12

Applications of Neural Networks to Biomedical Image Processing

Tülay Adali
University of Maryland

Yue Wang
Catholic University or America

Huai Li
University of Maryland

12.1 Introduction

Fueled by the rapid growth in the development of medical imaging technologies and the increasing availability of computing power, biomedical image processing emerged as one of the most active research areas of recent years. With their rich information content, biomedical images are opening entirely new areas of research and are posing new challenges to researchers. Neural networks, among other approaches, have demonstrated a growing importance in the area and are increasingly used for a variety of biomedical image processing tasks. These include detection and characterization of disease patterns, analysis (quantification and segmentation), compression, modeling, motion estimation, and restoration of images from a variety of imaging modalities such as magnetic resonance (MR), positron emission tomography (PET), ultrasound, radiography, and mammography images. For a recent collection of examples of neural network applications to biomedical signal and image processing, see the papers in references [1] and [9].

This chapter concentrates on two specific application areas that are increasingly important: image analysis (quantification and segmentation) and computer assisted (aided) diagnosis (CAD) system design. Both applications demonstrate the unique ways neural structures and learning algorithms can effectively be used in the biomedical domain.

The image analysis application discussed here demonstrates an example of unsupervised learning with a finite mixture network that is intimately related to the radial basis function network. We show that the image context can be modeled by a localized mixture model and that the final image segmentation can be achieved by a probabilistic constraint relaxation network. We give examples of the application of the framework in analyses of MR and mammographic images.

The second application, CAD design, demonstrates two specific ways neural networks can be used for classification (detection) type tasks in biomedical image processing. In the first application, meaningful features that identify the disease patterns of interest are extracted and input into a neural classifier. The second CAD system introduced herein relies on a convolutional neural network

0-8493-2359-2/01/$0.00+$1.50

(CNN) to extract features of disease patterns internally and, hence, to learn to distinguish them from non-disease patterns. We show application of the first CAD to mass detection in mammograms and the second CAD, based on CNN, is applied to detection of clustered microcalcifications.

12.2 Biomedical Image Analysis

Model-based image analysis aims at capturing the intrinsic character of images with few parameters and is also instrumental in helping to understand the nature of the imaging process. Key issues in image analysis include model selection, parameter estimation, imaging physics, and the relationship of the image to the task (how the image is going to be utilized) [2, 3]. Stochastic model-based image analysis has been the most popular among the model-based image analysis methods as, most often, imaging physics can be modeled effectively with a stochastic model. For example, the suitability of standard finite normal mixture models has been verified for a number of medical imaging modalities [4]–[7]. This section discusses a complete treatment of stochastic model-based image analysis that includes model and model order selection, parameter estimation, and final segmentation. We focus on models that use finite mixtures and show examples in MR and mammographic image analysis.

In image analysis, we can treat pixel and context modeling separately, assuming that each pixel can be decomposed into a pixel image and a context image. Pixel image is defined as the observed gray level associated with the pixel, and finite mixture models have been the most popular pixel image models. In particular, standard finite normal mixtures (SFNMs) have been widely used in statistical image analysis, and efficient algorithms are available for calculating the parameters of the model. Furthermore, by incorporating statistical properties of context images, where context image is defined as the membership of the pixel associated with different regions, a localized SFNM formulation can be used to impose local consistency constraints on context images in terms of a stochastic regularization scheme [8].

The next section describes the finite mixtures model and addresses identification of the model, i.e., estimation of the parameters of the model and the model order selection. Section 12.2.2 discusses approaches to modeling context and address segmentation of the MR image into different tissue components.

12.2.1 Pixel Modeling

Imagine a digital image of $N \equiv N_1 \times N_2$ pixels. Assume that this image contains K regions and that each pixel is decomposed into a pixel image x and a context image l. By ignoring information regarding the spatial ordering of pixels, we can treat context images (i.e., pixel labels) as random variables and describe them using a multinomial distribution with unknown parameters π_k. Since this parameter reflects the distribution of the total number of pixels in each region, π_k can be interpreted as a prior probability of pixel labels determined by the global context information. Thus, the relevant (sufficient) statistics are the pixel image statistics for each component mixture and the number of pixels of each component. The marginal probability measure for any pixel image, i.e., the finite mixtures distribution, can be obtained by writing the joint probability density of x and l and then summing the joint density over all possible outcomes of l, i.e., by computing $p(x_i) = \Sigma_l p(x_i, l)$, resulting in a sum of the following general form:

$$p_r(x_i) = \sum_{k=1}^{K} \pi_k p_k(x_i), \quad i = 1, \ldots, N \tag{12.1}$$

where x_i is the gray level of pixel i. $p_k(x_i)$s are conditional region probability density functions (PDFs) with the weighting factor π_k, satisfying $\pi_k > 0$, and $\Sigma_{k=1}^{K} \pi_k = 1$. The generalized Gaussian

PDF given region k is defined by [10]:

$$p_k(x_i) = \frac{\alpha\beta_k}{2\Gamma(1/\alpha)} \exp\left[-\left|\beta_k(x_i - \mu_k)\right|^{\alpha}\right], \quad \alpha > 0, \quad \beta_k = \frac{1}{\sigma_k}\left[\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}\right]^{1/2} \tag{12.2}$$

where μ_k is the mean, $\Gamma(\cdot)$ is the gamma function, and β_k is a parameter related to the variance σ_k by

$$\beta_k = \frac{1}{\sigma_k}\left[\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}\right]^{1/2} . \tag{12.3}$$

When $\alpha \gg 1$, the distribution tends to a uniform PDF; for $\alpha < 1$, the PDF becomes sharper; for $\alpha = 2.0$, one has the Gaussian (normal) PDF; and for $\alpha = 1.0$, one has the Laplacian PDF. Therefore, the generalized Gaussian model is a suitable model to fit the histogram distribution of those images whose statistical properties are unknown since the kernel shape can be controlled by selecting different α values. The finite Gaussian mixture model for $\alpha = 2$ is also commonly referred to as the standard finite normal mixture model which is the identification we adopt in this chapter. It can be written as

$$p_k(x_i) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right) \quad i = 1, 2, \ldots, N \tag{12.4}$$

where μ_k and σ_k^2 are the mean and variance of the kth normal kernel and K is the number of normal components.

