ARTICLE IN PRESS

ELSEVIER

Medical Image Analysis 1 (2003) 000–000

MEDICAL IMAGE ANALYSIS

www.elsevier.com/locate/media

A multi-scale probabilistic network model for detection, synthesis and compression in mammographic image analysis

Paul Sajda[a,*], Clay Spence[b], Lucas Parra[b]

[a]Department of Biomedical Engineering, Columbia University, 351 Engineering Terrace, MC 8904, New York, NY 10027, USA
[b]Sarnoff Corporation, 201 Washington Road, Princeton, NJ, USA

Accepted 7 January 2003

Abstract

We develop a probabilistic network model over image spaces and demonstrate its broad utility in mammographic image analysis, particularly with respect to computer-aided diagnosis. The model employs a multi-scale pyramid decomposition to factor images across scale and a network of tree-structured hidden variables to capture long-range spatial dependencies. This factoring makes the computation of the density functions local and tractable. The result is a hierarchical mixture of conditional probabilities, similar to a hidden Markov model on a tree. The model parameters are found with maximum likelihood estimation using the expectation-maximization algorithm. The utility of the model is demonstrated for three applications: (1) detection of mammographic masses for computer-aided diagnosis; (2) qualitative assessment of model structure through mammographic synthesis; and (3) compression of mammographic regions of interest.

Keywords: Probabilistic network model; Multi-scale pyramid decomposition; Mammographic computer-aided diagnosis; Image synthesis; Image compression

1. Introduction

Computer-aided diagnosis (CAD) can be defined as a diagnosis made by a radiologist who incorporates the results of computer analysis of the radiographs (Doi et al., 1993). The goal of CAD is to improve radiologists' performance by indicating the sites of potential abnormalities, to reduce the number of missed lesions, and/or by providing quantitative analysis of specific regions in an image to improve diagnosis. CAD systems typically operate as automated "second-opinion" or "double-reading" systems that indicate lesion location and/or type. Since individual human observers overlook different findings, it has been shown that double reading (the review of a study by more than one observer) increases the detection rate of breast cancers by 5–15% (Bird, 1990; Metz and Shen, 1992; Thurfjell et al., 1994). Double reading, if not done efficiently, can significantly increase the cost of screening, given the need for a second radiologist/mammographer. Methods to provide improved detection with little increase in cost will have significant impact on the benefits of screening. Automated CAD systems are a promising approach for low-cost double-reading.

Several CAD systems have been developed for mammographic screening and the first have been approved by the FDA.[1] Complete systems have been rigorously characterized, both in retrospective and prospective trials (Burhenne et al., 2000). Though many have demonstrated clinical utility, there is still a need to reduce false-positive rates generated by CAD systems. For example, prospective clinical studies have shown lower sensitivities and specificities than originally found in retrospective studies—80%

*Corresponding author. Tel.: +1-212-854-5270.
E-mail address: sajda@columbia.edu (P. Sajda).

[1]R2 Technology M1000 Image Checker, CADx Medical Second-Look, and Intelligent Systems Software MammoReader.

1361-8415/03/$ – see front matter
doi:10.1016/S1361-8415(03)00003-3

cancers detected with 2.4 false positives per case in prospective studies versus 85–90% sensitivity at one to two false positives per case in retrospective studies (Nishikawa et al., 1996).

1.1. The role of statistical pattern recognition and neural networks in CAD

CAD systems usually consist of two distinct subsystems, one designed to detect microcalcifications and one to directly detect masses (Giger et al., 2000). A common element in both subsystems is a statistical pattern recognition model, used to improve detection and reduce false-positive rates introduced by earlier stages of processing. Neural networks are a particularly important class of statistical model in CAD because they are able to capture complicated, often nonlinear, relationships in high dimensional feature spaces not easily captured by heuristic or rule-based algorithms. Several groups have developed neural networks architectures for CAD. Many of these architectures exploit well-known features that might also be used by radiologists (Floyd et al., 1994; Jiang et al., 1996; Huo et al., 1998), while others utilize more generic feature sets (Zhang et al., 1994; Lo et al., 1996; Chan et al., 1998; Sajda et al., 2002). In general, these neural networks are *recognition* or *discriminative* probabilistic models (Dayan and Abbott, 2002) in that they estimate $\Pr(C \mid I)$, the conditional probability of class C (e.g., mass vs. non-mass) given image I or a set of features extracted from I. An alternative approach is to construct a *generative* probabilistic model of the data, which, using the aforementioned formulation, would be a model that estimates the class conditional distribution, $\Pr(I \mid C)$. Such a model has several attractive features for mammographic image analysis. For example, classification is possible by training a distribution for each class and using Bayes' rule to obtain $\Pr(C \mid I) = \Pr(I \mid C)\Pr(C)/\Pr(I)$. In addition, novel examples, relative to the training data used to build the model, can be detected by computing the absolute likelihood over each model. In terms of CAD, the ability to identify novel examples is useful for establishing confidence measures on the CAD output (e.g., should the output of the classifier be ''trusted'' given that the current data is very different from the training data). In addition, novelty detection can be used to identify new clinical data that might be used to re-train/refine the CAD system. Since essentially any type of image analysis can be formulated given knowledge of the distribution of the data, the generative probabilistic model can also be used to compress (Cover and Thomas, 1991), suppress noise (Romberg et al., 2001), interpolate, increase or extend resolution (Freeman et al., 2002), etc.

1.2. Generative probabilistic models for images

Previous research has focused on developing probabilistic models of biological and natural shapes, much of which is based on the work of Grenander et al. (1991). This has in turn led to the development of active shape and appearance models for medical image analysis, most notably those of Cootes and Taylor (Cootes et al., 1994; Cootes and Taylor, 2001). These approaches construct a statistical description of object shape over a set of landmarks that are often extracted by a human expert (e.g., radiologist). The statistical descriptions are formulated as generative models and can be sampled to construct new instances of a given shape. The approach has demonstrated great utility in localizing structure in medical imagery, particularly in cases where the structure is well described by its contours/borders (e.g., ventricles in the brain and heart). However, in the case of mammography, lesions are not well characterized by border shape alone. Radiologists typically integrate evidence which includes texture, homogeneity, and spiculation, as well as contextual information such as vascularization and proximity to mammillary ducts (Kopans, 1989). Therefore, a concise set of landmarks is not easily extracted and instead a classification system must learn the set of shape and non-shape features which are correlated with disease (or absence of disease).

Significant efforts have also focused on the construction of generative probabilistic models for directly modeling images. Grenander (1983) was one of the first to propose a Bayesian framework for image analysis. This framework led to the development of a series of image distribution models, most notably the Markov Random Field (MRF) developed by Geman and Geman (1984) and further developed and studied by others (e.g., Chellappa and Chatterjee, 1985). MRFs model distributions by assuming that images are locally smooth except for relatively sparse intensity gradients and edges. The underlying assumption of an MRF is that local image structure is sufficient for global image representation. However, these models tend to be computationally expensive, have limited forms for the distributions/potential functions, and have difficulty capturing more global structure and long-range dependencies in images. Zhu et al. (1997) attempted to overcome some of the limitations in MRFs by computing distributions over a set of features constructed from the histograms of filtered images (e.g., using Gabor filters). In their approach they compute the maximum entropy distribution given the statistics across these features. Though this approach works well for textures, it is not clear how well it models the appearance of more structured objects.

De Bonet and Viola proposed a flexible histogram approach (De Bonet and Viola, 1998; De Bonet et al., 1998), where features are extracted at multiple image scales, with the resulting feature vectors treated as a set of independent samples drawn from a distribution. The distribution of feature vectors is subsequently modeled using Parzen windows. Though they report good results, their model treats the feature vectors from neighboring pixels as independent samples, when in fact they share exactly the same components from lower resolutions. One

solution to this is to build a model in which the features at one pixel of one pyramid level condition the features at each of several child pixels at the next higher-resolution pyramid level. The multiscale stochastic process (MSP) methods do exactly that. Luettgen and Willsky (1995), for example, applied a scale-space auto-regression (AR) model to texture discrimination. They use a quadtree or quadtree-like organization of the pixels in an image pyramid, and model the features in the pyramid as a stochastic process from coarse-to-fine levels along the tree. The variables in the process are hidden, and the observations are sums of these hidden variables plus noise. However, the assumed Gaussian distributions are a limitation of MSP models as well as the fact that the model is of the probability of the observations on the tree, not of the image. Once again, these methods appear well suited for modeling texture, but it is unclear how one might build models to capture the appearance of more structured objects, such as mammographic masses.

Recently, several groups have developed what are essentially extensions of the MSP model by adding hidden variables. These can be seen as improving the model's ability to capture non-local dependencies in the image. For example, Crouse et al. (1998) developed their Hidden Markov Tree (HMT) models for signals and images. A primary motivation of these models is to capture the tendency for wavelet coefficients to group into two classes, one with large and the other with small coefficient magnitudes. Thus their hidden states have one of two values corresponding to large and small wavelet coefficients. This is well suited to the many signal and image types that have homogeneous regions with boundaries. These models have been successfully applied to several problems, especially image enhancement and texture segmentation (Romberg et al., 2001; Coi and Baraniuk, 2001). Cheng and Bouman (2001) applied a similar model for segmentation, in which the observed class labels play the role of hidden variables, and therefore are no longer hidden.

