Computerized Radiographic Mass Detection—Part II: Decision Support by Featured Database Visualization and Modular Neural Networks

Huai Li, Yue Wang, K. J. Ray Liu*, Shih-Chung B. Lo, and Matthew T. Freedman

***Abstract*—Based on the enhanced segmentation of suspicious mass areas, further development of computer-assisted mass detection may be decomposed into three distinctive machine learning tasks: 1) construction of the featured knowledge database; 2) mapping of the classified and/or unclassified data points in the database; and 3) development of an intelligent user interface. A decision support system may then be constructed as a complementary machine observer that should enhance the radiologists performance in mass detection. We adopt a mathematical feature extraction procedure to construct the featured knowledge database from all the suspicious mass sites localized by the enhanced segmentation. The optimal mapping of the data points is then obtained by learning the generalized normal mixtures and decision boundaries, where a is developed to carry out both soft and hard clustering. A visual explanation of the decision making is further invented as a decision support, based on an interactive visualization hierarchy through the probabilistic principal component projections of the knowledge database and the localized optimal displays of the retrieved raw data. A prototype system is developed and pilot tested to demonstrate the applicability of this framework to mammographic mass detection.**

***Index Terms*—Feature extraction, knowledge database, mass detection, neural network, visual explanation.**

I. Introduction

IN ORDER to improve mass lesion detection and classification in clinical screening and/or diagnosis of breast cancers, many sophisticated computer-assisted diagnosis (CAD) systems have been recently developed [1]–[10]. Although the clinical roles of the CAD systems may still be debatable, the fundamental role should be complementary to the radiologists' clinical duties, where the pathways of achieving ultimate performance enhancement taken by the machine observer and human observer may not necessarily be close. For example, CAD systems may attack the 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 with microcalcifications, feature-based approaches are largely adopted in many CAD systems [1]–[4], [6], [7]. Kegelmeyer has first reported promising results for detecting spiculated tumors based on local edge characteristics and Laws texture features [7]. Zwiggelaar *et al.*developed a statistical model to describe and detect the abnormal pattern of linear structures of spiculated lesions [1]. Karssemeijer *et al.* [2] proposed to identify stellate distortions by using the orientation map of line-like structures. Petrick *et al.* presented to reduce the false positive detection by combining the breast tissue composition information [4]. Zhang *et al.* used the Hough spectrum to detect spiculated lesions [6].

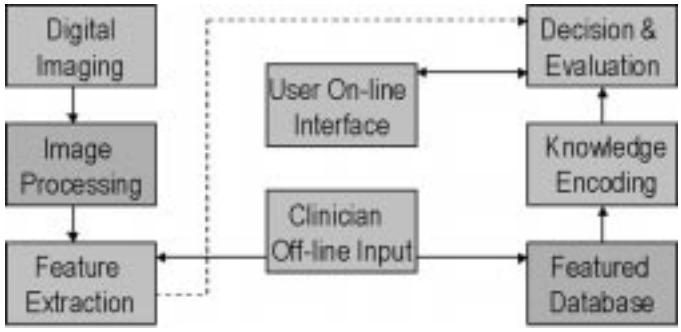


Fig. 1. Major components in CAD.

Although many previously proposed approaches have led to impressive results [1]–[5], [7], several fundamental issues remain unresolved in the application of CAD systems. Fig. 1 shows a general block diagram of CAD systems. Previous research has demonstrated that: 1) breast cancer is missed on mammograms in part because the optical density and contrast of the cancer is not optimal for human observer; 2) computer-based detection appears to be more affected by different criteria than human perception; 3) the challenges and pathways to the human or machine observers may be quite different, and 4) decision making by the CAD systems are largely not transparent to the user. For example, the training cases contributing to the database are often selected by the human observer while the featured knowledge database is constructed through mathematical pathways of feature extraction. The mismatch

Manuscript received February 3, 1997; revised January 9, 2001. This work was supported in part by the Department of Defense under Grants DAMD17-98-1-8045 and DAMD17-96-1-6254 through a subcontract from University of Michigan, Ann Arbor, and by the National Science Foundation (NSF) under NYI Award MIP-9457397. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was M. Giger. *Asterisk indicates corresponding author.*

H. Li is with the Electrical Engineering Department and Institute for Systems Research, University of Maryland at College Park, College Park, MD 20742 USA. He is also with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

Y. Wang is with the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 USA. He is also with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

*K. J. Ray Liu is with the Electrical Engineering Department and Institute for Systems Research University of Maryland at College Park, College Park, MD 20742 USA (e-mail: kjrliu@eng.umd.edu).

S.-C. B. Lo and M. T. Freedman are with the Department of Radiology, Georgetown University Medical Center, Washington, DC 20007 USA.

Publisher Item Identifier S 0278-0062(01)02830-0.

0278–0062/01$10.00 © 2001 IEEE

between the human supervised case selection in training and the machine dominant mass candidates selection in testing may exist. Second, the featured knowledge database is often high-dimensional with complex internal structures. Imposing a heuristically designed neural network for learning from the training data set may prevent a correct identification of the intrinsic data structure and an accurate estimation of the class boundaries. There may also exist the mismatch between the data structure and classifier architecture or between the class boundaries and decision boundaries. Furthermore, since the machine observer and human observer may not detect the same set of masses, the "black box" nature of most CAD systems to the clinical users will prevent a natural on-line integration of human intelligence and further upgrade of a CAD system. An interactive user interface should be considered to leverage the complementary roles of the CAD in the clinical practice.

As a step toward improving the performance of a CAD system, we have put considerable efforts to conduct various studies and develop reliable image enhancement and lesion selection techniques. The methods and results have been reported in [24], where the purposes of the research were to localize the potential mass sites and help accurate feature extraction. This paper addresses the further development of computer-assisted mass detection based on the 1) construction of the featured knowledge database; 2) mapping of the classified and/or unclassified data points in the database; and 3) development of an intelligent user interface (IUI). The clinical goal is to eliminate the false positive sites that correspond to normal dense tissues with *mass-like* appearances through featured discrimination. We adopt a mathematical feature extraction procedure to construct the featured knowledge database from all the suspicious mass sites localized by the enhanced segmentation. The optimal mapping of the data points is then obtained by learning the generalized normal mixtures and decision boundaries, where a probabilistic modular neural network (PMNN) is developed to carry out both soft and hard clustering. A visual explanation of the decision making is further invented as a decision support tool, based on an interactive visualization hierarchy through the probabilistic principal component projections of the knowledge database and the localized optimal displays of the retrieved raw data. The motivation of this work comes from the following considerations. First, though both human and machine observers use the same set of raw data in the diagnostic stage, the construction of the knowledge database for training machine classifiers and that accomplished by human brains are indeed different. Thus, the knowledge database should be established with both machine and expert organized representative cases. Second, a quantitative understanding of the knowledge database used by the machine observer should be acquired to logically compare and/or predict the performance of CAD systems with respect to the human observers without possible under- or over-estimation, and to optimize the feature extraction and design of the machine learner for best final performance. Finally, since the human and machine observers indeed take different learning and intelligence pathways, an IUI should be developed to visually (e.g., transparently) explain the entire internal decision making process of the CAD system to the human observer to enhance the clinical decision when facing either consistent or conflicting opinions.

The major differences between our work and the previous work [1]–[10] are as follows.

1) We construct a knowledge database by combining both expert and machine selected cases where the assignment of class memberships (e.g., mass and nonmass classes) is supervised by the radiologists or pathological report *after* all the cases are collected.
2) We impose a model identification procedure to determine the optimal number and kernel shape of the local clusters within each of the two classes in a high-dimensional feature space. The model is then estimated using the expectation–maximization (EM) algorithm and information theory.
3) We develop a PMNN, which is considered as a nonlinear classifier, to carry out the mapping function of the knowledge database. In the knowledge database, the decision likelihood boundaries and the class prior probabilities are determined in a separate fashion, and the structure of PMNN is optimized by adapting to the database structure.
4) We derive a probabilistic principal component projection scheme to reduce the dimensionality of the feature space for natural human perception. The scheme leads to a hierarchical visualization algorithm allowing the complete data set to be analyzed at the top level, with best separated clusters and subclusters of data points analyzed at deeper levels.