The whole image can be closely approximated by an independent and identically distributed random field $\mathbf{X}$. The corresponding joint PDF is

$$P(\mathbf{x}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k p_k(x_i) \tag{12.5}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_N]$, and $\mathbf{x} \in \mathbf{X}$. Based on the joint probability measure of pixel images, the likelihood function under finite mixture modeling can be expressed as

$$\mathcal{L}(\mathbf{r}) = \prod_{i=1}^{N} p_{\mathbf{r}}(x_i) \tag{12.6}$$

where $\mathbf{r} : \{K, \alpha, \pi_k, \mu_k, \sigma_k, k = 1, \ldots, K\}$ denotes the model parameter set. Note that $p_{\mathbf{r}}(x_i)$ refers to the joint density defined in Equation (12.1); however, we have added the subscript $\mathbf{r}$ to emphasize that it is a parameterized density.

Maximization of the likelihood yields parameters for the chosen distribution given the observations, i.e., the pixel images. Once the model is chosen, identification addresses the estimation of local region parameters $(\pi_k, \mu_k, \sigma_k, k = 1, \ldots, K)$ and the structural parameters (K, α). In particular, the estimation of the order parameter K is referred to as model order selection.

12.2.1.1 Parameter Estimation

With an appropriate system likelihood function, the objective of model identification is to estimate the model parameters by maximizing the likelihood function. This is equivalent to minimization of relative entropy between the image histogram $p_{\mathbf{x}}(u)$ and the estimated PDF $p_{\mathbf{r}}(u)$, where u is the gray level [11, 12] as relative entropy measures the information theoretic distance between two distributions and is zero only when the two distributions match.

There are a number of approaches to perform the maximum likelihood (ML) estimation of finite mixture distributions [13]. The most popular method is the expectation–maximization (EM) algorithm [14, 15]. EM algorithm first calculates the posterior Bayesian probabilities of the data based on the observations and obtains the current parameter estimates (E-step). It then updates parameter estimates using generalized mean ergodic theorems (M-step). The procedure moves back and forth between these two steps. The successive iterations increase the likelihood of the model parameters being estimated. A neural network interpretation of this procedure is given by Perlovsky and McManus [16].

We can use relative entropy (the Kullback–Leibler distance) [17] for parameter estimation, i.e., we can measure the information theoretic distance between the histogram of the pixel images, denoted by $p_{\mathbf{x}}$, and the estimated distribution $p_{\mathbf{r}}(u)$ which we define as the global relative entropy (GRE):

$$D\left(p_{\mathbf{x}} \| p_{\mathbf{r}}\right) = \sum_{u} p_{\mathbf{x}}(u) \log \frac{p_{\mathbf{x}}(u)}{p_{\mathbf{r}}(u)} . \tag{12.7}$$

It can be shown that, when relative entropy is used as the distance measure, distance minimization is equivalent to the ML estimation of the model parameters [11, 12].

For the case of the FGGM model, the EM algorithm can be applied to the joint estimation of the parameter vector and the structural parameter α as follows [14]:

EM Algorithm

1. For $\alpha = \alpha_{\min}, \ldots, \alpha_{\max}$:
   - $m = 0$, given initialized $\mathbf{r}^{(0)}$.
   - E-step: for $i = 1, \ldots, N,\ k = 1, \ldots, K$, compute the probabilistic membership:

$$z_{ik}^{(m)} = \frac{\pi_k^{(m)} p_k(x_i)}{\sum_{k=1}^{K} \pi_k^{(m)} p_k(x_i)} \tag{12.8}$$

   - M-step: for $k = 1, \ldots, K$, compute the updated parameter estimates:

$$\begin{cases} \pi_k^{(m+1)} = \frac{1}{N} \sum_{i=1}^{N_1 N_2} z_{ik}^{(m)} \\ \mu_k^{(m+1)} = \frac{1}{N \pi_k^{(m+1)}} \sum_{i=1}^{N} z_{ik}^{(m)} x_i \\ \sigma_k^{2(m+1)} = \frac{1}{N \pi_k^{(m+1)}} \sum_{i=1}^{N} z_{ik}^{(m)} \left(x_i - \mu_k^{(m+1)}\right)^2 \end{cases} \tag{12.9}$$

   - When $|GRE^{(m)}(p_{\mathbf{x}} \| p_{\mathbf{r}}) - GRE^{(m+1)}(p_{\mathbf{x}} \| p_{\mathbf{r}})| \leq \epsilon$ is satisfied, go to step 2. Otherwise, $m = m + 1$ and go to E-Step.
2. Compute GRE, and go to step 1.
3. Choose the optimal $\hat{\mathbf{r}}$ which corresponds to the minimum GRE.

However, the EM algorithm generally has the reputation of being slow, since it has a first order convergence in which new information acquired in the expectation step is not used immediately [18]. Recently, a number of online versions of the EM algorithm were proposed for large scale sequential learning [11, 13, 19, 20]. Such a procedure eliminates the need to store all the incoming observations and changes the parameters immediately after each data point, allowing for high data rates. Titterington et al. [13] present a stochastic approximation procedure that is closely related to the probabilistic self-organizing mixtures (PSOM) algorithm we introduce below. Other similar formulations for normal mixture parameter estimation are due to Marroquin and Girosi [19] and Weinstein et al. [20].

For the adaptive estimation of the SFNM model parameters, we can derive an incremental learning algorithm by simple stochastic gradient descent minimization of $D(p_{\mathbf{x}}||p_{\mathbf{r}})$ [5, 11] given in Equation (12.7) with $p_{\mathbf{r}}$ given by Equation (12.4):

$$\begin{aligned} \mu_k{}^{(t+1)} &= \mu_k^{(t)} + a(t)\left(x_{t+1} - \mu_k^{(t)}\right) z_{(t+1)k}^{(t)} && (12.10) \\ \sigma_k^{2(t+1)} &= \sigma_k^{2(t)} + b(t)\left[\left(x_{t+1} - \mu_k^{(t)}\right)^2 - \sigma_k^{2(t)}\right] z_{(t+1)k}^{(t)} && (12.11) \\ k &= 1, \ldots, K \end{aligned}$$

where $a(t)$ and $b(t)$ are introduced as learning rates, two sequences converging to zero, and ensuring unbiased estimates after convergence.

Updates for the constrained regularization parameters, π_k in the SFNM model, are obtained using a recursive sample mean calculation based on a generalized mean ergodic theorem [21]:

$$\pi_k^{(t+1)} = \frac{t}{t+1}\pi_k^{(t)} + \frac{1}{t+1} z_{(t+1)k}^{(t)} . \qquad (12.12)$$

Hence, the updates given by Equations (12.10), (12.11), and (12.12), together with an evaluation of Equation (12.8) using Equation (12.4), provide the incremental procedure for computing the SFNM component parameters. Their practical use requires a strong mixing condition and a decaying annealing procedure (learning rate decay) [21]–[23]. For details of the derivation, see Wang et al. [9, 11]. In finite mixture parameter estimation, the algorithm initialization must also be carried out carefully. Wang et al. [24] introduced an adaptive Lloyd–Max histogram quantization (ALMHQ) algorithm for threshold selection which is also well suited to initialization in an ML estimation. It can be used for initializing the network parameters: μ_k, σ_k^2, and π_k, k, $1, 2, \ldots, K$.

