid \boldsymbol{\theta}_k)$.

For a given partition $\boldsymbol{r}_4$, $\boldsymbol{\theta}_k$ is computed for the 1×1 pixel intensities $y_{i,j}$ in all the blocks $b_4^{u,v}$ assigned to $\boldsymbol{r}_{4,k}$ as indicated. For a given $\boldsymbol{\theta}_k, k = 1, \cdots, s$ the partition is updated as follows

Assign $b^{u,v}$ to $\boldsymbol{r}_{4,k}$ if

$$\left[\sum_{0 \leq i,j \leq 3} f_k(y_{u+i,v+j}; \boldsymbol{\theta}_k) \leq \sum_{0 \leq i,j \leq 3} f_l(y_{u+i,v+j}; \boldsymbol{\theta}_l),\ \forall l \right]$$

Thus we get the best partition $\boldsymbol{r}_4^* = \{\boldsymbol{r}_{4,1}^*, \cdots, \boldsymbol{r}_{4,s}^*\}$.

Partition at coarse level 2×2: We divide the 4×4 blocks into 2 groups, the boundary or B blocks and non-boundary or NB blocks. Each block has 4 immediate neighbors: top, bottom, left, right. Two neighboring blocks are labeled NB if their region labels are not different. The important idea here is the region assignments made to the 1×1 pixels in the 4×4 NB blocks are fixed and not altered in subsequent iterations. Only the assignments of pixels in B blocks are altered. All the 2×2 blocks derived from 4×4 NB retain the region type and their region labels are not altered in iteration. At every iteration every member of the block in the B type is assigned to $\boldsymbol{r}_{2,1}, \cdots, \boldsymbol{r}_{2,s}$ as the case may be. Computation of $\boldsymbol{\theta}$ and updating of $\boldsymbol{r}_2$ is similar to the earlier case of 4×4. After arriving at 1×1 level, the final result is cleaned by averaging over a 5×5 window.

Choice of s

The best value of s is that which minimizes H_s. However, in our multiscale scheme, it is more robust to decide on the value of s at a coarser scale itself. So the value of s is decided at the scale where all the pixels in blocks of size 4×4 have the same region assigned to them.

Example: (Synthetic Image)

We consider a synthetic image made up of three textures from the Brodatz album. The reason for considering a synthetic texture image is that the ground truth of the underlying segmentation is available. The image is 80×80 made of 5 segments and 3 regions. The original image is in Figure 1(a).

The segmentation at level 4×4 involves $N = 400$ blocks. We begin with a random initial partition and derive the associated local minimum. Several different local minima are derived. The best local minimum is displayed in Figure 1(b) and the associated initial partition in Figure 1(c). For segmentation at level 2×2 the $\{\boldsymbol{\theta}_k, k = 1, \cdots, s\}$ obtained from 4×4 level can serve as the starting point. The final result is given in Figure 1(d). The result of segmentation at the lowest level is displayed in Figure 1(e) and the cleaned image in Figure 1(f). The number of errors in the final segmentation at the pixel level is 63 which corresponds to 1% misclassification error. We note that the boundaries are visually perfect.