We have developed a class of models for probability distributions of images that we call hierarchical image probability (HIP) models. The HIP model can be viewed as a development of the HMT model, with several differences. The main elements of both the HIP and HMT models include:

- Capturing local dependencies in a coarse-to-fine factoring of the image distribution over scale and position.
- Capturing non-local and scale dependencies through a set of discrete hidden variables whose dependency graph is a tree.
- Optimizing model parameters to match the natural image statistics using strict Maximum Likelihood.
- Enabling both evaluation of the likelihood and sampling from the distribution.

In addition, the HIP model differs from the HMT model in the following ways:

- The coefficients of the different subbands at each node are modeled jointly, using mixtures of multivariate Gaussian distributions.[2]
- The number of hidden states in each level is adjusted separately in an attempt to better fit the image distribution.
- The hidden states capture complex structure in the image through the use of mixture, hierarchy and scale components.
- The probability of a child state value at a node, conditioned on the state at the parent node, also depends on the child node's relative position, e.g. upper-left, lower-right, etc.
- The mean of each normal distribution depends on the corresponding coefficient vector in the unsampled wavelet coefficient subbands from the next coarsest resolution pyramid level. (The HIP model resembles a simple MSP model in this way.)

In the following we begin by presenting the structure of the HIP model, along with an EM algorithm used to estimate its parameters. We first describe the most simple form of the HIP model, namely with a single component in the hidden variable structure, and then augment the model to include mixture, hierarchy and scale components. We then demonstrate the broad utility of the complete model by presenting results for several applications in mammographic image analysis, including mass detection in CAD, mammographic synthesis, and compression of mammographic ROIs. In all cases we compare results to a traditional HMT (Crouse et al., 1998).

2. Structure of the HIP model

2.1. Coarse-to-fine factoring of image distributions

Similar to previous work (Luettgen and Willsky, 1995; De Bonet and Viola, 1998; De Bonet et al., 1998; Crouse et al., 1998; Cheng and Bouman, 2001) we model dependencies in an image over a range of scales. We begin by representing the image as a set of feature images, for example computed using a set of filters with limited spatial support. Coarse-scale image structure is captured by applying the set of filters at a low-resolution in a pyramid decomposition of the image. Long-range dependencies of fine-scale structure are modeled by conditioning fine scales on coarse scales. Denoting the set of feature images at pyramid level l by $\mathbf{F}_l$, the goal is to write the image distribution in a form similar to $\Pr(I) \sim \Pr(\mathbf{F}_0 \mid \mathbf{F}_1)\Pr(\mathbf{F}_1 \mid \mathbf{F}_2) \ldots$.

We first prove that a coarse-to-fine factoring of this form is exact. From an image I build a low-pass (e.g., Gaussian) pyramid. Call the lth level of this pyramid I_l, with the

[2] Arbitrarily complex distributions can be modeled as mixtures of Gaussians.

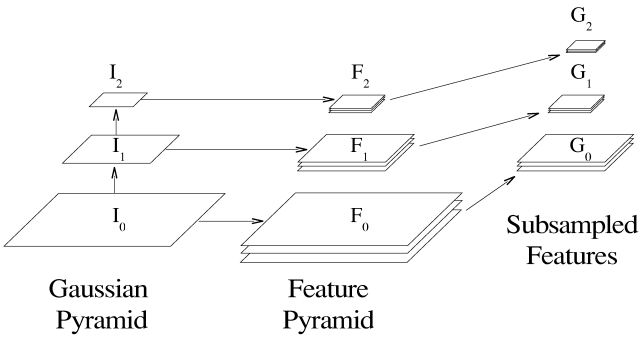


Fig. 1. Pyramids and feature notation used to demonstrate coarse-to-fine factoring.

original full resolution image denoted as I_0 (see Fig. 1). For each low-pass image I_l at level l extract a set of feature images $\mathbf{F}_l$. Appropriate sub-sampling of these features results in $\mathbf{G}_l$, having the same dimensions as I_{l+1}. Denote by $\tilde{\mathbf{G}}_l$ the set of images containing I_{l+1} and the images in $\mathbf{G}_l$. Finally, denote the mapping from I_l to $\tilde{\mathbf{G}}_l$ as $\tilde{\mathcal{G}}_l$. We continue this decomposition up to layer L.[3] Wavelet transforms are well-known examples of such mappings. Alternately, one could build Gaussian pyramids (Burt and Adelson, 1983) to obtain I_l and then filter these with several carefully chosen band-pass filters, followed by subsampling, as shown in Fig. 1.

Suppose now that $\tilde{\mathcal{G}}_l : I_l \mapsto \tilde{\mathbf{G}}_l$ is invertible. Then $\tilde{\mathcal{G}}_l$ is a change of variables, and we can relate the distributions on the two different sets of variables through multiplication by the Jacobian, i.e., $\Pr(I_l) = |\tilde{\mathcal{G}}_l| \Pr(\tilde{\mathbf{G}}_l)$. Since $\tilde{\mathbf{G}}_l = (\mathbf{G}_l, I_{l+1})$, $\Pr(\tilde{\mathbf{G}}_l)$ can be factored to obtain $\Pr(I_l) = |\tilde{\mathcal{G}}_l| \Pr(\mathbf{G}_l \mid I_{l+1}) \Pr(I_{l+1})$. If $\tilde{\mathcal{G}}_l$ is invertible for all $l \in \{0, \ldots, L\}$ then we can recursively apply this change of variables and factoring procedure to obtain[4]

$$\Pr(I) = \left[\prod_{l=0}^{L} |\tilde{\mathcal{G}}_l| \Pr(\mathbf{G}_l \mid I_{l+1}) \right]. \tag{1}$$

This is a very general result, valid for all $\Pr(I)$, requiring only that the mapping be invertible and unique.

If our features are the outputs of linear filters, the determinants $|\tilde{\mathcal{G}}_l|$ depend only on the filters used, and not on the image or model parameters. Therefore, we can drop the determinants if we write Eq. (1) as a proportionality

$$\Pr(I) \propto \prod_{l=0}^{L} \Pr(\mathbf{G}_l \mid I_{l+1}). \tag{2}$$

Note that for comparing the likelihoods with different features or sampling from the distribution (e.g., synthesis) it will be important to keep the determinants.

2.2. Hidden variables for modeling non-local dependencies

For the sake of computational tractability we would like to factor $\Pr(\mathbf{G}_l \mid I_{l+1})$ over position. However, this is problematic due to non-local dependencies that remain after coarse-to-fine factoring. Fig. 2 illustrates these dependencies. Assume that the presence of a particular object, O_A (e.g., mammographic mass), may be inferred with high probability at a coarse scale, $l+1$, of the image pyramid. Assume further that the presence of O_A implies with high probability the presence of a texture at position x at a finer scale l. Since we may need to examine an extended spatial area at level $l+1$ to detect the object, the presence of the texture at x in level l can depend upon an extended spatial area in level $l+1$, i.e., the dependence between scales is non-local. Similarly (see Fig. 2(B)), the information at level $l+1$ may be sufficient for detecting an object but not for distinguishing its class, O_A or O_B (e.g., mass vs. non-mass). However, the detection of a single object imposes constraints of structure at high resolutions, for instance that distant positions in l have similar texture. Once again, conditioning fine scales on coarser scales cannot capture these long-range dependencies, which are entirely within the finer scale in this example.

To capture these dependencies we introduce a hidden variable A that takes on values in some set $\mathcal{A}$. We assume that A contains sufficient information for $\Pr(\mathbf{G}_l \mid I_{l+1}, A)$ to factor over position x

$$\begin{aligned} \Pr(I) &\propto \sum_{A \in \mathcal{A}} \left[\prod_{l=0}^{L} \Pr(\mathbf{G}_l \mid I_{l+1}, A) \right] \Pr(A) \\ &= \sum_{A \in \mathcal{A}} \left[\prod_{l=0}^{L} \prod_{x \in \mathcal{P}_l} \Pr(\mathbf{g}_l(x) \mid I_{l+1}, A) \right] \Pr(A). \end{aligned} \tag{3}$$

Here, $\mathcal{P}_l$ is the set of all positions in resolution level l in the wavelet/multi-resolution decomposition.[5] Note that by replacing uppercase letters (e.g., $\mathbf{G}_l$) with lowercase letters which are functions of x (e.g., $\mathbf{g}_l(x)$) we are indicating a factoring of the features over position.