12.2.1.2 Model Order Selection

The determination of the region parameter K directly affects the quality of the resulting model parameter estimation and, in turn, affects the result of segmentation. In a statistical problem formulation such as the one introduced in the previous section, the use of information theoretic criteria for the problem of model determination arises as a natural choice. Two popular approaches are Akaike's information criterion (AIC) [25], and Rissanen's minimum description length (MDL) [26]. Akaike proposes the selection of the model that gives the minimum AIC, which is defined by

$$\mathrm{AIC}\,(K_a) = -2\log\left(\mathcal{L}\left(\hat{\mathbf{r}}_{ML}\right)\right) + 2K_a \qquad (12.13)$$

where $\hat{\mathbf{r}}_{ML}$ is the maximum likelihood estimate of the model parameter set $\mathbf{r}$, and K_a is the number of free adjustable parameters in the model [4, 25] and is given by $3K - 1$ for the SFNM model. The AIC selects the correct number of the image regions K_0 such that

$$K_0 = \arg\left\{\min_{1\le K\le K_{\mathrm{MAX}}} \mathrm{AIC}\,(K_a)\right\} . \qquad (12.14)$$

Rissanen addresses the problem from a quite different point of view. He reformulates the problem explicitly as an information coding problem in which the best model fit is measured such that high probabilities are assigned to the observed data, while at the same time the model itself is not too complex to be described [26]. The model is selected by minimizing the total description length defined by

$$\mathrm{MDL}\,(K_a) = -\log\left(\mathcal{L}\left(\hat{\mathbf{r}}_{ML}\right)\right) + 0.5K_a\log(N) . \qquad (12.15)$$

Similarly, the correct number of distinctive image regions K_0 can be estimated as

$$K_0 = \arg\left\{ \min_{1 \leq K \leq K_{\mathrm{MAX}}} \mathrm{MDL}(K_a) \right\} . \tag{12.16}$$

It is also worth noting that Schwartz arrives at the same formulation by using Bayesian arguments [27]. Wax and Kailath [28] provide a good introduction to model order selection for signal processing.

12.2.2 Context Modeling and Segmentation

Once the pixel model is estimated, the segmentation problem is the assignment of labels to each pixel in the image. A straightforward solution is to label pixels into different regions by maximizing the individual likelihood function $p_k(x)$, i.e., by performing ML classification. Usually, this method may not achieve a good performance since it does not use local neighborhood information in the decision. The CBRL algorithm [29] is one approach that can incorporate the local neighborhood information into the labeling procedure and, thus, improve the segmentation performance. The CBRL algorithm to perform/refine pixel labeling based on the localized FGGM model can be defined as follows [7].

Let ∂i be the neighborhood of pixel i with an $m \times m$ template centered at pixel i. An indicator function is used to represent the local neighborhood constraints $R_{ij}(l_i, l_j) = I(l_i, l_j)$, where l_i and l_j are labels of pixels i and j, respectively. Note that pairs of labels are now either compatible or incompatible. Similar to the procedure described in Hummel and Zucker [29], one can compute the frequency of neighbors of pixel i which has the same label values k as at pixel i:

$$\pi_k^{(i)} = p\left(l_i = k | \mathbf{l}_{\partial i}\right) = \frac{1}{m^2 - 1} \sum_{j \in \partial i, j \neq i} I\left(k, l_j\right) \tag{12.17}$$

where $\mathbf{l}_{\partial i}$ denotes the labels of the neighbors of pixel i. Since $\pi_k^{(i)}$ is a conditional probability of a region, the localized FGGM PDF of gray-level x_i at pixel i is given by:

$$p\left(x_i | \mathbf{l}_{\partial i}\right) = \sum_{k=1}^{K} \pi_k^{(i)} p_k\left(x_i\right) \tag{12.18}$$

where $p_k(x_i)$ is given in Equation (12.2). Assuming gray values of the image are conditional independent, the joint PDF of $\mathbf{x}$, given the context labels $\mathbf{l}$, is

$$P(\mathbf{x}|\mathbf{l}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k^{(i)} p_k\left(x_i\right) \tag{12.19}$$

where $\mathbf{l} = (l_i : i = 1, \ldots, N)$.

It is important to note that the CBRL algorithm can obtain a consistent labeling solution based on the localized FGGM model Equation (12.18). Since $\mathbf{l}$ represents the labeled image, it is consistent if $S_i(l_i) \geq S_i(k)$, for all $k = 1, \ldots, K$ and for $i = 1, \ldots, N$ [29], where

$$S_i(k) = \pi_k^{(i)} p_k\left(x_i\right) . \tag{12.20}$$

Now we can define

$$A(\mathbf{l}) = \sum_{i=1}^{N} \left(\sum_{k} I(l_i, k) S_i(k) \right) \tag{12.21}$$

as the average measure of local consistency, and

$$LC_i = \sum_k I\,(l_i, k)\, S_i(k), \quad i = 1, \ldots, N \tag{12.22}$$

represents the local consistency based on $\mathbf{l}$. The goal is to find a consistent labeling $\mathbf{l}$ which can maximize Equation (12.21). In the real application, each local consistency measure LC_i can be maximized independently. Hummel and Zucker have shown that when $R_{ij}(l_i, l_j) = R_{ji}(l_j, l_i)$, if $A(\mathbf{l})$ attains a local maximum at $\mathbf{l}$, then $\mathbf{l}$ is a consistent labeling [29].

Based on the localized FGGM model, $l_i^{(0)}$ can be initialized by the ML classifier,

$$l_i^{(0)} = \arg\left\{\max_k \; p_k\,(x_i)\right\}, \quad k = 1, \ldots, K\,. \tag{12.23}$$

Then, the order of pixels is randomly permutated and each label l_i is updated to maximize LC_i, i.e., classify pixel i into kth region if

$$l_i = \arg\left\{\max_k \; \pi_k^{(i)} p_k\,(x_i)\right\}, \quad k = 1, \ldots, K \tag{12.24}$$

where $p_k(x_i)$ is given in Equation (12.2), $\pi_k^{(i)}$ is given in Equation (12.17). By considering Equations (12.23) and (12.24), we can give a modified CBRL algorithm as follows [7]:

CBRL Algorithm

1. Given $\mathbf{l}^{(0)}$, $m = 0$.
2. Update pixel labels:
   - Randomly visit each pixel for $i = 1, \ldots, N$.
   - Update its label l_i according to:

$$l_i^{(m)} = \arg\left\{\max_k \pi_k^{(i)(m)} p_k\,(x_i)\right\}\,.$$

3. When $\frac{\Sigma(\mathbf{l}^{(m+1)} \oplus \mathbf{l}^{(m)})}{N_1 N_2} \leq 1\%$, stop; otherwise, $m = m + 1$, and repeat step 2.