The framework of the proposed method for mass detection is illustrated in Fig. 2. A detailed description of this paper is organized as follows. In Section II, the procedure of the knowledge database construction is described. The data mapping process for decision making is presented in Section III. Section IV presents the design of the IUI for the CAD systems. Finally, major results and discussions are summarized in Section V.

II. Knowledge Database Construction

Given the available information contained in the raw data of mass sites and in order to establish machine intelligence carried out by various machine observers, a knowledge database may be constructed in a multidimensional feature space. It should be emphasized however that the knowledge acquired by the human brain uses much more sophisticated processes than the artificial systems. Though feature extraction has been a key step in most pattern analysis tasks, the mathematical procedures are often done 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 the 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.

Many useful image features have been suggested previously by both image processing and pattern analysis communities [11]–[13]. These features can be divided into three categories, namely, intensity features, geometric features, and texture

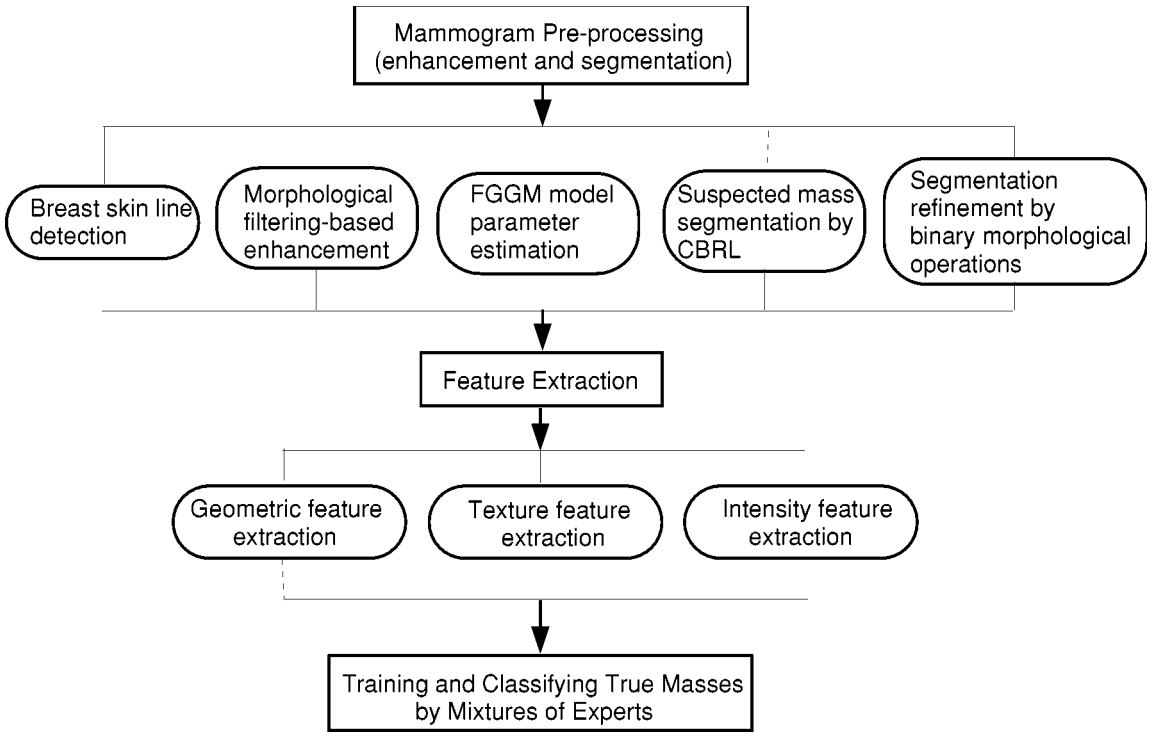


Fig. 2. The flow diagram of mass detection in digital mammograms.

features, whose values are calculated from the pixel matrices of the regions of interest (ROIs). Though these features are mathematically well defined, they may not be complete since they cannot capture all of the capable aspects of human perception nature. Thus, in this study, we have included several additional expert-suggested features to reflect the radiologists' experience. The typical features are summarized in Table I, where Fig. 3 shows the raw image of corresponding featured sites.

The joint histogram of the feature point distribution extracted from true and false mass regions are investigated, 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, i.e., the site area, two measured compactness (circularity), and difference entropy, were having better discrimination and reliability properties. Their definitions are given as follows.

1) *Compactness 1*

$$C_1 = \frac{A_1}{A} \tag{1}$$

where A is the area of the actual suspected region, and A_1 is the area of the overlapped region of A and the effective circle A_c, which is defined as the circle whose area is equal to A and is centered about the corresponding centroid of A.

2) *Compactness 2*

$$C_2 = \frac{P}{4\pi A} \tag{2}$$

where P is the boundary perimeter, and A is the area of region.

TABLE I
THE SUMMARY OF MATHEMATICAL FEATURES

Feature Sub-Space	Features
A. Intensity Features	1. contrast measure of ROIs; 2. standard derivation inside ROIs; 3. mean gradient of ROIs 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 surface of ROI;

3) *Difference Entropy*

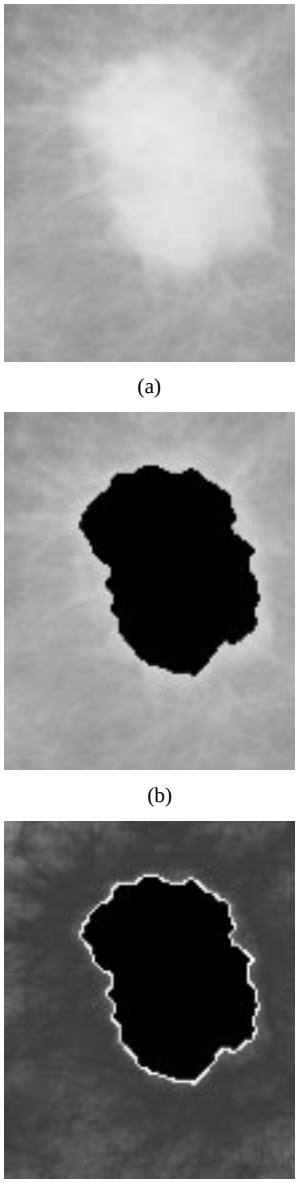
$$DH_{d,\theta} = -\sum_{k=0}^{L-1} p_{x-y}(k) \log p_{x-y}(k) \tag{3}$$

where

$$p_{x-y}(k) = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} p_{d,\theta}(i,j), \quad |i-j| = k. \tag{4}$$

Several important observations are worth reiteration:

1) The knowledge database that will be used by the CAD system are constructed from the cases selected by both lesion localization procedure and human expert's experience. This joint set provides more complete knowledge to



(a) (b) (c)

Fig. 3. One example of mass segmentation and boundary extraction. (a) Mass patch; (b) segmentation; (c) boundary extraction.

the machine observer. In particular, during the interactive decision making, CAD system can still provide opinion when the cases are missed by the localization procedure but presented to the system by the radiologists.

2) The knowledge database is defined quantitatively in a high dimensional feature space. It provides not only the knowledge for training the machine observer, but also an objective base for evaluating the quality of feature extraction or network's learning capability, and the on-line visual explanation possibility.
3) The assignment of the cases' class memberships (e.g., mass and nonmass classes) is supervised by the radiologists or pathological reports. A complete knowledge database includes three subsets: raw data of mass-like sites, corresponding feature points, and class membership labels.