To simplify, we assume that, given A, $\mathbf{g}_l(x)$ depends only on the local information in I_{l+1} which is captured by $\mathbf{f}_{l+1}(x)$, the features of I_{l+1} at position x. To be precise, the complete decomposition of I_{l+1} requires in addition to the high-pass features F_{l+1} also the low-pass information I_{l+2}. To simplify the presentation, we drop this, essentially assuming that A carries all of the coarse-scale intensity information from I_{l+2} that is relevant for $\mathbf{G}_l$. (In practice, it

[3] It will prove convenient to define $\mathbf{G}_L$ to be the same as $\tilde{\mathbf{G}}_L$.

[4] For the last layer L the conditioning on I_{L+1} is to be ignored, since we defined $\mathbf{G}_L$ to include I_{L+1}.

[5] In the following we frequently simplify expressions by omitting the limits of the sums and products, since they should be clear from context.

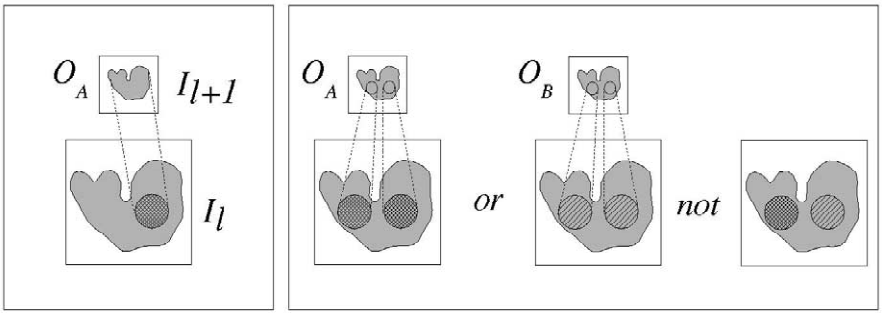


Fig. 2. Example of dependencies that cannot be captured by a coarse-to-fine factoring. (A) An object O_A, detectable at level $l+1$, implies the presence of particular texture at location x at level l. Since we place no constraints on the spatial extent of O_A in level $l+1$, the presence of the texture at x can depend upon an extended region in $l+1$. (B) The information in level $l+1$ may be insufficient to discriminate between objects O_A and O_B, however the detection of a single object imposes global dependencies that constrain the textures at distant positions to be homogeneous (either B-left or B-center, but not B-right).

is not difficult to include it, and we do this in the experiments presented later.) This gives

$$\Pr(I) \propto \sum_A \left[\prod_l \prod_x \Pr(\mathbf{g}_l(x) \mid \mathbf{f}_{l+1}(x), A) \right] \Pr(A). \quad (4)$$

In principle, *any* distribution of images can be written in this form since variables A, their joint distribution $\Pr(A)$, and the dependence of the image features on A can have arbitrarily complex structure capturing any non-local behavior.

Before we propose a specific structure for the variables A let us point out that the conditioning of $\mathbf{g}_l(x)$ on $\mathbf{f}_{l+1}(x)$ already captures some of the coarse-to-fine dependency of image statistics. Many image structures, such as edges, persist across scale, and so it is found that modeling this dependency of features across scales is essential for the synthesis of natural texture images (Portilla and Simoncelli, 2000). Note that we choose to condition $\mathbf{G}_l$ on $\mathbf{F}_{l+1}$ instead of $\mathbf{G}_{l+1}$ as in Luettgen and Willsky (1995). We believe that this better captures local correlation since it is consistent with empirically established natural image statistics (Buccigrossi and Simoncelli, 1998), and with equal image dimensions the conditioning becomes straightforward.

Eq. (4) can be seen as a mixture model with mixture labels A conditioning the entire image. We remind the reader that mixture models group samples with common statistics by assigning them a common label (Duda et al., 2001). In this case a sample corresponds to the entire image. However, instead of entire images we intend to group individual pixels in the pyramid. We consider, therefore, the set of hidden variables as an unsupervised segmentation. As such, we assign to each position and layer in the pyramid a variable $a_l(x)$ that conditions the features only locally. This gives

$$\Pr(I) \propto \sum_A \left[\prod_l \prod_x \Pr(\mathbf{g}_l(x) \mid \mathbf{f}_{l+1}(x), a_l(x)) \right] \Pr(A). \quad (5)$$

The structure of the joint distribution $\Pr(A)$ of variables, $A = \{a_l(x) \mid x \in \mathcal{P}_l,\ l = 0, \ldots, L\}$, captures the statistical relation between the segmentation in different regions and scales. In addition, due to the factorization over space the dependency structure of $\Pr(A)$ has to communicate non-local information over different regions of the image and across scale. A tree, as shown in Fig. 3(A), satisfies that requirement and makes the necessary computations tractable. With this choice the joint distribution is given by[6]

$$\Pr(A) = \prod_l \prod_x \Pr(a_l(x) \mid a_{l+1}(Px)), \quad (6)$$

where the probability $\Pr(a_l(x) \mid a_{l+1}(Px))$ is that of finding a_l at x given a_{l+1} at the parent of x, Px. We allow the number of possible values for the labels a_l to be different

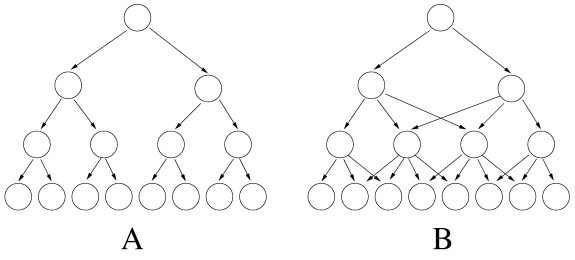


Fig. 3. Dependency structure for the label pyramid. (A) In a binary tree probabilities can be propagated efficiently. The disadvantage is that some neighboring nodes are very weakly linked, while others are very tightly linked. (B) Dense graph where the smallest clique is the entire graph and probability computations increase exponentially with the tree size.

[6] Variable a_{L+1} has not been defined and can be thought of as a label with a single possible value. The conditional distribution $\Pr(a_L \mid a_{L+1})$ then turns into a prior $\Pr(a_L)$. The reader should note that this footnote applies to the remainder of the paper, most notably in the derivation of the expectation in Section 3.2.

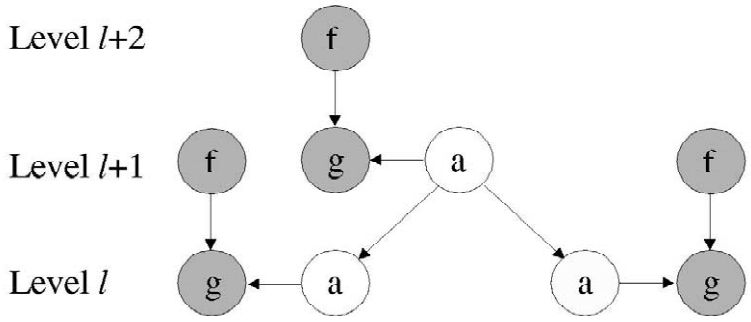


Fig. 4. Dependency structure of the HIP model corresponding to Eq. (7). To simplify the diagram, we show the dependency graph for a single parent node conditioning two of its children. In practice, each parent has four children, i.e., a quadtree. Dark shaded nodes represent observable data. We also omit the subscripts which indicate position. White nodes are hidden variables.

for each level l. Combining Eq. (6) with Eq. (5), and using shorter notation[7] for the position x, we obtain

$$\Pr(I) \propto \sum_{A} \prod_{l} \prod_{x} \Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l, x) \Pr(a_l \mid a_{l+1}, x). \quad (7)$$

The dependency graph of expression (7) is shown in Fig. 4. Note that this is not the only way to introduce hidden variables to capture non-local dependence. The more general model is still given by expression (4). However, expression (7) represents a fairly general class of models with several desired properties, in particular dependencies proceeding from coarse-to-fine scales that are local in both space and scale. The integration of the hidden variable structure into the pyramid framework is depicted in Fig. 5.

3. Training the HIP model with an EM algorithm

We adjust the parameters of our model to match the statistics of a given set of images by using Maximum Likelihood (ML) parameter estimation. The structure of the model in Eq. (7) and illustrated in Fig. 4 permits the exact and efficient computation of all marginal probabilities required for the expectation-maximization (EM) algorithm (Dempster et al., 1977). The algorithm first computes the expectations, over the hidden variables, of the log-likelihood for a given set of parameters and observations (E-step). Then, using these expectations, the likelihood is maximized with respect to the parameters of the model (M-step):

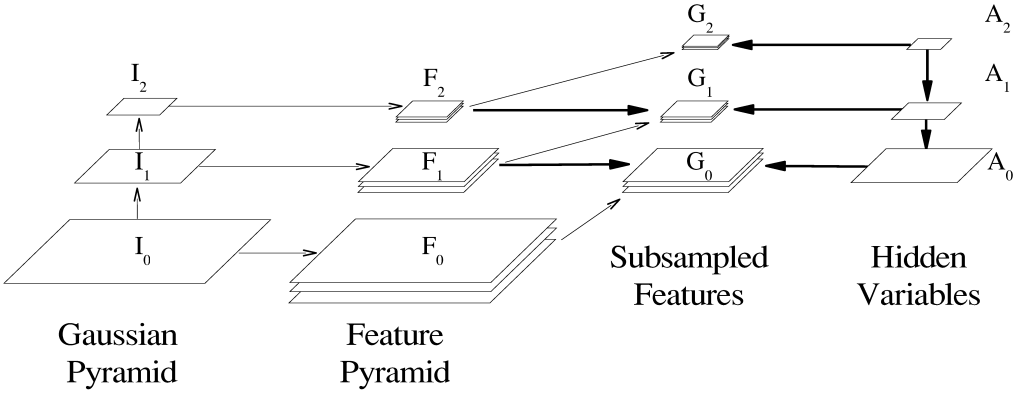


Fig. 5. The addition of hidden variable images for capturing long-range dependencies. Conditioning is shown with thick arrows while construction of features is shown with thin arrows. In this example $L = 2$.