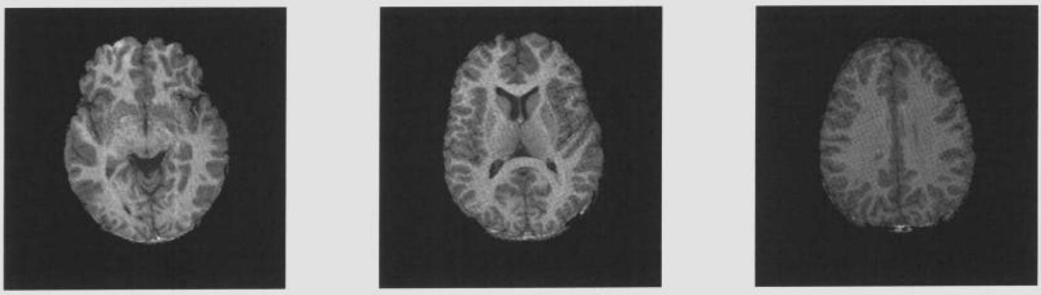
12.2.3 Application Examples

We present two examples to demonstrate the application of the stochastic model-based image analysis scheme described in Sections 12.2.1 and 12.2.2. Although tissue quantification and image segmentation may be simultaneously performed [30, 32, 33], a more accurate result can be achieved if the two objectives are considered separately [11, 34]. Guided by the two information theoretic criteria, our algorithm proceeds by fitting an SFNM with model order selection to the histogram of pixel images and then constructing a consistent relaxation labeling of the context images. A summary of the major steps in implementation is

1. For each value of K, perform an ML tissue quantification by applying the EM algorithm (Equations (12.8) and (12.9)).
2. Scan the values of $K = K_{\min}, \ldots, K_{\max}$ by using AIC in Equation (12.13) and MDL Equation (12.15) to determine the suitable number of tissue types K_0.
3. Select the result of tissue quantification corresponding to the value of K_0 determined in step 2.

4. Initialize image segmentation using the ML classification method Equation (12.23).
5. Finalize tissue segmentation by CBRL (implementing Equation (12.24)).

For this study, we use data consisting of three adjacent, T1-weighted MR images parallel to the AC-PC line. Since the skull, scalp, and fat in the original brain images do not contribute to the brain tissue, we edited the MR images to exclude nonbrain structures prior to tissue quantification and segmentation, as explained by Wang and colleagues [8, 9]. This also helps us achieve better quantification and segmentation of brain tissues by delineation of other tissue types that are not clinically significant [30, 31, 34]. The extracted brain tissues are shown in Figure 12.1.



12.1 Three sample MR brain tissues.

Evaluation of different image analysis techniques is a particularly difficult task, and the dependability of evaluations by simple mathematical measures such as squared error performance is questionable. Therefore, most of the time, the quality of the quantified and segmented image usually depends heavily on the subjective and qualitative judgments. Besides the evaluation performed by radiologists, we use the GRE value to reflect the quality of tissue quantification.

The brain is generally composed of three principal tissue types, i.e., white matter (WM), gray matter (GM), cerebrospinal fluid (CSF), and their combinations, called the partial volume effect. We consider the pair-wise combinations as well as the triple mixture tissue, defined as CSF-white-gray (CWG). More importantly, since the MRI scans clearly show the distinctive intensities at local brain areas, the functional areas within a tissue type need to be considered. In particular, the caudate nucleus and putamen are two important local brain functional areas. In our complete image analysis framework, we allow the number of tissue types to vary from slice to slice, i.e., we do consider adaptability to different MR images. We let $K_{\text{min}} = 2$ and $K_{\text{max}} = 9$ and calculate AIC(K) (Equation (12.13)) and MDL(K) (Equation (12.15)) for $K = K_{\text{min}}, \ldots, K_{\text{max}}$. The results with these three criteria all suggest that the three sample brain images chosen contain six, eight, and six tissue types, respectively. According to the model fitting procedure using information theoretic criteria, the minima of these criteria indicate the most appropriate number of tissue types, which is also the number of hidden nodes in the corresponding PSOM (mixture components in SFNM).

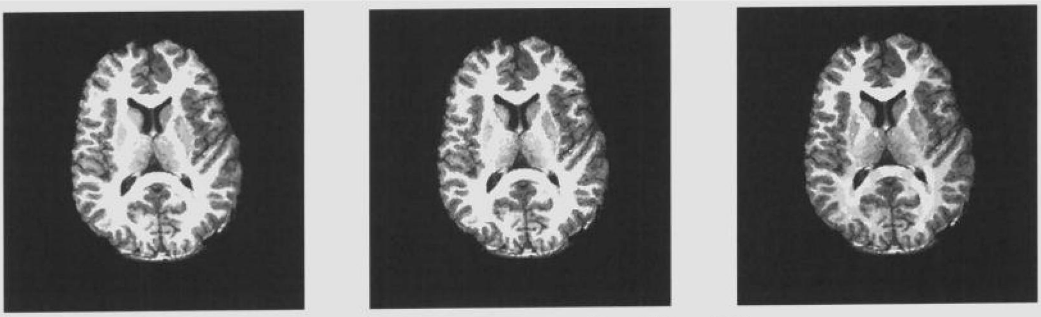
When performing the computation of the information theoretic criteria, we use PSOM to iteratively quantify different tissue types for each fixed K. The PSOM algorithm is initialized by the adaptive Lloyd–Max histogram quantization [24]. For slice 2, the results of final tissue quantification with $K_0 = 7, 8, 9$ are shown in Table 12.1 corresponding to $K_0 = 8$, where a GRE value of 0.02–0.04 nats is achieved. These quantified tissue types agree with those of a physician's qualitative analysis results [11].