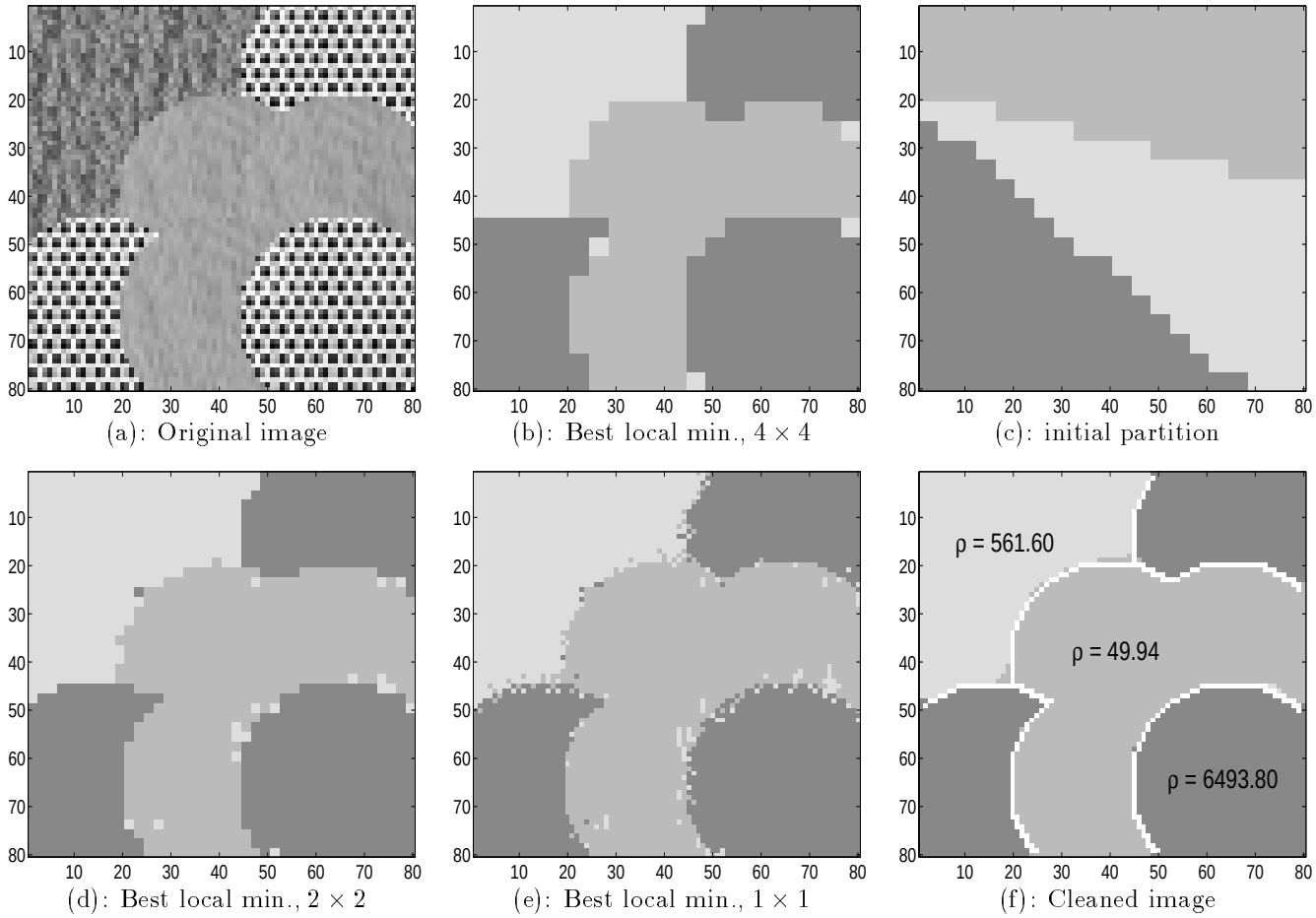

Figure 1: Image segmentation on texture image, $N = 6400$, for $s = 3$. (b) Classification at scale 4×4. (c) Initial partition that gave (b). (d) Classification at scale 2×2 starting from (b). (e) Classification at scale 1×1 starting from (d). (f) Cleaned version of (e).

Choice of s

The values of H_s for $s = 2, 3$ and 4 are $63032.97, 61550.89$ and 61707.41 respectively. The value of H_s is minimum for $s = 3$ which is the actual number of distinct textures present in the image.

4. CONCLUSION

We proposed a solution to the problem of image segmentation based on Bayesian estimation. The new feature of our method is, we regard the data partition as a variable to be estimated. We developed a Bayesian framework to estimate the number of classes, the class parameters and the data partition simultaneously. We presented a synthetic image example. Results on real images are also quite encouraging [4].

5. REFERENCES

[1] H. Derin and H. Elliot, "Modeling and segmentation of noisy and textured images using gibbs random fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 9, No. 1, pp. 39–55, January 1987.

[2] D. A. Langan, J. W. Modestino and J. Zhang, "Cluster validation for unsupervised stochastic model-based image segmentation," *IEEE Trans. on Image Processing*, Vol. 7, No. 2, pp. 180–195, February 1998.

[3] C. A. Bouman and B. Liu, "Multiple Resolution Segmentation of Textured Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 13, No. 2, pp. 99–113, February 1991.

[4] R. L. Kashyap and Srinivas Sista, "Unsupervised Classification and Choice of Classes: Bayesian Approach," *Technical Report TR-ECE 98-12*, School of Electrical and Computer Engineering, Purdue University, July 1998.