[7] In the following we will write $\Pr(a_l(x) \mid a_{l+1}(Px))$ as simply $\Pr(a_l \mid a_{l+1}, x)$. For brevity we also write $\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l, x)$ for $\Pr(\mathbf{g}_l(x) \mid \mathbf{f}_{l+1}(x), a_l(x))$, the probability distribution for finding the feature vector $\mathbf{g}_l(x)$ at position x given that the feature vector $\mathbf{f}_{l+1}(x)$ and hidden variable $a_l(x)$ were also found at x. Similar notation will be used for other expressions. The argument x in $\Pr(\cdot \mid \cdot, x)$ selects the random variables associated with position x and should not be understood as a random variable by itself.

$$\text{E-step:} \quad Q(\theta \mid \theta^t) = \sum_{A} \Pr(A \mid I, \theta^t) \ln \Pr(I, A \mid \theta), \quad (8)$$

$$\text{M-step:} \quad \theta^{t+1} = \arg\max_{\theta} Q(\theta \mid \theta^t). \quad (9)$$

Here we have summarized all parameters of the model in θ, and θ^t represents the values of the parameters in the current iteration step t.

The main challenge for this model lies in computing the expectations over the unknown labels. In this section, only the resulting equations will be given. For the derivation of the probability propagation in this hierarchical model readers are referred to Appendix A.

3.1. Maximization

We start with the M-step by inserting Eq. (7) into Eq. (8):

$$Q(\theta \mid \theta^t)$$

$$= \sum_A \Pr(A \mid I,\theta^t) \sum_l \sum_x \ln \Pr(\mathbf{g}_l, a_l \mid \mathbf{f}_{l+1}, a_{l+1}, x, \theta) + \text{const.}, \tag{10}$$

$$= \sum_l \sum_x \sum_{a_l, a_{l+1}} \{\Pr(a_l, a_{l+1} \mid I,x,\theta^t) \ln \Pr(\mathbf{g}_l, a_l \mid \mathbf{f}_{l+1}, a_{l+1}, x)\} + \text{const.} \tag{11}$$

Here, $\Pr(a_l, a_{l+1} \mid I,x,\theta^t)$ represents the marginal probabilities of pairs of labels from neighboring layers at position x for given image data and the current parameter values. The additive constant is due to the proportionality factors of Eq. (7). Assuming we know the probability $\Pr(a_l, a_{l+1} \mid I,x,\theta^t)$ for all parent/child label pairs, a_l, a_{l+1}, we can search for the optimal parameters. At this point we must commit to a parameterization of $\Pr(a_l \mid a_{l+1}, x)$ and $\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l, x)$. We use the same parameters for all positions so that we obtain homogeneous behavior across the image, a constraint which is often referred to as ''tying'' in the HMM literature (Rabiner, 1989), and is also used in the HMT model (Crouse et al., 1998). However, we allow our model to have different parameters at different pyramid levels—we tie across position but not scale. We allow $\Pr(a_l \mid a_{l+1}, x)$ to depend on the position of the child relative to the parent, e.g. the probability is different for the upper-right child than for the lower-left child, etc. We also choose to parameterize $\Pr(a_l \mid a_{l+1}, x)$ as

$$\Pr(a_l \mid a_{l+1}) = \frac{\pi_{a_l, a_{l+1}}}{\sum_{a_l} \pi_{a_l, a_{l+1}}}. \tag{12}$$

The values of the parameters $\pi_{a_l, a_{l+1}}$ are determined during the updates in the M-step of the EM algorithm. Note that with this definition $\Pr(a_l \mid a_{l+1})$ is always properly normalized. There is an arbitrary scale in the $\pi_{a_l, a_{l+1}}$ for each a_{l+1}, but this is fixed by choosing a particular form for the update equation. Note also that we omit x in this notation as the parameterization is independent of the position within a layer.

We assume a simple model for the distribution of subsampled features conditioned on the features of the next highest pyramid level. Our model represents a mixture where the label a selects the mixture component. We choose a Gaussian distribution where the parameters are indexed by the labels and the dependency of the features is parameterized as a linear relationship in the mean:

$$\Pr(\mathbf{g} \mid \mathbf{f}, a) = \mathcal{N}(\mathbf{g}, M_a \mathbf{f} + \bar{\mathbf{g}}_a, \Lambda_a). \tag{13}$$

If different features at a given spatial location (a pixel) are independent, then diagonal M and Λ are sufficient. The parameter set is now defined as

$$\theta = \cup_l \{\pi_{a_l, a_{l+1}}, M_{a_l}, \bar{\mathbf{g}}_{a_l}, \Lambda_{a_l} \mid a_l \in \{1, \ldots, N_{a_l}\}\}.$$

With the choices (12) and (13) the M-step is easily solved. The maximum of (11) with respect to θ can be found by setting the derivatives with respect to the different parameters equal to zero and solving for the corresponding parameter. For $\pi^{t+1}_{a_l, a_{l+1}}$ we find

$$\pi^{t+1}_{a_l, a_{l+1}} = \sum_x \Pr(a_l, a_{l+1} \mid I,x,\theta^t). \tag{14}$$

For the remaining update equations we define the following weighted average:

$$\langle X \rangle_{t, a_l} = \frac{\sum_x \Pr(a_l \mid I,x,\theta^t) X(x)}{\sum_x \Pr(a_l \mid I,x,\theta^t)}. \tag{15}$$

The weights $\Pr(a_l \mid I,x,\theta^t)$ represent the marginal probabilities of finding label value $a_l(x)$ at position x given the image data and the current parameter values.

The update equations are

$$\bar{\mathbf{g}}^{t+1}_{a_l} = \langle \mathbf{g}_l \rangle_{t, a_l} - M^{t+1}_{a_l} \langle \mathbf{f}_{l+1} \rangle_{t, a_l}, \tag{16}$$

$$M^{t+1}_{a_l} = (\langle \mathbf{g}_l \mathbf{f}^T_{l+1} \rangle_{t, a_l} - \bar{\mathbf{g}}^{t+1}_{a_l} \langle \mathbf{f}^T_{l+1} \rangle_{t, a_l}) \times \langle \mathbf{f}_{l+1} \mathbf{f}^T_{l+1} \rangle^{-1}_{t, a_l}, \tag{17}$$

and

$$\Lambda^{t+1}_{a_l} = \langle (\mathbf{g}_l - M^{t+1}_{a_l} \mathbf{f}_{l+1})(\mathbf{g}_l - M^{t+1}_{a_l} \mathbf{f}_{l+1})^T \rangle_{t, a_l} - \bar{\mathbf{g}}^{t+1}_{a_l} \bar{\mathbf{g}}^{t+1\,T}_{a_l}. \tag{18}$$

Since these expressions are mutually dependent, we must insert Eq. (16) into Eq. (17) and solve for $M^{t+1}_{a_l}$ to obtain

$$M^{t+1}_{a_l} = (\langle \mathbf{g}_l \mathbf{f}^T_{l+1} \rangle_{t, a_l} - \langle \mathbf{g}_l \rangle_{t, a_l} \langle \mathbf{f}^T_{l+1} \rangle_{t, a_l})(\langle \mathbf{f}_{l+1} \mathbf{f}^T_{l+1} \rangle_{t, a_l} - \langle \mathbf{f}_{l+1} \rangle_{t, a_l} \langle \mathbf{f}^T_{l+1} \rangle_{t, a_l})^{-1}. \tag{19}$$

To summarize, the update procedure is:

1. compute $M^{t+1}_{a_l}$ according Eq. (19),
2. compute $\bar{\mathbf{g}}^{t+1}_{a_l}$ according to Eq. (16), then
3. compute $\Lambda^{t+1}_{a_l}$ according to Eq. (18).

If we assume diagonal M and Λ we can ignore the off-diagonal terms in these expressions. In fact, the component densities $\mathcal{N}(\mathbf{g}, M_a \mathbf{f} + \bar{\mathbf{g}}_a, \Lambda_a)$ factor into individual densities for each component of $\mathbf{g}$. We can replace Eqs. (19), (16), and (18) with their scalar versions and apply them to each component of $\mathbf{g}$ independently.