The CBRL tissue segmentation for slice 2 is performed with $K_0 = 7, 8, 9$, and the algorithm is initialized by ML classification (Equation (12.23)) [13]. CBRL updates are terminated after five to ten iterations since further iterations produce almost identical results. The segmentation results are

TABLE 12.1 Result of Parameter Estimation for Slice 2

Tissue Type	1	2	3	4	5	6	7	8
π	0.0251	0.0373	0.0512	0.071	0.1046	0.1257	0.2098	0.3752
μ	38.848	58.718	74.400	88.500	97.864	105.706	116.642	140.294
σ^2	78.5747	42.282	56.5608	34.362	24.1167	23.8848	49.7323	96.7227

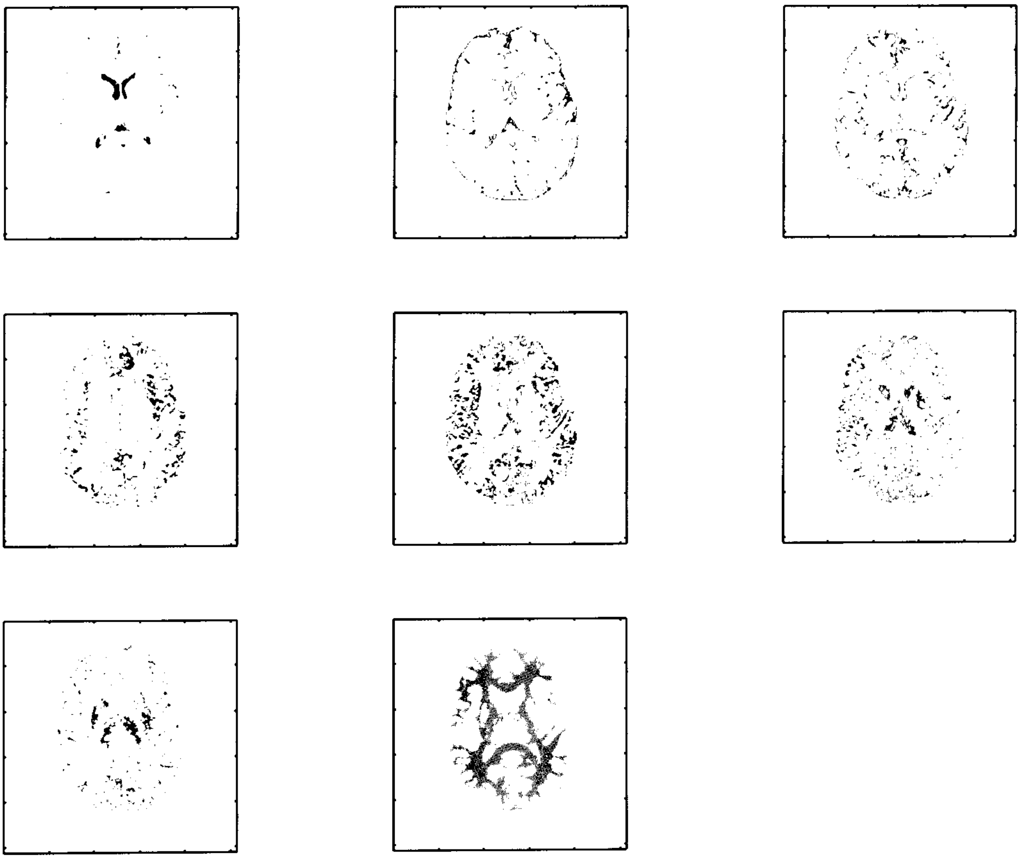
shown in Figure 12.2. It is seen that the boundaries of WM, GM, and CSF are successfully delineated. To see the benefit of using information theoretic criteria in determining the number of tissue types, the decomposed tissue type segments are given in Figure 12.3 with $K_0 = 8$. As can be observed in Figures 12.2 and 12.3, the segmentation with eight tissue types provides a very meaningful result. The regions with different gray levels are satisfactorily segmented; specifically, the major brain tissues are clearly identified. If the number of tissue types were underestimated by one, tissue mixtures located within putamen and caudate areas would be lumped into one component, though the results are still meaningful. When the number of tissue type is overestimated by one, there is no significant difference in the quantification result, but white matter has been divided into two components. For $K_0 = 8$, the segmented regions represent eight types of brain tissues: (a) CSF, (b) CG, (c) CGW, (d) GW, (e) GM, (f) putamen area, (g) caudate area, and (h) WM, as shown in Figure 12.3. These segmented tissue types again agree with the results of radiologists' evaluations [11].



12.2 Results of tissue segmentation for slice 2 with $K_0 = 7, 8, 9$ (from left to right).

Another possible application area for the image analysis framework we introduced is in segmentation and extraction of suspicious mass areas from mammographic images. With an appropriate statistical description of various discriminate characteristics of both true and false candidates from the localized areas, an improved mass detection can be achieved in computer aided diagnosis. Preprocessing is an important step in image analysis for most applications. In this example, one type of morphological operation is derived to enhance disease patterns of suspected masses by cleaning up unrelated background clutters, and then image segmentation is performed to localize the suspected mass areas using a stochastic relaxation labeling scheme [7, 35]. Results are shown in Figure 12.4.

The mammograms for this study were selected from the Mammographic Image Analysis Society (MIAS) database and the Brook Army Medical Center (BAMC) database created by the Department of Radiology at Georgetown University Medical Center. The areas of suspicious masses were identified by an expert radiologist based on visual criteria and biopsy proven results. The BAMC films were digitized with a laser film digitizer (Lumiscan 150) at a pixel size of 100 μm × 100 μm and 4096 gray levels (12 bits). Before the method was applied, the digital mammograms were smoothed by averaging 4 × 4 pixels into one pixel. According to radiologists, the size of small masses is 3–15 mm in effective diameter. A 3 mm object in an original mammogram occupies 30 pixels in a digitized image with a 100 μm resolution. After reducing the image size by four times, the object