III. Data Mapping For Decision Making

The decision making support by a CAD system addresses the problem of mapping a knowledge database, given a finite set of data examples. The mapping function can therefore be interpreted as a quantitative representation of the knowledge about the mass lesions contained in the database [14]. Instead of mapping the whole data set using a single complex network, it is more practical to design a set of simple class subnets with local mixture clusters, each one of which represents a specific region of the knowledge space. Inspired by the principle of *divide-and-conquer* in applied statistics, PMNN has become increasingly popular in machine learning research [14], [15], [19]–[22]. In this section, we present its applications to the problem of mapping from databases in mass detection, with a constructive criterion for designing the network architecture and the learning algorithm that are governed by information theory [25].

A. Statistical Modeling

The quantitative mapping of a database may be decomposed into three distinctive learning tasks: the detection of the structure of each class model with local mixture clusters; the estimation of the data distributions for each induced cluster inside each class; and the classification of the data into classes that realizes the data memberships. Recently, there has been considerable success in using finite mixture distributions data mapping [15], [17], [18], [20]. Assume that the data points $\vec{x}_i$ in a multidimensional database come from M classes $\{\vec{\omega}_1, \ldots, \vec{\omega}_r, \ldots, \vec{\omega}_M\}$, and each class contains K_r clusters $\{\vec{\theta}_1, \ldots, \vec{\theta}_k, \ldots, \vec{\theta}_{K_r}\}$, where $\vec{\omega}_r$ is the model parameter vector of class r, and $\vec{\theta}_k$ is the kernel parameter vector of cluster k within class r. The class conditional probability measure for any data point inside the class r, i.e., the standard finite mixture distribution (SFMD), can be obtained as a sum of the following general form:

$$f(\vec{u}|\vec{\omega}_r) = \sum_{k=1}^{K_r} \pi_k g(\vec{u}|\vec{\theta}_k) \tag{5}$$

where $\pi_k = P(\vec{\theta}_k|\vec{\omega}_r)$ with a summation equal to one, and $g(\vec{u}|\vec{\theta}_k)$ is the kernel function of the local cluster distribution. For the model of global class distributions, we denote the Bayesian prior for each class by $P(\vec{\omega}_r)$. Then the sufficient statistics according to the Bayes' rule, are the posterior probability $P(\vec{\omega}_r|\vec{x}_i)$ given a particular observation $\vec{x}_i$

$$P(\vec{\omega}_r|\vec{x}_i) = \frac{P(\vec{\omega}_r) f(\vec{x}_i|\vec{\omega}_r)}{p(\vec{x}_i)} \tag{6}$$

where $p(\vec{x}_i) = \sum_{r=1}^{M} P(\vec{\omega}_r) f(\vec{x}_i|\vec{\omega}_r)$.

B. Class Distribution Learning

Class distribution learning addresses the combined estimation of regional parameters $(\pi_k, \vec{\theta}_k)$ and detection of the structural parameter K_r and the kernel shape of $g(\cdot)$ in (5) based on the observations $\mathbf{x}_r$. One natural criterion used for learning the optimal parameter values is to minimize the distance between the SFMD, denoted by $f_r(\vec{u})$, and the class data histogram, denoted by $f_{\mathbf{x}_r}(\vec{u})$ [17]. In this paper, we use relative entropy (Kullback–Leibler distance), suggested by information theory

[25], as the distance measure (for simplicity we use $f_r(\vec{u})$ to denote $f(\vec{u}|\vec{\omega}_r)$ in our formulation), given by

$$D(f_{\mathbf{x}_r}\|f_r)=\sum_{\vec{u}} f_{\mathbf{x}_r}(\vec{u})\log\frac{f_{\mathbf{x}_r}(\vec{u})}{f(\vec{u}|\vec{\omega}_r)}. \tag{7}$$

We have previously shown that when relative entropy is used as a distance measure, the distance minimization method is equivalent to the soft-split classification-based method under the criterion of maximum likelihood (ML) [23].

Another important issue concerning unsupervised distribution learning is the detection of the structural parameters of the class distribution, called model selection [15]. The objective here is to propose a systematic strategy for determining the optimal number and kernel shape of local clusters, when the prior knowledge is not available. This is indeed the case when the structure of the mass lesion patterns for a particular type of cancer may be arbitrarily complex, so correct identification of the database structure is very important. Thus, it will be desirable to have a neural network structure that is adaptive, in the sense that the number and kernel shape of local clusters are not fixed beforehand. In this paper, we applied two popular information theoretic criteria, i.e., the Akaike information criterion and minimum description length to guide the model selection procedure [24].

As the counterpart for adaptive model selection, there are many numerical techniques to perform ML estimation of cluster parameters [17]. For example, EM algorithm first calculates the posterior Bayesian probabilities of the data through the observations and the current parameter estimates (E-step) and then updates parameter estimates using generalized mean ergodic theorems (M-step). The procedure cycles back and forth between these two steps. The successive iterations increase the likelihood of the model parameters. The scheme provides winner-takes-in probability (Bayesian "soft") splits of the data, hence allowing the data to contribute simultaneously to multiple clusters. For the sake of simplicity, we assume the kernel shape of local clusters to be a multidimensional Gaussian with mean $\vec{\mu}_{kr}$ and variance Γ_{kr}. We summarize the EM algorithm as follows.

1) **E-Step:** for training sample $\vec{x}^{(t)}, t=1,\ldots,N$, compute the probabilistic membership

$$h_{kr}^{(m)}(t)=\frac{\pi_{kr}^{(m)}p_k^{(m)}(\vec{x}^{(t)}|\vec{\omega}_r)}{\sum_{k=1}^{K_r}\pi_{kr}^{(m)}p_k^{(m)}(\vec{x}^{(t)}|\vec{\omega}_r)}. \tag{8}$$

2) **M-Step:** compute the updated parameter estimates

$$\pi_{kr}^{(m+1)}=\frac{1}{N}\sum_{t=1}^{N}h_{kr}^{(m)}(t) \tag{9}$$

$$\vec{\mu}_{kr}^{(m+1)}=\frac{1}{N\pi_{kr}^{(m+1)}}\sum_{t=1}^{N}h_{kr}^{(m)}(t)\vec{x}^{(t)} \tag{10}$$

$$\Gamma_{kr}^{(m+1)}=\frac{1}{N\pi_{kr}^{(m+1)}}\sum_{t=1}^{N}h_{kr}^{(m)}(t)\left[\vec{x}^{(t)}-\vec{\mu}_{kr}^{(m+1)}\right]\times\left[\vec{x}^{(t)}-\vec{\mu}_{kr}^{(m+1)}\right]^T. \tag{11}$$

C. Decision Boundary Learning

The objective of data classification is to realize the class membership l_{ir} for each data points based on the observation $\vec{x}_i$ and the class statistics $\{P(\vec{\omega}_r),\ f(\vec{u}|\vec{\omega}_r)\}$. It is well known that the optimal data classifier is the Bayes classifier since it can achieve the minimum rate of classification error [26]. Measuring the average classification error by the mean squared error E, many previous researchers have shown that minimizing E by adjusting the parameters of class statistics is equivalent to directly approximating the posterior class probabilities when dealing with the two class problem [13], [26]. In general, for the multiple class problem the optimal Bayes classifier (minimum average error) classifies input patterns based on their posterior probabilities: input $\vec{x}_i$ is classified to class $\vec{\omega}_r$ if

$$P(\vec{\omega}_r|\vec{x}_i)>P(\vec{\omega}_j|\vec{x}_i) \tag{12}$$

for all $j\neq r$. It should be noted that in the formulation of classifier design, the optimal criterion used for the future data classification has been intuitively and directly applied to the learning of class statistics from the training data set.