3.2. Expectation

In the E-step we compute the marginal probabilities of pairs of labels from neighboring layers $\Pr(a_l,a_{l+1} \mid I,x,\theta^t)$ for given image data. However, note that, in all occurrences of the re-estimation equations, i.e., Eqs. (12), (14) and (15), we need that quantity only up to an overall factor. We can choose that factor to be $\Pr(I \mid \theta^t)$ and therefore compute $\Pr(a_l,a_{l+1},I \mid x,\theta^t)$ using

$$\Pr(a_l,a_{l+1} \mid I,x,\theta^t)\Pr(I \mid \theta^t) = \Pr(a_l,a_{l+1},I \mid x,\theta^t) = \sum_{A\setminus a_l(x),a_{l+1}(x)} \Pr(I,A \mid \theta^t). \quad (20)$$

The complexity of computing these sums relates to the dependency structure of the variables A, which we have already defined in Eq. (7) and Fig. 4.

From the viewpoint of computational complexity, it is important to understand the rationale for this choice. From the literature on graphical models (Jordan, 1998) we know that the cost of evaluating these sums grows exponentially with the clique sizes in the graph and linearly with the number of cliques. If we choose the dependency such that every label is conditioned on only one label from the parent layer then the clique size is minimal (Fig. 3(A)). For an image pyramid with subsampling-by-two that corresponds to a quadtree structure. In a quadtree a location x_l has only one parent Px_l in layer $l+1$, and four children Cx_l in layer $l-1$. If we do not restrict the dependencies, and maintain instead a more general belief network structure between layers, with local connectivity (Fig. 3(B)), the entire label pyramid is one irreducible clique, and the exact evaluation of the sums becomes prohibitive.

We now compute the probability of hidden labels given the entire image pyramid. This computation will be essentially the same as propagating the probabilities of observations of the entire pyramid to a particular junction of label pairs. Probabilities first propagate upwards, and then downward to a particular label pair. During the propagation we marginalize over the other labels. We recursively define quantities u and d, representing the upwards and downwards propagating probabilities:

$$u_l(a_l,x) = \Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1},a_l,x) \prod_{x' \in Cx} \tilde{u}_{l-1}(a_l,x'), \quad (21)$$

$$\tilde{u}_l(a_{l+1},x) = \sum_{a_l} \Pr(a_l \mid a_{l+1})\, u_l(a_l,x), \quad (22)$$

$$d_l(a_l,x) = \sum_{a_{l+1}} \Pr(a_l \mid a_{l+1})\, \tilde{d}_l(a_{l+1},x), \quad (23)$$

$$\tilde{d}_l(a_{l+1},x) = \frac{u_{l+1}(a_{l+1},Px)}{\tilde{u}_l(a_{l+1},x)}\, d_{l+1}(a_{l+1},Px). \quad (24)$$

The upward recursion (Eqs. (21) and (22)) is initialized at $l=0$ with $u_0(a_0,x) = \Pr(\mathbf{g} \mid \mathbf{f}_1,a_0,x)$ and ends at $l=L$. At layer L, Eq. (22) reduces to $\tilde{u}_L(a_{L+1},x) = \tilde{u}_L(x)$. Since we do not model any further dependencies beyond layer L, the pixels at layer L are assumed independent. The product of all $\tilde{u}_L(x)$ is the total image probability:

$$\Pr(I \mid \theta^t) = \prod_{x \in \mathcal{P}_L} \tilde{u}_L(x) = u_{L+1}. \quad (25)$$

The downward recursion (Eqs. (23) and (24)) starts with Eq. (24) at $l=L$ with $d_{L+1}(a_{L+1},x) = d_{L+1}(x) = 1$, and ends at $l=0$ with Eq. (23).

With these quantities we can compute Eq. (20) as

$$\Pr(a_l,a_{l+1},I \mid x,\theta^t) = u_l(a_l,x)\tilde{d}_l(a_{l+1},x)\Pr(a_l \mid a_{l+1}), \quad (26)$$

$$\Pr(a_l,I \mid x,\theta^t) = u_l(a_l,x)d_l(a_l,x), \quad (27)$$

where the computations (21)–(27) in the E-step at iteration t are performed with fixed parameters θ^t.

3.3. Emission probabilities

The model described thus far uses the same labels A for modeling the distributions of the observables $\mathbf{G}$ as well as for propagating non-local information through the different scales. For the latter purpose it might be necessary to have many different possible label values that can encode for more complex information. In the levels of the pyramid the means and variances that are assigned to each label value may have very few pixels for training and therefore may be poorly estimated. It is thus reasonable to separate the functionality of the label A, for example as indicated in Fig. 6. In this case, labels A still code for the non-local information while labels B now are used for modeling the distribution of the features. Up to the conditioning on $\mathbf{F}$

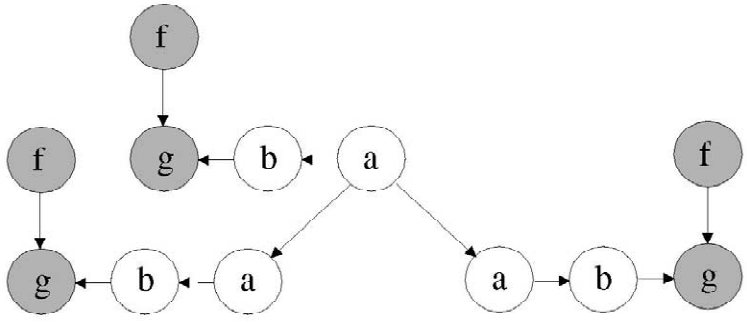


Fig. 6. Dependency structure of the HIP model with emission probabilities corresponding to Eq. (28).

this model now very closely resembles an HMM tree, with mixture densities as emission probabilities.

The expressions for the joint probability distributions as well as the corresponding probability propagation can be obtained by setting

$$\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l, x) = \sum_{b_l} \Pr(b_l \mid a_l) \Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, b_l, x) \tag{28}$$

in Eqs. (7) and (21).

Once again we parameterize $\Pr(b_l \mid a_l)$ as

$$\Pr(b_l \mid a_l) = \frac{\pi_{b_l, a_l}}{\sum_{b_l} \pi_{b_l, a_l}}. \tag{29}$$

The re-estimation equation in the M-step is then

$$\pi^{t+1}_{b_l, a_l} = \sum_x \Pr(b_l, a_l \mid I, x, \theta^t), \tag{30}$$

where we can use the joint

$$\Pr(b_l, a_l, I \mid x, \theta^t) = \frac{\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, b_l, x)}{\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l, x)} \Pr(b_l \mid a_l) \Pr(a_l, I \mid x, \theta^t). \tag{31}$$

3.4. Scale, mixture and hierarchy labels

An alternative to adding separate labels for mixture components as emission probabilities is to partition the labels in the simple model, in other words to view a label a, as in Fig. 4, as being composed of a label $m \in \{1, \ldots, N_m\}$ that specifies the mixture component and a hierarchy label $c \in \{1, \ldots, N_c\}$ that is intended to capture non-local information. We can relate these labels to each other in different ways, for example $a = (c-1)N_m + m$ while requiring $N_a = N_m N_c$. The mixture component at a location in the pyramid is given by m, whereas c influences image structure at finer scales through the model's conditional probability distribution $\Pr(a_l \mid a_{l+1})$. We can now choose a small value for N_m at low-resolution levels, and a larger value for N_c. Conversely, the only appropriate value for N_c at the finest-resolution level is one, since all information from other levels can be carried to m_l by a_{l+1}.

We can recover emission probabilities from this model by imposing a simplification, namely that $\Pr(a_l \mid a_{l+1}) = \Pr(m_l \mid c_l) \Pr(c_l \mid c_{l+1})$. In this case, N_c at the finest-resolution level should be greater than one.

We can go further and add more structure to the labels. In the models we use for mass detection, we further partition the mixture labels into a label m and a *scale* label z, so that

$$\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, m_l, z_l) = \mathcal{N}(\mathbf{g}_l, M_{m_l} \mathbf{f}_{l+1} + \sigma_{z_l} \bar{\mathbf{g}}_{m_l}, \sigma^2_{z_l} \Lambda_{m_l}). \tag{32}$$

The mixture components for a given value of m_l model the same type of image structure, but with means and variances of different magnitudes as determined by the scale factor σ_{z_l}. Such explicit representation of scale has been reported to be important in modeling natural image structure (Wainwright and Simoncelli, 1999; Wainwright et al., 2001; Romberg et al., 2001). As with these earlier models, we make the scale factors depend on their parents at lower-resolution levels, since the magnitude of wavelet coefficients tends to persist across pyramid level. To reduce the number of model parameters we chose to constrain the label probabilities so that

$$\Pr(a_l \mid a_{l+1}) = \Pr(m_l \mid c_{l+1}) \Pr(z_l \mid z_{l+1}, c_{l+1}) \Pr(c_l \mid c_{l+1}). \tag{33}$$

Note that there is ambiguity in this representation, since we can multiply $\bar{\mathbf{g}}_{m_l}$ by a factor λ and Λ_{m_l} by λ^2 for all m_l, and the mixture components will not change if we also multiply σ_{z_l} by λ^{-1} for all z_l. To remove the ambiguity we apply the constraint $\prod_{z_l} \sigma_{z_l} = 1$.