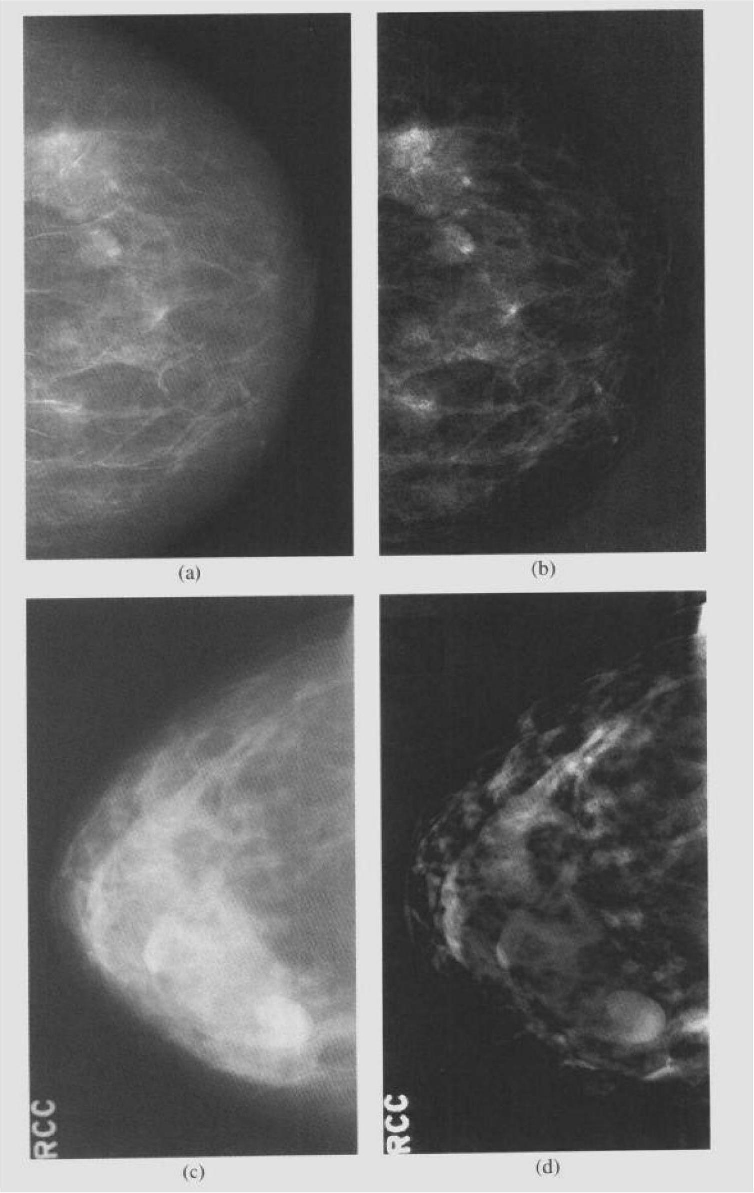


12.3 Results of tissue type decomposition for slice 2. They represent eight types of brain tissues: CSF, CG, CGW, GW, GM, putamen area, caudate area, and WM (left to right, top to bottom).

will occupy the range of about 7 to 8 pixels. The object with the size of 7 pixels is expected to be detectable by any computer algorithm. Therefore, size reduction can be applied for mass detection and can save computation time.

Consider the use of the FGGM model and the two information criteria, AIC and MDL, to determine the mixture number K. Tables 12.2 and 12.3 show the AIC and MDL values with different K and α of the FGGM model based on one sample original mammogram. As can be observed from the tables, even with different values for α, for all cases, AIC and MDL values are minimum when $K = 8$. This indicates that AIC and MDL are relatively insensitive to the change of α. With this observation, we can decouple the relation between K and α and choose the appropriate value of one while fixing the value of the other. Figure 12.5a and b shows two examples of AIC and MDL curves with different K and fixed $\alpha = 3.0$. Figure 12.5a is based on the original, and Figure 12.5b is based on the enhanced mammogram. With the original mammogram, both criteria achieve the minimum when $K = 8$. Figure 12.5b indicates that $K = 4$ is the appropriate choice for the number of meaningful regions for the mammogram enhanced by dual morphological operation, which is reasonable since the numbers of effective regions are expected to decrease after background correction. The results of parameter estimation with $K = 8$ and different values of α are shown in Figure 12.6.

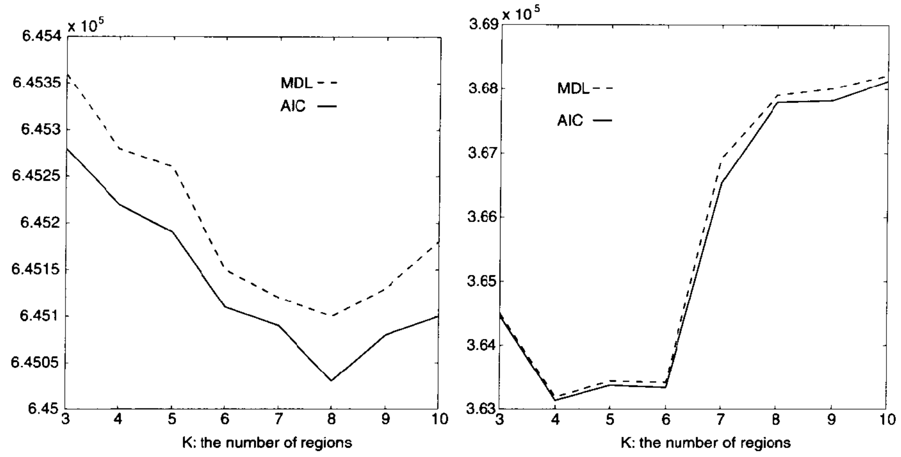
The order of the model is thus fixed at $K = 8$, and the value of α is changed for estimating the FGGM model parameters using the EM algorithm given in Section 12.2.2 with the original mammogram. The GRE value between the histogram and the estimated FGGM distribution is used



12.4 Examples of mass enhancement: (a) and (c) original mammograms; (b) and (d) enhanced mammograms.

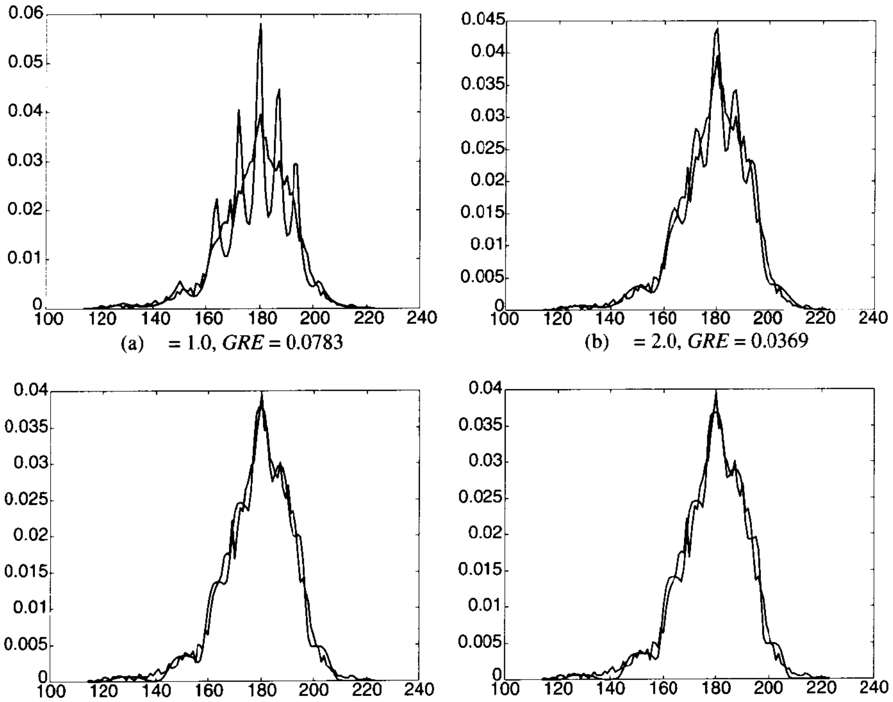
as a measure of the estimation bias, and it is noted that GRE achieves a minimum distance when the FGGM parameter $\alpha = 3.0$, as shown in Figure 12.6. A similar result is obtained when the EM algorithm is applied to the enhanced mammogram with $K = 4$. This indicates that the FGGM model might be better than the SFNM model ($\alpha = 2.0$) in modeling mammographic images when the true statistical properties of mammograms are generally unknown, though the SFNM has been successfully used in a large number of applications, as shown in our previous example. Hence, the choice of the best model to describe the data depends on the nature of the data for the given problem. For details of this experiment, see Li et al. [7].