Direct learning of posterior probability is a complex task. Great effort has been made in designing the classifier as an estimator of the posterior class probability [19]. By closely investigating the global class distribution modeling, we found that the classifier design for data classification can be dramatically simplified at the learning stage. Revisit (6), since the class prior probability $P(\vec{\omega}_r)$ is a known parameter when a supervised learning is applied, the posterior class probability $P(\vec{\omega}_r|\vec{x}_i)$ can be obtained without any further effort. Thus, by conditioning $P(\vec{\omega}_r)$, the problem is formulated as a supervised classification learning of the class conditional likelihood density $f(\vec{u}|\vec{\omega}_r)$. Thus, an efficient supervised algorithm to learn the class conditional likelihood densities called the "decision-based learning" [21] is adopted in this paper. The decision-based learning algorithm uses the *misclassified* data to adjust the density functions $f(\vec{u}|\vec{\omega}_r)$, which are initially obtained using the unsupervised learning scheme described previously, so that the minimum classification error can be achieved. Define the rth class discriminant function $\phi_r(\vec{x}_i,\mathbf{w})$ to be $P(\vec{\omega}_r)f(\vec{x}_i|\vec{\omega}_r)$. Given a set of training patterns $\mathbf{X}=\{\vec{x}_i;\ i=1,2,\ldots,M\}$. The set $\mathbf{X}$ is further divided into the "positive training set" $\mathbf{X}^+=\{\vec{x}_i;\ \vec{x}_i\in\vec{\omega}_r,\ i=1,2,\ldots,N\}$ and the "negative training set" $\mathbf{X}^-=\{\vec{x}_i;\ \vec{x}_i\notin\vec{\omega}_r,\ i=N+1,N+2,\ldots,M\}$. If the misclassified training pattern is from positive training set, reinforced learning will be applied. If the training pattern belongs to the negative training set, we anti-reinforce the learning, i.e., pull the kernels away from the problematic regions. The boundary refinement is summarized as follows:

$$\begin{aligned}&\text{Reinforced}\\&\text{Learning:}\quad \mathbf{w}^{(j+1)}=\mathbf{w}^{(j)}+\eta l'(d(t))\nabla\phi(\mathbf{x}(t),\mathbf{w})\\&\text{Antireinforced}\\&\text{Learning:}\quad \mathbf{w}^{(j+1)}=\mathbf{w}^{(j)}-\eta l'(d(t))\nabla\phi(\mathbf{x}(t),\mathbf{w})\end{aligned}. \tag{13}$$

PMNN is a probabilistic modular network designed especially for data classification where a Bayesian decomposition of the learning process provides a unique opportunity to optimize

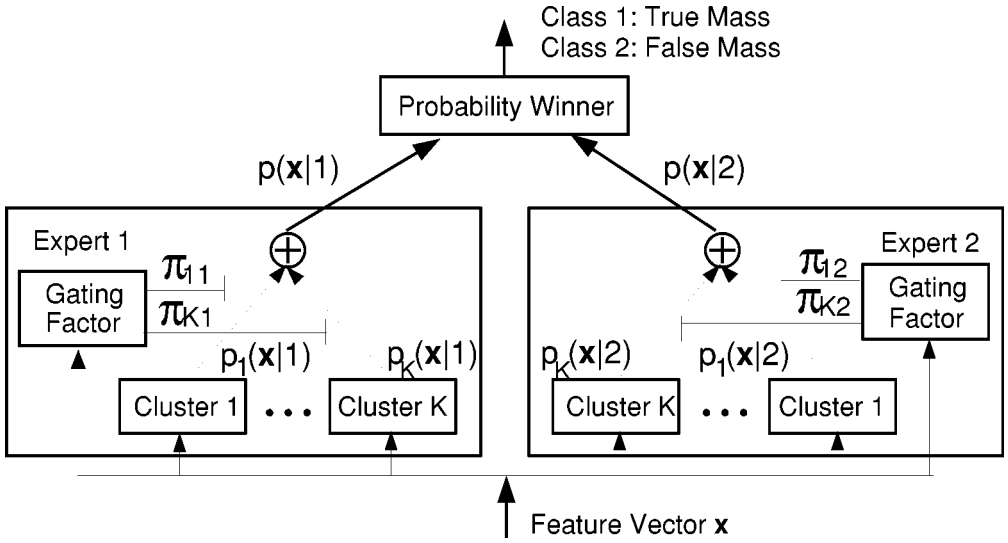


Fig. 4. The structure of the PMNN.

the structure of training scheme [14], [22]. Since the information about class population is, in general, physically uncorrelated with the conditional features about the individual class, a decoupled two-step training, in terms of both network structure and learning rule, makes much more sense than that in the conventional posterior-type neural networks, i.e., the conditional likelihood of each class and the class Bayesian prior should be adjusted separately in the classification spaces. Thus, PMNN consists of several disjoint subnets and a winner-takes-all network. The subnet outputs of the PMNN are designed to model the likelihood functions (likelihood-type network) which are first estimated from equally presented class samples, and the final decision boundaries are determined simply weighting the likelihood by the class populations. For a M-classification problem, PMNN contains M different class subnets, each of which represents one data class in the database. Within each subnet, several neurons (or clusters) are applied in order to handle problems which have complicated decision boundaries. The outputs of class subnets are fed into a winner-take-all network. The winner-take-all network categorizes the input pattern to the data class whose subnet produces the highest output value.

The structure of the PMNN used in this study is shown in Fig. 4. The PMNN consists of two subnets. Within each subnet, there are several neurons (or clusters). The outputs of class subnets are fed into a probability winner processor, which categorizes the input pattern to the data class whose subnet produces the highest probability value. The training scheme of the PMNN is based on the unsupervised learning. Each subnet is trained individually, and no mutual information across the classes may be utilized. In our study, one modular expert is trained to detect true masses, and the other is trained to detect false masses. After training, the feature vectors extracted from ROIsub are entered to this network to classify true or false masses. In both training and testing processes, we assume that the feature vectors $\vec{x}_i$ in class r ($r = 1, \ldots, M$) is a mixture of multidimensional Gaussian distributions, i.e.,

$$f(\vec{x}_i|\vec{\omega}_r) = \sum_{k=1}^{K_r} \pi_{kr} p_k(\vec{x}_i|\vec{\omega}_r) \tag{14}$$

where $\sum_{k=1}^{K_r} \pi_{kr} = 1$ and $p_k(\vec{\omega}_r) = N(\vec{\mu}_{kr}, \Gamma_{kr})$ is a multidimensional Gaussian distribution within cluster k of class r.

IV. Interactive Visual Explanation

In order to improve the utility of the CAD systems in clinical practice, an IUI is highly desired. Different from many previously proposed approaches, we have organized our database from both mathematical-localized and radiologist-selected mass-like cases, and formed the featured knowledge database based on both mathematical-based and radiologist-selected image features. This off-line effort should enhance the performance of the machine observer through better quality of training set and optimal design of neural network architecture. Our experience has suggested, however, that further improvement of CAD systems requires on-line natural integration of human intelligence with the computer' output, since human perception has and can play an important role in the clinical decision making. In this research, we have pilot developed an IUI where the major functions include: 1) interactive visual explanation of the CAD decision making process; 2) on-line retrieval of the optimally displayed raw data and/or similar cases; and 3) supervised upgrade of the knowledge database by radiologist-driven input of the "unseen" and/or "typical" cases. Our preliminary studies have shown that the visual presentation of both raw data and CAD results to radiologists may provide visual cues for improved decision making.

As a step toward understanding the complex information from data and relationships, structural and discriminative knowledge reveals insight that may prove useful in data mining. Hierarchical minimax entropy modeling and probabilistic principal component projection are proposed for data explanation, which is both statistically principled and visually effective at revealing all of the interesting aspects of the data set. The methods involve multiple use of standard finite normal mixture models and probabilistic principal component projections. The strategy is that the top-level model and projection should explain the entire data set, best revealing the presence of clusters and relationships, while lower-level models and

projections should display internal structure within individual clusters, such as the presence of subclusters and attribute trends, which might not be apparent in the higher-level models and projections. With many complementary mixture models and visualization projections, each level will be relatively simple while the complete hierarchy maintains overall flexibility yet still conveys considerable structural information. In particular, a probabilistic principal component neural network is developed to generate optimal projections, leading to a hierarchical visualization algorithm. This algorithm allows the complete data set to be analyzed at the top level, with best separated subclusters of data points analyzed at deeper levels.