The M-step of the EM algorithm must be modified in this model, since we cannot solve for all of the parameters of $\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, m_l, z_l)$ analytically. We choose to use a generalized EM algorithm, in which the M-step is iterative. Since the E-step is more computationally intensive than the M-step, the increase in training time is relatively small. In the iterative M-step, we fix σ_{z_l} for all z_l and re-estimate the other mixture parameters to maximize the expected likelihood. We then fix these other parameters (that depend only on m_l) and re-estimate σ_{z_l} for all z_l. Within the M-step we alternate repeating these two substeps several times. In practice, two to four iterations usually is adequate.

Finally, we allow for rotations into the label structure, although the use of wavelets restricts us to rotations by multiples of 90°. We do this simply by requiring that for every value of m_l, there are three other values whose mean and covariance ($\bar{\mathbf{g}}_{m_l}$ and Λ_{m_l}) are related to those of the original by the three rotations of 90°, 180°, and 270°. The EM algorithm is easily modified to handle this. In the E-step we can compute expectations for each individual mixture component. For a given set of four components related by rotations, we first perform the appropriate inverse rotation on each of three of the expectation values, so they agree with the fourth component, and sum them. In the M-step we then update the parameters for this component from these sums, and copy the results to the other three components after rotating them, as appropriate.

3.5. Preprocessing and training methods

We divide the data set into training and test sets of approximately equal size and, for the mass detection, we construct a jackknife (i.e. 10 different random splits) so as to demonstrate the robustness of the results. We use a set of approximately orthogonal wavelets to decompose the intensity images into feature images (see Appendix B for details). Before applying the wavelet decomposition we wrap images at edges in order to obtain perfect reconstruc-

tion for the compression and synthesis. We crop images so that they are square with objects approximately centered.

We train the HIP model using the EM algorithm described in Section 3. The number of labels was chosen through a splitting procedure, using the minimum description length (MDL) cost criterion to compute the optimal model. The MDL cost is given as $-\log \Pr(I \mid H) + d/2 \log(N)$ (Rissanen, 1996; Deco and Obradovic, 1996), where $\Pr(I \mid H)$ is the probability of the training data under the model H, d is the number of parameters in H, and N is the number of images in the training set (note that the form of the HIP model makes it well suited for MDL model selection). For the splitting procedure, we begin with only one hidden label value at each level. We then duplicate each label along with its parameters, i.e., $\Pr(a_l \mid a_{l+1})$, $\bar{\mathbf{g}}_{a_l}$, M_{a_l}, and Λ_{a_l}, randomly perturbing the duplicate parameters. We re-train the new, larger model and compare its MDL cost with the previous model. We repeat this, successively duplicating labels, retraining, and evaluating the MDL cost, until the MDL cost increases. The model with the lowest MDL cost is then used in the applications presented below. Fig. 7 shows how the area under the ROC curve A_z for the test data tracks the MDL cost. Note that best performance tends to be for lowest costs, though the tracking is not monotonic.

Such an MDL-based training procedure is feasible, however it is computationally expensive. On a Sun Ultrasparc-2 workstation the entire splitting procedure required roughly two weeks of computer time. This is partly due to the large number of parameters being adjusted, 12,995 for the optimal model for the masses (see Appendix C for discussion on the number of parameters in the model). In spite of this, over-fitting does not appear to be a problem, as evidenced from our jackknife results presented below. We believe this is so because every location in each level of the wavelet decomposition provides examples for the parameters used to model the coefficients at that level. For example, as part of fitting the image distribution the model must fit the marginal distributions of the wavelet coefficients at each level. For this purpose there is one example per location in each image, so that there are many more effective examples than the number of images. Also, once the models are trained, there is minimal computational cost/overhead in applying them for detection, synthesis and compression.

4. Experimental results

In this section we report results for applying a HIP model, with complete scale, hierarchy and mixture labels, to mammographic image analysis, in particular detection of mammographic masses. As an experimental paradigm, we choose to demonstrate the utility of the approach for a dataset representing the output of the University of Chicago's CAD system for mass detection. This is a state-of-the-art mammographic screening system which includes a set of signal enhancement, pre-processing, rule-based and statistical-based classification schemes for detecting masses in digitized mammograms (Doi et al., 1993; Nishikawa et al., 1996; Giger et al., 2000). We choose this

[8] For example, the Digital Database for Screening Mammography (DDSM), the Mammographic Image Analysis Society (MIAS) database, and the Lawrence Livermore National Laboratories (LLNL)/University of California at San Francisco (UCSF) database.

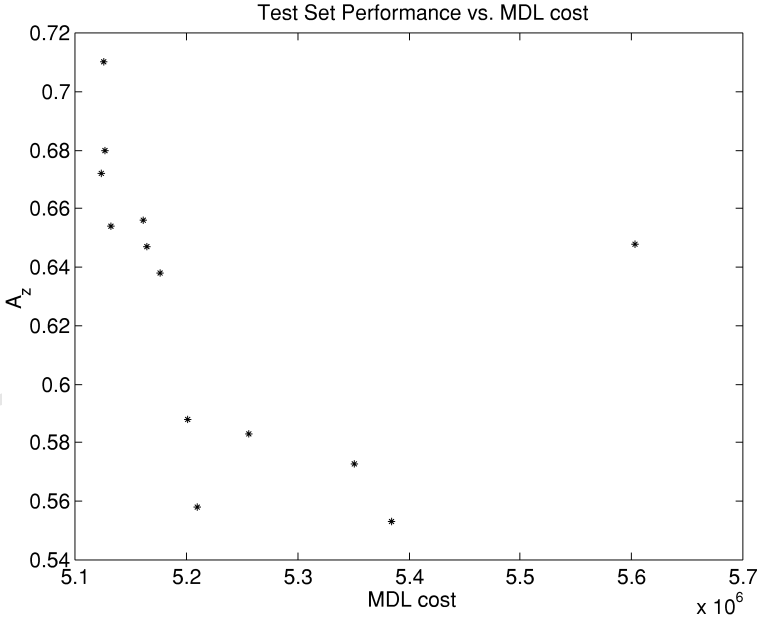


Fig. 7. Performance (area under the ROC curve, or A_z) versus MDL cost for the HIP model pair used to generate the experimental results.

paradigm over an alternative, such as performance on a public database of digitized mammograms,[8] since we can better estimate the clinical impact of the model in terms of reducing difficult false positives as well as demonstrating performance relative to a well-characterized clinical system. As an additional demonstration of the utility of HIP, we compare results to that of an HMT model using a single set of hidden labels to model two component mixtures over a wavelet tree. Details of this model can be found in Crouse et al. (1998) and Romberg et al. (2001). The comparison with the HMT enables us to characterize performance relative to another hierarchical probabilistic model for images, specifically the utility of the additional hidden label structure.

In the following we first describe the dataset used in the experiments and then present our results for classification, synthesis and compression.

4.1. Mammographic dataset

The dataset used in these experiments consists of regions of interest (ROIs) selected from digitized mammograms by the mammographic mass CAD system developed by the Rossmann Laboratories of the University of Chicago (UofC) (Nishikawa et al., 1995, 1996; Giger et al., 2000). The CAD system, consisting of a series of classification/detection modules, places an indicator (e.g., ''arrow'') next to potential masses on a digital image of the mammogram. The location of the indicator is determined by dividing the mammogram into ROIs and then eliminating false-positive ROIs using pattern recognition techniques. The output of the CAD system therefore can be seen as a set of ROIs, of which all are assumed to be positive for masses. Since this is a screening system, both malignant and benign masses are considered ''true positives''. ROIs output by the CAD system which do not contain masses (i.e. non-masses) are UofC false positives.

For the experiments in this paper, 169 ROIs were available, of which 72 contained masses (true positives) and 97 were false positives of the UofC CAD system. The detected objects (apparent masses) are not necessarily centered in the ROI, since they may lie close to the edge of the mammogram. The original ROIs are 512-by-512 pixels. Examples of mass and non-mass ROIs are shown in Fig. 8.

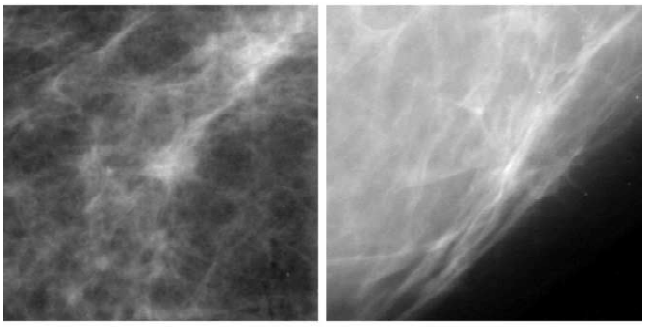


Fig. 8. Examples of data used in experiments. (A) Mammographic mass (true positive). (B) False positive generated by the UofC CAD system.