After determination of all model parameters, every pixel of the image is labeled to one region (from 1 to K) based on the CBRL algorithm. Then, the brightest region, which corresponds to label



(a) (b)

12.5 AIC and MDL curves as functon of the number of regions K: (a) results based on the original mammogram (Optimal $K = 8$); (b) results based on the enhanced mammogram (Optimal $K = 4$).



(a) = 1.0, *GRE* = 0.0783

(b) = 2.0, *GRE* = 0.0369

(c) = 3.0, *GRE* = 0.0251

(d) = 4.0, *GRE* = 0.0282

12.6 The comparison of learning curves and histograms of the original mammogram with different α values for $K = 8$ (Optimal $\alpha = 3.0$).

TABLE 12.2 Computed AIC Values for the FGGM Model with Different α

K	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$	$\alpha = 4.0$
2	651250	650570	650600	650630
3	646220	644770	645280	646200
4	645760	644720	645260	646060
5	645760	644700	645120	646040
6	645740	644670	645110	645990
7	645640	644600	645090	645900
8	645550 (min)	644570 (min)	645030 (min)	645850 (min)
9	645580	644590	645080	645880
10	645620	644600	645100	645910

TABLE 12.3 Computed MDL Values for the FGGM Model with Different α

K	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$	$\alpha = 4.0$
2	651270	650590	650630	650660
3	646260	644810	645360	646350
4	645860	644770	645280	646150
5	645850	644770	645280	646100
6	645790	644750	645150	646090
7	645720	644700	645120	645930
8	645680 (min)	644690 (min)	645100 (min)	645900 (min)
9	645710	644710	645140	645930
10	645790	644750	645180	645960

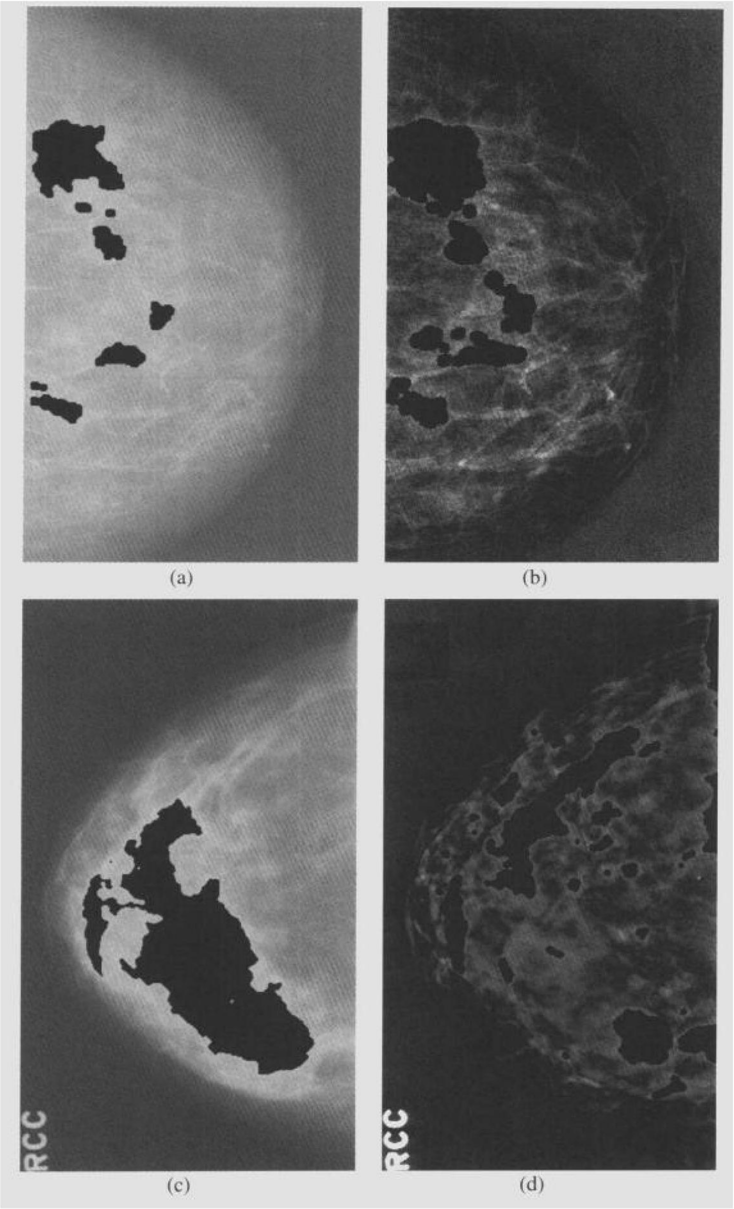
TABLE 12.4 Comparison of Segmentation Error Resulting from Noncontextual and Contextual Methods

Method	Soft Classification	Bayesian Classification	CBRL
GRE Value	0.0067	0.4406	0.1578

K, plus a criterion of closed isolated area, is chosen as the candidate region of suspicious masses, as shown in Figure 12.7. These results are noted to be highly satisfactory when compared to outlines of the lesions [7]. Also, similar to the previous example, GRE values can be used to assess the performance of the final segmentation. Table 12.4 shows our evaluation data from three different segmentation methods when applied to these real images.