Research evidence suggests that for analysis of complex and high-dimensional data sets, structure decomposition and dimensionality reduction are the natural strategies in which the model-based approach and visual explanation have proven to be powerful and widely-applicable [27]. However, there is a trade-off between maximizing (structure decomposition) and minimizing (dimensionality reduction) the entropy of the system. In this research, a minimax entropy approach is adopted through the use of progressive model identification and principal component projection. The complete visual explanation hierarchy is generated by performing principal projection (dimensionality reduction) and model identification (structure decomposition) in two iterative steps using information theoretic criteria, EM algorithm, and probabilistic principal component analysis (PCA). Hierarchical probabilistic principal component visualization involves: 1) evaluation of posterior probabilities for mixture data set; 2) estimation of multiple principal component axes from probabilistic data set; and 3) generation of a complete hierarchy of visual projections.

Suppose the data space is d-dimensional with coordinates $y_1, \ldots, y_d$ and the data set consists of a set of d-dimensional vectors $\{\mathbf{t}_i\}$ where $i = 1, \ldots, N$. Now consider a three-dimensional (3-D) latent space $\mathbf{x} = (x_1, x_2, x_3)^T$ together with a linear function which maps the latent space to the data space by $\mathbf{y} = \mathbf{W}\mathbf{x}+\mathbf{b}$ where $\mathbf{W}$ is a $d\times 3$ matrix and $\mathbf{b}$ is a d-dimensional mean vector. If we introduce a probability distribution $p(\mathbf{x})$ over the latent space given by a Gaussian estimated from the latent variables $\{\mathbf{x}_i\}$, then a similar full-dimensional Gaussian distribution in data space can be defined by convolving this distribution with a general diagonal Gaussian conditional probability distribution $p(\mathbf{t}|\mathbf{x}, \Lambda_d)$ in data space where Λ_d is the covariance matrix, resulting in a final form of

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{15}$$

where the log likelihood function for this model is given by $L = \sum_i \log p(\mathbf{t}_i)$. Suppose $\mathbf{W}$ is determined by the PCA, ML can be used to fit the model to the data and hence determine values for the parameters $\mathbf{b}$ and Λ_d [27]. Using a soft clustering of the data set and multiple PCAsub corresponding to the clusters, a *mixture* of latent models takes the form of $p(\mathbf{t}) = \sum_{k=1}^{K_0} \pi_k p(\mathbf{t}|k)$ where K_0 is the number of components in the mixture, and the parameters π_k are the prior probabilities corresponding to the components $p(\mathbf{t}|k)$. Each component is an independent latent model with PCA projection $\mathbf{W}_k$ and parameters $\mathbf{b}_k$ and Λ_{dk}. This procedure can be further extended to a hierarchical mixture model formulated by $p(\mathbf{t}) = \sum_{k=1}^{K_0} \pi_k \sum_j \pi_{j|k} p(\mathbf{t}|k, j)$

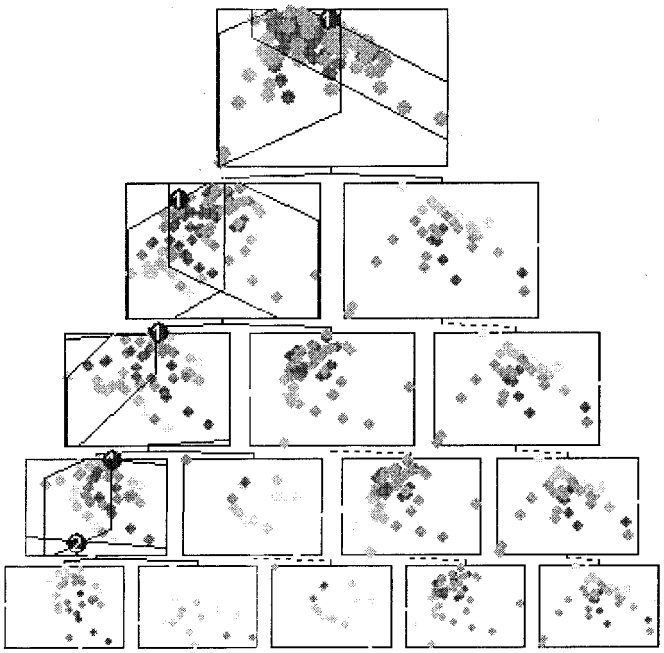


Fig. 5. The hierarchical view of computed features for mass and nonmass samples (Database A, see Table II).

where $p(\mathbf{t}|k, j)$ again represent independent latent models [27]. With a soft partitioning of the data set via EM algorithm, data points will effectively belong to more than one cluster at any given level. This step is automatically available in our approach since the estimation of parent latent model involves the calculation of posterior probabilities denoted by z_{ik}. Thus, the effective input values are $z_{ik}\mathbf{x}_i$ for an independent visualization space k, corresponding to the visualization space k in the hierarchy. It should be emphasized that *probabilistic* means both neural network based learning and posterior probability weighted inputs. Further projections can again be performed by using the effective input values $z_{ik}z_{j|k}\mathbf{t}_i$ for the visualization subspace j. Fig. 5 shows the hierarchical view of computed features for mass and nonmass samples. In Fig. 5, a hierarchical visualization view of a high dimensional feature data set was generated using hierarchical data visualization algorithm. One hundred and 25 real cases were involved, among them 75 are mass sites, 50 are nonmass sites. Nine features were computed on 125 cases. The dimension of the resulted feature data set became 125 $\times$ 9 (Database A, see Table II). Hierarchical visualization tool enables the visualization of high dimensional data set through dimension reduction and data modeling so that data distribution features of the data set can be well recognized. For instance, the clusters and subclusters of mass and nonmass data points and the boundaries of the clusters can be revealed for further research purpose.

In the use of a hierarchical minimax entropy mixture model, an interactive visualization environment is required to enable a flexible computerized experiment such that a human-database interaction can be performed effectively. We have developed an interactive environment for visualizing five-dimensional (5-D) data sets, based on state-of-the-art computer graphics toolkits such as object-oriented OpenGL and OpenInventor. With a sophisticated set of various kinds of simulated lights, color

TABLE II
THE SUMMARY OF EXPERIMENTAL DATABASES

Database	Descriptions
A	Nine features extracted from 75 mass sites and 50 non-mass sites. Used for visualizing hierarchically projected high dimensional feature space. Result is presented in Figure 5.
B	A simulated two-dimensional feature space. Used to show the effect of model selection on decision boundary estimation. Result is shown in Figure 6.
C	ORL standard database. Used to show the improvement of PMNN with decision-based learning. Result is discussed in the text.
D	The training data set consisting of 50 mammograms, with 50 true mass sites and 50 false mass sites. Three most discriminatory features are extracted. Used for both PMNN training and visualization. Result is given in Figure 7.
E	The testing data set consisting of 46 mammograms, with 23 normal cases and 23 biopsy proven mass cases with each of them having at least one true mass site. Three most discriminatory features, the same as database D, are extracted. Used to test the overall performance of our CAD system prototype where the mass candidates were selected using the method reported in Part I, automatically. Result is shown in Figure 8 and also discussed in the text.

texturing editors, and 3-D manipulator and viewers (we have integrated 3-D mouse and stereo glass units into our existing system), our system allows one to examine the volumetric data sets with any viewpoint and dynamically walk through its internal structures to better understand the spatial relationships among clusters and decision surfaces present. One of the most important features in our approach is to attach the decision surface to the 3-D probability cloud in support of decision making, and to link each data point in the visualization space to its raw data so that the user can on-line retrieve the corresponding raw data such as an original image for interim decision making.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we present the experimental results using the information theoretic criteria and PMNNs to generate the mapping function of the featured database, and the preliminary results using the hierarchical minimax entropy projections to conduct visual explanation of the decision making. For the validation of the database mapping using the proposed algorithms, global relative entropy (GRE) value between the (SFMD)and the joint histogram is used as an objective measure to evaluate the fitness of the mapping function. A summary of the databases we used in our study is presented in Table II.