4.2. Mass detection

We first consider using HIP as a post-processor (i.e. adjunct) to the UofC CAD system (Nishikawa et al., 1996). The goal was to determine if the HIP model could be used to reduce false positives without reducing sensitivity. In addition, the performance of the HIP model was compared to an HMT. A 10-way jackknife was used to compute the results.

Two HIP models were trained for each of the jackknife sets. Each jackknife set consisted of 36 randomly chosen ROIs that contained masses, and 48 randomly chosen ROIs without masses. One model was trained for the mass ROIs and another model for the non-mass ROIs. Similarly, two HMTs were trained, using the same split of the data.

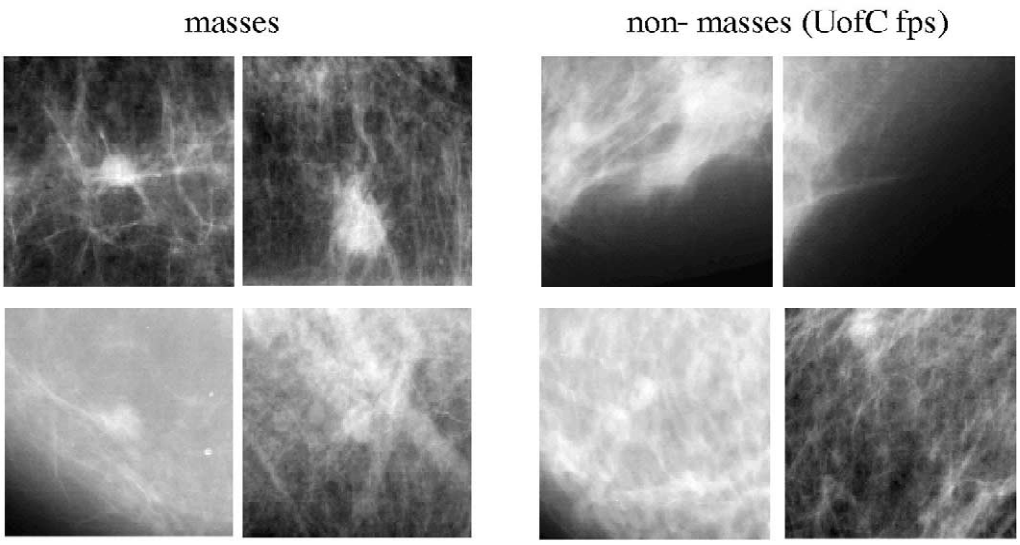
The likelihood ratio under the two models was used as the test criterion, i.e., a threshold on this ratio is used to decide which ROIs will be detected as masses. The true and false-positive fractions as a function of the threshold were measured on the split of the jackknife that contained the test set. This set also consisted of 36 mass and 49 non-mass ROIs.

Table 1 summarizes the results for the jackknife experiments. On average, the receiver operating characteristic (ROC) curve (Metz, 1988) for the HIP model applied to the test images has an area under the curve (A_z) equal to 0.78 and a 16% reduction in false positives generated by the UofC CAD system, without loss in sensitivity. By comparison the HMT model has a mean A_z equal to 0.55, with only a 3% reduction in false positives. Given the difficulty of this dataset (i.e. it represents the most difficult false positives that could not be eliminated by the UofC system) a 16% reduction in the false-positive rate is significant. Nonetheless, the HIP model is not capable of learning subtle differences for distinguishing between masses and non-masses. Fig. 9 shows examples of ROIs correctly and incorrectly classified by HIP. From these examples we see that the model trained to detect masses performs well for ROIs containing localized, and somewhat isolated, homogeneous ''mass-like'' structure. For non-masses (UofC false positives) the HIP model correctly characterizes ROIs devoid of mass-like structure, and in

Table 1
Jackknife results for mass detection

	HIP	HMT
Mean (std) A_z	0.78 (0.04)	0.55 (0.05)
Mean (std) FPF@100% TPF	0.84 (0.15)	0.97 (0.05)
Mean (std) FPF@95% TPF	0.73 (0.17)	0.93 (0.05)

FPF, false-positive fraction; TPF, true-positive fraction.



900

901 Fig. 9. Example of ROIs which the HIP model correctly (top row) and incorrectly (bottom row) characterizes. Note that the difference between the two
902 classes of ROI (mass versus non-mass) is much more apparent in the top row than in the bottom row, consistent with model performance.

903 fact learns that many of the non-mass false positives are in
904 fact at the breast border. For ROIs incorrectly character-
905 ized by HIP, we see a striking similarity in the ROI
906 structure for masses and non-masses, which is consistent
907 with the fact that the HIP model would have difficulty
908 putting these ROIs into one of the two classes. More
909 insight into the structure captured by the mass and non-
910 mass HIP models, and how it relates to detection per-
911 formance, can be seen through mammographic synthesis.

912 *4.3. Mammographic synthesis*

913 Since the HIP model is a generative model, we can
914 sample the model and synthesize new images. In the
915 context of ROI classification, synthesized images can
916 provide qualitative insight into what features the model is
917 extracting and representing for both positive and negative
918 ROIs. Using the same ROI database used for classification,
919 we constructed HIP models for positives (masses) and
920 negatives (no masses). The trained HIP models were
921 sampled to synthesize new ROI images. The sampling
922 procedure begins at the coarsest resolution, where the
923 hidden labels are randomly sampled from the distribution
924 $\Pr(A_L)$. The feature images $\mathbf{G}_L$ are then sampled from
925 $\Pr(\mathbf{G}_L \mid A_L)$. The $\mathbf{G}_L$ are used to construct I_{L-1}, from
926 which the $\mathbf{F}_L$ are constructed. We then sample A_{L-1} from
927 $\Pr(A_{L-1} \mid A_L)$, and then $\mathbf{G}_{L-1}$ from $\Pr(\mathbf{G}_{L-1} \mid \mathbf{F}_L, A_{L-1})$.
928 This is repeated until the finest resolution is reached and I_0
929 is constructed.

930 Fig. 10 shows examples of these images. Inspection of
931 the synthesized positive ROIs shows more focal structure,
932 with more well-defined borders and higher spatial fre-
933 quency content than the negative ROIs. Comparison to the
934 HMT synthesized images, constructed with a similar
935 sampling procedure, shows the HMT images for positive
936 and negative ROIs. Though positive and negative ROIs are
937 different, the positive ROI does not capture the focal
938 structure of the mass, as is the case for the HIP generated
939 images. This is likely due to the flexibility of the hidden
940 variable architecture of the HIP, with scale, mixture and
941 hierarchy labels able to capture more structure in the
942 image. As a test, we sampled a HIP model constructed
943 using only a single hidden label structure (similar to that of
944 an HMT). Fig. 10(C) shows that the focal structure has
945 now disappeared in the positive ROI.

946 It is equally important to consider the mammographic
947 structure that is not well represented in the synthesized
948 images. A comparison of Fig. 8(A) and Fig. 10(A) indi-
949 cates that the model is not accurately representing the
950 extended linear structure of the breast parenchyma. One
951 possible reason is that the tree structure of the model is not
952 ideal for capturing colinear dependencies across space,
953 since there is no direct conditioning between neighboring
954 nodes. Such dependencies can be captured only indirectly
955 via propagation up and down the tree.

956 *4.4. Mammographic image compression*

957 A stream of random variables can be optimally com-
958 pressed if we know their distribution. A HIP model of a
959 source of images should therefore allow us to compress
960 examples of those images with high efficiency. Here we
961 demonstrate compression with HIP and HMT models using
962 a simple technique.