12.3 CAD System Design

In order to improve detection and classification in clinical screening and/or diagnosis using radiographic images, many CAD systems have been developed within the last decade. Among others, the development of CAD systems for breast cancer screening and/or diagnosis has received particular attention [36]–[47]. This attention might be attributed to the fact that the role of a CAD system is better defined as complementary to radiologists' clinical duties, i.e., CAD systems can be used for tasks that the radiologists cannot perform well or find difficult to perform. Because of generally larger size and complex appearance of masses, especially the existence of spicules in malignant lesions, as compared to microcalcifications, feature-based approaches are widely adopted in many CAD systems for breast cancer screening [36]–[39], [41, 44]. Kegelmeyer et al. has reported initially promising results for detecting spiculated tumors based on local edge characteristics and Laws texture features [44]. Zwiggelaar et al. developed a statistical model to describe and detect the abnormal pattern of linear structures of spiculated lesions [36]. Karssemeijer and te Brake [37] proposed to identify stellate distortions by using the orientation map of line-like structures, while Petrick et al. showed reduction of false positive detection by combining breast tissue composition information [39]. Zhang et al. used the Hough spectrum to detect spiculated lesions [41].



12.7 Segmentation results for suspected masses based on the original mammogram (a) and (c); (b) and (d) results based on the enhanced mammogram, $K = 4, \alpha = 3.0$.

As is the case in most classification type applications, the first and most critical step in CAD design is the construction of the database (choice of cases to include preprocessing of data, extraction and definition of relevant of features, and labeling of data, i.e., identification of classes). The next step involves design of the classifier, which typically is a problem in a high-dimensional space, requiring use of complex classifier structures due to the inherent complexity of the problem. For this reason, modular network structures that aim at partitioning the problem into simpler sets that are then weighted to form the decisions (e.g., as those discussed in Chapter 5) are particularly attractive solutions for the task. Section 12.3.1 introduces such a modular network and discusses its application to a carefully constructed featured knowledge database of suspicious mass sites for breast cancer screening/detection.

Another approach in CAD systems is to choose a network structure that can extract relevant features of disease patterns internally rather than defining explicit features. The second part of this chapter introduces an example of this approach and shows that convolutional neural networks can be effectively used for this task and applied to CAD design for breast mammography.

12.3.1 Feature-Based Modular Networks for CAD

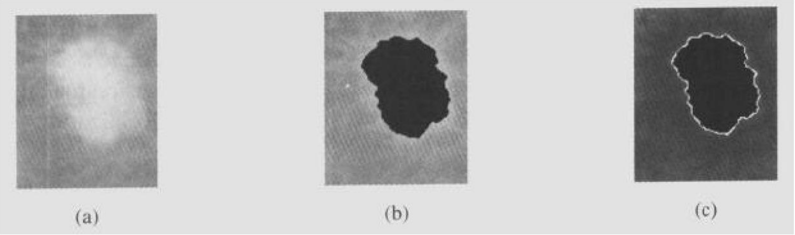
In the following sections we describe the two main steps in the design of a feature-based CAD system: construction of a featured database and the database mapping, i.e., design of the modular neural network for performing the classification task. Even though our concentration in the discussion to follow is on CAD for breast cancer, most of the discussion and methodology is applicable to a variety of CAD design tasks such as electrocardiogram beat classification [42] and detection of malignant melonoma [43].

12.3.1.1 Feature Extraction

Even though feature extraction has been a key step in most pattern analysis tasks, the procedure is often carried out intuitively and heuristically. The general guidelines are

1. Discrimination — features of patterns in different classes should have significantly different values.
2. Reliability — features should have similar values for patterns of the same class.
3. Independence — features should not be strongly correlated to each other.
4. Optimality — some redundant features should be deleted. A small number of features is preferred for reducing the complexity of the classifier. Among a number of approaches for the task, principal component analysis has, by far, been the most widely used approach.

Many useful image features have been suggested by the image processing and pattern analysis communities [3, 48, 49]. These features can be divided into three categories, intensity features, geometric features, and texture features, whose values are calculated from the pixel matrices of the region of interest (ROI). Though these features are mathematically well defined, they may not be complete since they cannot capture all relevant aspects of human perception. Thus, we suggest inclusion of several additional expert-suggested features to reflect the radiologists' experiences. The typical features are summarized in Table 12.5, while Figure 12.8 shows the raw image of the corresponding featured sites.



12.8 One example of mass segmentation and boundary extraction: (a) mass patch; (b) segmentation; and (c) boundary extraction.

TABLE 12.5 Summary of Mathematical Features

Feature Subspace	Features
A. Intensity Features	1. Contrast measure of ROI 2. Standard derivation inside ROI 3. Mean gradient of ROI's boundary
B. Geometric Features	1. Area measure 2. Circularity measure 3. Deviation of the normalized radial length 4. Boundary roughness
C. Texture Features	1. Energy measure 2. Correlation of co-occurrence matrix 3. Inertia of co-occurrence matrix 4. Entropy of co-occurrence matrix 5. Inverse difference moment 6. Sum average 7. Sum entropy 8. Difference entropy 9. Fractal dimension of ROI surface

The joint histogram of the feature point distribution extracted from true and false mass regions is studied, and the features that can better separate the true and false mass regions are selected for further study. Our experience has suggested that three features, the site area, two measures of compactness (circularity), and difference entropy, led to better discrimination and reliability. They are defined as:
1. Compactness 1:

$$C_1 = \frac{A_1}{A} \tag{12.25}$$

where A is the area of the actual suspected region and A_1 is the area of the overlapping region of A and the effective circle A_c, defined as the circle whose area is equal to A and is centered at the corresponding centroid of A.
2. Compactness 2:

$$C_2 = \frac{P}{4\pi A} \tag{12.26}$$

where P is the boundary perimeter and A is the area of region.
3. Difference Entropy:

$$DH_{d,\theta} = -\sum_{k=0}^{L-1} p_{x-y}(k) \log p_{x-y}(k) \tag{12.27}$$

where

$$p_{x-y}(k) = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_{d,\theta}(i,j), \quad |i-j| = k\,. \tag{12.28}$$

Several important observations are worth reiteration:

- The knowledge database that will be used by the CAD system is constructed from the cases selected by both lesion localization and human expert experience. This joint set provides more complete information and, during the interactive decision making, allows the CAD system to provide input when the cases are missed by the localization procedure but presented to the system by the radiologists.
- The knowledge database is defined quantitatively in a high-dimensional feature space. It provides not only the knowledge for training the neural network classifier, but also an objective base for evaluating the quality of feature extraction or a network's learning capability and the online visual explanation possibility.