As we have discussed in Sections III and IV, model selection is the first and a very important learning task in mapping a database and the objective of the procedure is to determine both the number and the kernel shape of local clusters in each class. This procedure is used not only in the data mapping for decision making but also in the structure decomposition for hierarchical visual explanation. Our experience has suggested that an incorrect model selection will affect the performance of data-classification based decision making. For the sake of simplicity, we discuss this conclusion in the following 2-D example. Let us form a simulated featured database with two major features that well characterize the two targeted classes, as it shown in Fig. 6 (Database B, see Table II). The ground truth is that class 1 contains only one local cluster while class 2 contains two local clusters. With a model selection procedure using the proposed criteria, the intrinsic data structure was correctly identified. According to the principle of designing the optimal structure of PMNN and visual explanation hierarchy, the result of these criteria also determines the most appropriate number of mixture components in the corresponding PMNN and projected cluster decomposition. Two PMNN with different architecture orders were designed and trained to determine the classification boundaries between the two classes. The classification results are shown in Fig. 6(a) and (b). The result in Fig. 6(a) is with the right cluster number in Class 2, while the result in Fig. 6(b) is with the wrong cluster number in Class 2. From this simple experiment, we have shown that the decision boundary with the right cluster number may be much more accurate than that with heuristically determined cluster number, since the decision boundary between class 1 and class 2 will be determined by four cross points in the first case while in the second case the decision boundary will be determined by only two cross points. It should be emphasized that the error of data classification is theoretically controlled by the accuracy in estimating the decision boundaries between classes, and the quality of the boundary estimates is indeed dependent upon the correct structure of the class likelihood function.

As we have discussed before, although the knowledge database contains both machine-localized and human-selected cases, in clinical settings "unseen" and/or subtle cases contribute the major false positives. We have also pilot tested the PMNN method to the so-called "$M + 1$ classes" problem, in which the disease pattern under testing could be either from one of the M classes, or from some other unknown classes (the "unknown" class or the "intruder" class). Note that the unknown class probability is often very hard to estimate because of the lack of sufficient training samples (for example, in the mass detection problem, the unknown classes include the ROIsub over the normal tissues). In our experiment, PMNN uses different decision rule from that of the "M classes" problem: pattern $\vec{x}_i$ belongs to class r if both of the following conditions are true: a) $\phi(\vec{\omega}_r, \vec{x}_i) > \phi(\vec{\omega}_j, \vec{x}_i),\ \forall j \neq r$, and b) $\phi(\vec{\omega}_r, \vec{x}_i) > T$. T is a threshold obtained by decision-based

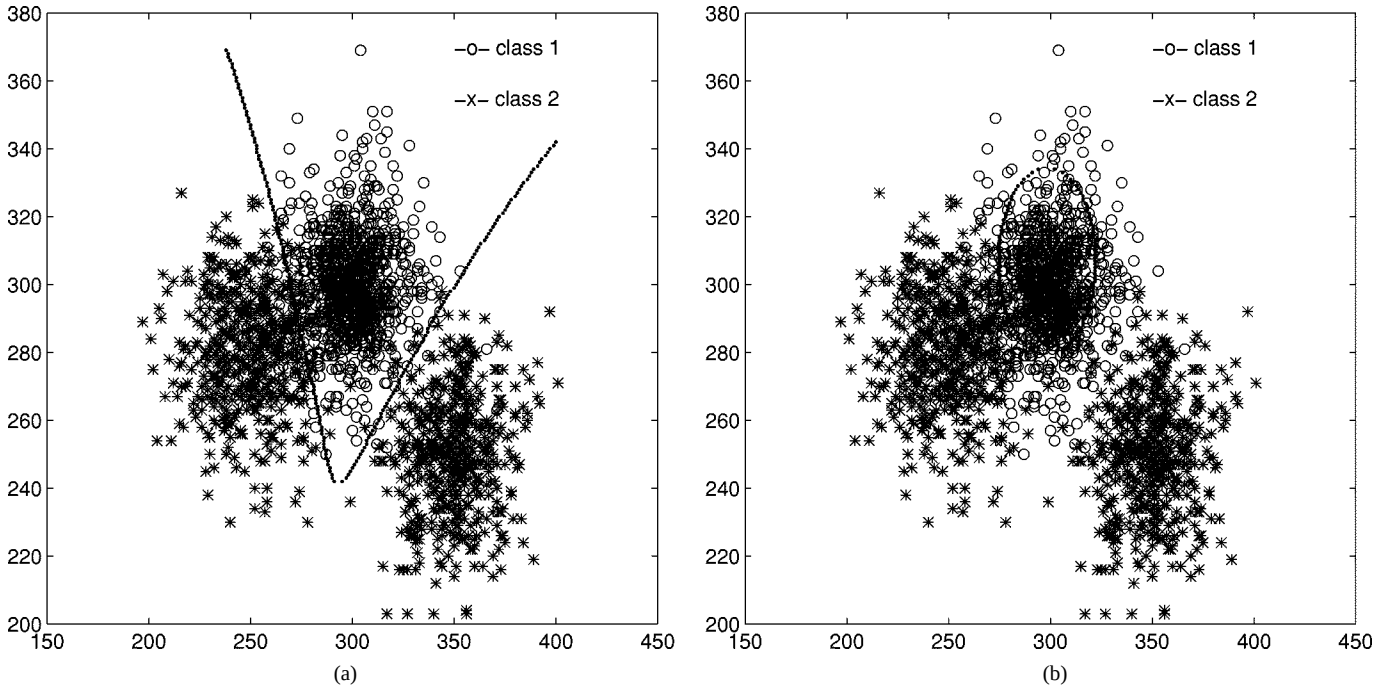
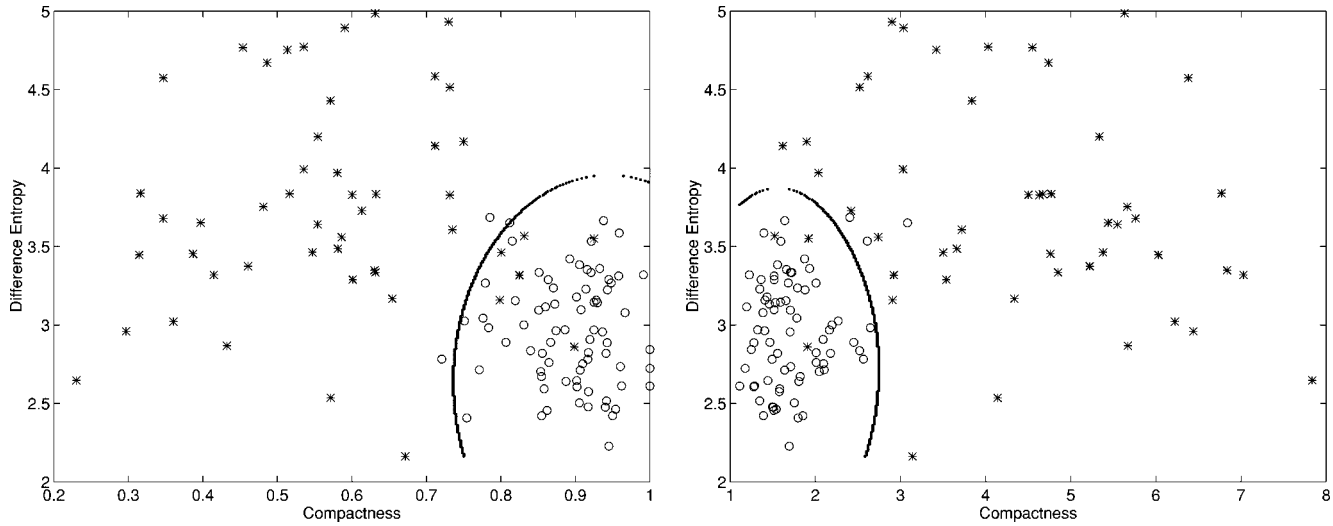


Fig. 6. The classification examples with a two-dimensional (2-D) simulated database (Dabase B, see Table II). (a) Class 2 contains two local clusters. (b) Class 2 contains one local cluster.

learning. Otherwise pattern $\vec{x}_i$ belongs to the unknown class. We observed consistent and significant improvement in classification results compared with the pure Bayesian decision. Using the ORL (Olivetti Research Laboratory, Cambridge, U.K.) standard database (Database C, see Table II), our experience has shown an increase of correct detection rate from 70% to 90% [14].

In the third experiment, we use the proposed classifier to distinguish true masses from false masses based on the features extracted from the suspected regions. The objective is to reduce the number of suspicious regions and identify the true masses. 150 mammograms, each of them contains at least one mass case of varying size and location, were selected in our study. The areas of suspicious masses were identified following the proposed procedure with biopsy proven results. Fifty mammograms with biopsy proven masses were selected from the 150 mammograms for training (Database D, see Table II). The mammogram set used for testing contained 46 single-view mammograms: 23 normal cases and 23 with biopsy proven masses (Database E, see Table II) which were also selected from the 150 mammograms. All mammograms were digitized with an image resolution of 100 μm $\times$ 100 μm/pixel by the laser film digitizer (Model: Lumiscan 150). The image sizes are 1792 $\times$ 2560 $\times$ 12 bpp. For this study, we shrunk the digital mammograms with the resolution of 400 μm by averaging 4 $\times$ 4 pixels into one pixel. According to radiologists, the size of the small masses is 3–15 mm. The middle size of masses is 15–30 mm. The large size of masses is 30–50 mm, which are rare in mammograms. 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 will occupy the range of about seven to eight pixels. The object with the size of seven pixels is expected to be detectable by any computer algorithm. Therefore, the shrinking step is applicable for mass cases and can save computation time.

After the segmentation, the area index feature was first used to eliminate the nonmass regions. In our study, we set $A_1 = 7 \times 7$ pixels and $A_2 = 75 \times 75$ pixels as the thresholds. A_1 corresponds to the smallest size of masses (3 mm), and an object with a area of 75 $\times$ 75 pixels corresponds to 30 mm in the original mammogram. This indicates that the scheme can detect all masses with sizes up to 30 mm. Masses larger than 30 mm are rare cases in the clinical setting. When the segmented region satisfied the condition $A_1 \leq A \leq A_2$, the region was considered to be suspicious for mass. For the purpose of representative demonstration, we have selected a 3-D feature space consisting of compactness I, compactness II, and difference entropy. According to our investigation, these three features have the better separation (discrimination) between the true and false mass classes. It should be noticed that the feature vector can easily extend to higher dimensionality. A training feature vector set was constructed from 50 true mass ROIsub and 50 false mass ROIsub (Database D, see Table II). The training set was used to train two modular probabilistic decision-based neural networks separately. In addition to the decision boundaries recommended by the computer algorithms, a visual explanation interface has also been integrated with 3-D to 2-D hierarchical projections. Fig. 7(a) shows the database map projection with compactness definition I and difference entropy. Fig. 7(b) shows the database map projection with compactness definition II and difference entropy. Our experience has suggested that the recognition rate with compactness I are more reliable than that with compactness II. In order to have more accurate texture information, the computation of the second-order joint probability matrix $p_{d,\theta}(i, j)$ is only based on the segmented region of the original mammogram. For the shrunk mammograms, we found that



(a) (b)

Fig. 7. The data mapping results (Database D, see Table II). -o- denotes true mass cases; -*- denotes false mass cases. (a) The mapping using compactness I. (b) The mapping using compactness II.

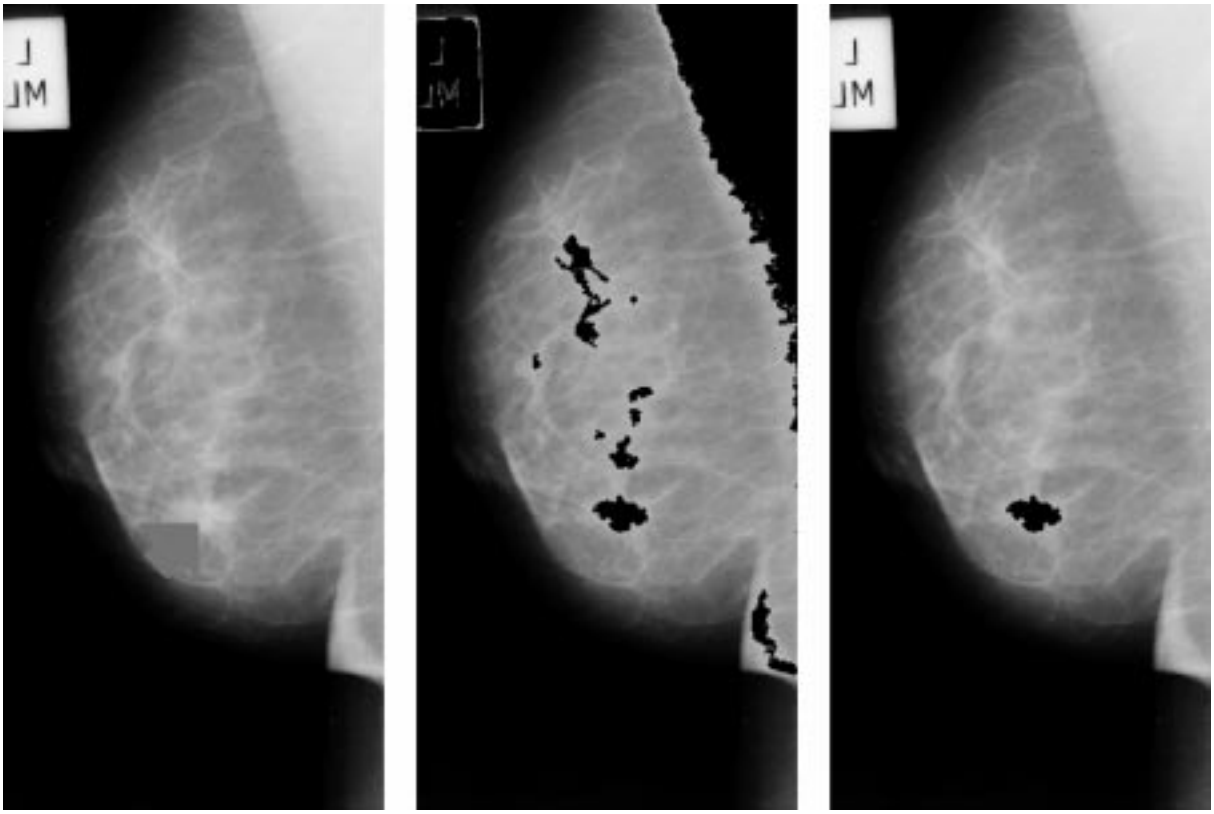


Fig. 8. One example of the mass detection using the proposed approach (Database E, see Table II).

the difference entropy had better discrimination with $d = 1$. The difference entropy used in this study was the average of values at θ =0°, 45°, 90°, and 135°.

We have conducted a preliminary study to evaluate the performance of the algorithms in real case detection, in which 6–15 suspected masses/mammogram were detected and required further clinical decision making. We found that the proposed classifier can reduce the number of suspicious masses with a sensitivity of 84% at 1.6 false positive findings/mammogram based on the testing data set containing 46 mammograms (23 of them have biopsy proven masses) (Database E, see Table II). Fig. 8 shows a representative mass detection result on one mammogram with a stellate mass. After the enhancement, ten regions with brightest intensity were segmented. Using the area criterion, too large and too small regions were eliminated first and the rest regions were submitted to the PMNN for further evaluation. The results indicated that the stellate mass lesion was correctly detected.

For further evaluation, receiver operating characteristic (ROC) method may be employed. However, we do not feel

ROC analysis will provide really a better evaluation but an alternative method to this case. First, most ROC analysis reported by others were based on different database thus are not comparable since ROC results are highly data-dependent. Second, ROC analysis only indicate an "overall" performance with limitations at least in twofold: it is for multithreshold thus the corresponding system may not be optimal to a particular application where only one threshold is needed; and it cannot provide a mathematically traceable feedback to improve the performance of the system or the one component in the system. Third, currently used FROC analysis package imposes several assumptions on the distributions of the cases which are invalid in most applications and particularly untrue in our situation. For example, our assumptions about the data distributions is SFNM that is clearly different from the restricted conditions imposed by the application of existing FROC analysis algorithm. In our approach, a quantitative mapping of the knowledge database is performed with hierarchical SFMD modeling and should be perfectly (at least in the theoretical sense) carried out by the corresponding PMNN classifier. In other words, optimal decision making should have already been achieved according to the Bayesian rule. It is reasonable to acknowledge that in order to compare the overall performance with the other systems, an ROC study may be further conducted. We are currently working on developing a new generation of FROC analysis package with a caution to remove the forementioned problems.

Another important consideration with the present approach is the measure of quality in visual explanation [29]. This is not a glamorous area, but progress in this area is eminently critical to the future success of visual exploration [28]. What is the correct matrix for a direct projection of a particular multimodal data set? How effective was a particular visualization tool? Did the user come to the correct conclusion? It may be agreeable that the benchmark criteria in visual exploration are very different and difficult [28]. As shared by Bishop and Tipping [27], we believe that in data visualization there is no objective measure of quality, and so it is difficult to quantify the merit of a particular data visualization technique, and the effectiveness of such a techniques is often highly data-dependent. The possible alternative is to perform a rigorous psychological evaluation using simple and controlled environment, or to invite domain experts to direct evaluate the efficacy of the algorithm for a specified task. For example, we can compare the domain expert's performances with and without the system aid. In that case, the ROC method may be used to evaluate the performance of our algorithm when used by the radiologists. While the optimality of these new techniques is often highly data-dependent, we would expect the hierarchical visualization model to be a very effective tool for the data visualization and exploration in many applications.

In summary, we employed a mathematical feature extraction procedure to construct the featured knowledge database from all the suspicious mass sites localized by the enhanced segmentation. The optimal mapping of the data points was then obtained by learning the generalized normal mixtures and decision boundaries. A visual explanation of the decision making was further invented as a decision support, based on an interactive visualization hierarchy through the probabilistic principal component projections of the knowledge database and the localized optimal displays of the retrieved raw data. A prototype system was developed and pilot tested to demonstrate the applicability of this framework to mammographic mass detection.

ACKNOWLEDGMENT

The authors would like to thank R. F. Wagner of the Food and Drug Administration and S.-Y. Kung of the Princeton University for their valuable scientific input.

REFERENCES

[1] R. Zwiggelaar, T. C. Parr, J. E. Schumm, I. W. Hutt, C. J. Taylor, S. M. Astley, and C. R. M. Boggis, "Model-based detection of spiculated lesions in mammograms," *Med. Image Anal.*, vol. 3, no. 1, pp. 39–62, 1999.

[2] N. Karssemeijer and G. M. te Brake, "Detection of stellate distortions in mammogram," *IEEE Trans. Med. Imag.*, vol. 15, pp. 611–619, Oct. 1996.

[3] L. Miller and N. Ramsey, "The detection of malignant masses by nonlinear multiscale analysis," *Excerpta Medica*, vol. 1119, pp. 335–340, 1996.

[4] N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computer-aided breast mass detection: False positive reduction using breast tissue composition," *Excerpta Medica*, vol. 1119, pp. 373–378, 1996.

[5] W. K. Zouras, M. L. Giger, P. Lu, D. E. Wolverton, C. J. Vyborny, and K. Doi, "Investigation of a temporal subtraction scheme for computerized detection of breast masses in mammograms," *Excerpta Medica*, vol. 1119, pp. 411–415, 1996.

[6] M. Zhang, M. L. Giger, C. J. Vyborny, and K. Doi, "Mammographic texture analysis for the detection of spiculated lesions," *Excerpta Medica*, vol. 1119, pp. 347–351, 1996.

[7] W. P. Kegelmeyer Jr., J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," *Radiology*, vol. 191, pp. 331–337, 1994.

[8] R. N. Strickland, "Tumor detection in nonstationary backgrounds," *IEEE Trans. Med. Imag.*, vol. 13, pp. 491–499, June 1994.

[9] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Alder, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," *Phys. Med. Biol.*, vol. 40, pp. 857–876, 1995.

[10] M. L. Giger, C. J. Vyborny, and R. A. Schmidt, "Computerized characterization of mammographic masses: Analysis of spiculation," *Cancer Lett.*, vol. 77, pp. 201–211, 1994.

[11] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[12] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[13] R. Schalkoff, *Pattern Recognition: Statistical, Structural, and Neural Approaches*. New York: Wiley, 1992.

[14] Y. Wang, S. H. Lin, H. Li, and S. Y. Kung, "Data mapping by probabilistic modular networks and information theoretic criteria," *IEEE Trans. Signal Processing*, vol. 46, pp. 3378–3397, Dec. 1998.

[15] L. Perlovsky and M. McManus, "Maximum likelihood neural networks for sensor fusion and adaptive classification," *Neural Networks*, vol. 4, pp. 89–102, 1991.

[16] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1990, pp. 1361–1364.

[17] D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[18] C. E. Priebe, "Adaptive mixtures," *J. Amer. Stat. Assoc.*, vol. 89, no. 427, pp. 910–912, 1994.

[19] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: MacMillan College, 1994.

[20] M. I. Jordan and R. A. Jacobs, "Hierarchical mixture of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[21] S. Y. Kung and J. S. Taur, "Decision-based neural networks with signal/image classification applications," *IEEE Trans. Neural Networks*, vol. 1, pp. 170–181, Jan. 1995.

[22] S. H. Lin, S. Y. Kung, and L. J. Lin, "Face recognition/detection by probabilistic decision-based neural network," *IEEE Trans. Neural Networks (Special issue on Artificial Neural Networks and Pattern Recognition)*, vol. 8, Jan. 1997.

[23] Y. Wang, L. Luo, H. Li, and M. T. Freedman, "Hierarchical minimax entropy modeling and probabilistic principal component visualization for data explanation and exploration," presented at the SPIE Medical Imaging Conf., San Diego, CA, Feb. 20–26, 1999.

[24] H. Li, Y. Wang, K. J. R. Liu, S.-C. B. Lo, and M. T. Freedman, "Computerized Radiographic Mass Detection—Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation," *IEEE Trans. Med. Imag.*, vol. 20, no. 4, pp. 289–301, Apr. 2001.

[25] T. W. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[26] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Berlin, Germany: Springer–Verlag, 1988.

[27] C. M. Bishop and M. E. Tipping, "A hierarchical latent variable model for data visualization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 281–293, Mar. 1998.

[28] G. M. Nielson, "Challenges in visualization research," *IEEE Trans. Visual. Comput. Graphics*, vol. 2, pp. 97–99, 1996.

[29] E. R. Tufte, *Visual Explanation: Images and Quantities, Evidence and Narrative*. Cheshire, U.K.: Graphics, 1996.