963 Given an image and a HIP model, we compress the
964 image as follows. First, we compute the most likely value
965 of each hidden label, $a_l^*(x) = \arg\max_{a_l} \Pr(a_l, I \mid x, \theta^t)$,
966 using Eq. (27). These most likely values are then encoded
967 with arithmetic coders, which require a probability dis-
968 tribution for the symbols they are to encode. For this we

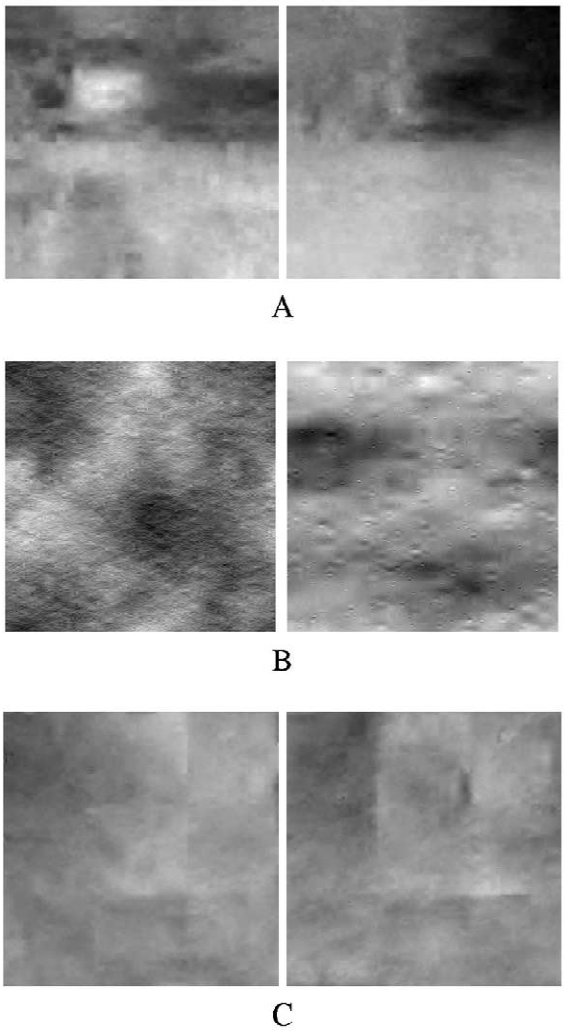


Fig. 10. Mammographic ROI images synthesized from positive and negative HIP and HMT models. (A) Synthesized ROIs from HIP model with scale, hierarchy and mixture labels. Positive ROIs (left) tend to have more focal structure, with more defined borders and higher spatial frequency content. Negative ROIs (right) tend to be more amorphous with lower spatial frequency content. (B) Synthesized ROIs from HMT model. Though ROIs of positive and negative models appear different, the positive ROI does not appear to capture the focal structure of masses. (C) HIP model with single label architecture. As with the HMT, this architecture does not capture the focal structure of the masses.

use the HIP model distributions $\Pr(a_l^*(x) \mid a_{l+1}^*(x))$. Given the label value $a_l^*(x)$, we then encode the feature vector $\mathbf{g}_l(x)$ using $\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l^*, x)$. The latter is used by decomposing $\mathbf{g}_l(x)$ into its components along the eigenvectors of the covariance matrix of $\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l^*)$, $\Lambda_{a_l^*}$. These components are independent under $\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l^*)$, so they can be encoded independently. Each component is encoded with a specified precision by dividing the real line into intervals of width equal to twice the precision. Using an arithmetic coder we then encode an index for the interval containing the component. The probability of each bin is provided by the integral of the univariate Gaussian distribution of the component implied by $\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l^*)$, i.e., with variance given by the corresponding eigenvalue of $\Lambda_{a_l^*}$. This procedure is computationally expensive, and is not necessarily optimal even if the HIP model exactly matches the image distribution, but it serves to demonstrate the capability. We used the analogous procedure for compression with the HMT models, except the DC residual needs to be stored separately, without compression.

We compress the images at several values of the precision. The range of precisions was chosen to roughly match the errors given by JPEG with a range of quality factors.

To compare with JPEG, we first convert the images to unsigned byte pixels. We divide the pixel values by four before compressing with JPEG, since the maximum value of the pixels in the mass ROIs is a little less than 1024, and multiply by four after decompressing. The results, averaged across all images, are shown in Fig. 11. The HIP model performs better then HMT for higher precision or lower loss, suggesting that the hidden labels capture useful information, allowing better compression. The better performance of HMT at higher compression ratios comes from our method for encoding the hidden labels. In our scheme these use the same number of bits no matter what the precision is, putting a lower limit on the compressed image size. The HMT model has fewer and simpler hidden values to be encoded. It still has a lower limit, but this is much smaller than for the HIP model. In fact, the HMT model performs better than JPEG at these high compression ratios. A more sophisticated compression algorithm with the HIP model would group labels, since some mixture components can become indistinguishable when coding at low precision. This grouping would effectively adjust the complexity of the HIP model with coding precision.

5. Discussion and conclusion

We have developed a class of multi-scale probabilistic network models for images which we call hierarchical image probability or HIP models. To justify these, we show that image distributions can be exactly represented as products over pyramid levels of distributions of sub-sampled feature images conditioned on coarser-scale image information. We argue that hidden variables are needed to capture long-range dependencies while allowing us to further factor the distributions over position. In our current model the hidden variables include scale, hierarchy and mixture labels which enable a more flexible modeling of natural images, compared to the structure of an HMT. This was demonstrated by comparison of the two approaches for mammographic mass detection, synthesis and compression, with the HIP model giving superior results. However, the current structure of HIP is not well suited for capturing dependencies between oriented spatial structure. Future work will investigate methods for more direct modeling of

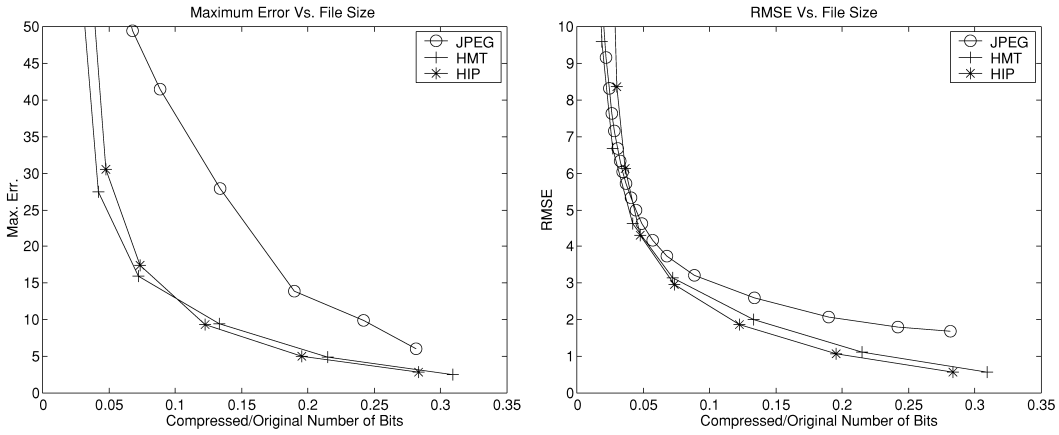


Fig. 11. Pixel error vs. size of compressed files for JPEG, HIP, and HMT. (Left) Maximum error (L_∞ norm). (Right) RMS error. These curves represent averages across all of the mammographic ROIs.

spatial orientation dependencies, which are obvious in the structure of mammograms and, in general, natural images.

Because HIP models are probability distributions over images, they can be used for a wide range of image processing tasks, e.g. classification, compression, noise suppression, up-sampling, error correction, etc. In fact, any image analysis problem can be approached in a principled way using such distributions. Here we have presented results for mammographic image analysis. However, there are obviously other modalities and medical application areas where HIP models would be useful. One in particular is multi-modal fusion, where the problem is to bring a set of images, acquired using different imaging modalities, into alignment. One method that has demonstrated particularly good performance uses mutual information as an objective criterion (Wells et al., 1996). The computation of mutual information requires an estimate of entropies, which in turn requires an estimate of the underlying densities of the images. The HIP model potentially provides a framework for learning those densities.

Some of the results we have obtained with the HIP model are comparable to those given by other approaches rather than being superior to them (e.g., for detection the HPNN (Sajda et al., 2002) gives similar if not better results). However, we obtain our results for several different problems using a single model, rather than training very different models for each problem. This flexibility and the principled approach provided by HIP models to image analysis are quite useful. We believe that, with further development, models of image probability distributions will give superior performance in a variety of medical image processing tasks.

Acknowledgements

We thank Robert Nishikawa and Maryellen Giger for useful discussions and providing the data and Adam Gerson for assistance in the simulations. We also thank the three anonymous reviewers for their helpful comments which greatly improved the manuscript. This work was funded by the U.S. Army Medical Research and Material Command (DAMD17-98-1-8061). P.S. was also supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under grant N00014-01-1-0625 as well as a grant from the National Imagery and Mapping Agency, NMA201-02-C0012.

Appendix A. Belief propagation in HIP

Here we show how to obtain the upwards and downwards propagation rules (21)–(24). All the computations can be executed locally. Consider the subgraph presented in Fig. A.1. In this subgraph, every node X can take on a discrete number of values, with Σ_X indicating a sum over those values. Assigned to every node X is also an evidence node g_X, with a fixed value for given image data. $g_X \ldots$ refers to g_X and all the evidence in the rest of the graph that can be reached through node X. Using this notation the entire evidence provided by the image I is the collection $\{g_A \ldots, g_B \ldots, g_C \ldots\}$. The probability required in the EM algorithm is

$$\Pr(B,A,I) = \Pr(B,A,g_A \ldots, g_B \ldots, g_C \ldots), \quad \text{(A.1)}$$

$$= \Pr(A,g_A \ldots, g_C \ldots)\Pr(B,g_B \ldots \mid A), \quad \text{(A.2)}$$

$$= \Pr(A,g_A \ldots, g_C \ldots)\Pr(B \mid A)\Pr(g_B \ldots \mid B), \quad \text{(A.3)}$$

$$= d_B(A)\Pr(B \mid A)u(B), \quad \text{(A.4)}$$

where in Eq. (A.4) the quantities $d_B(A)$ and $u(B)$ represent

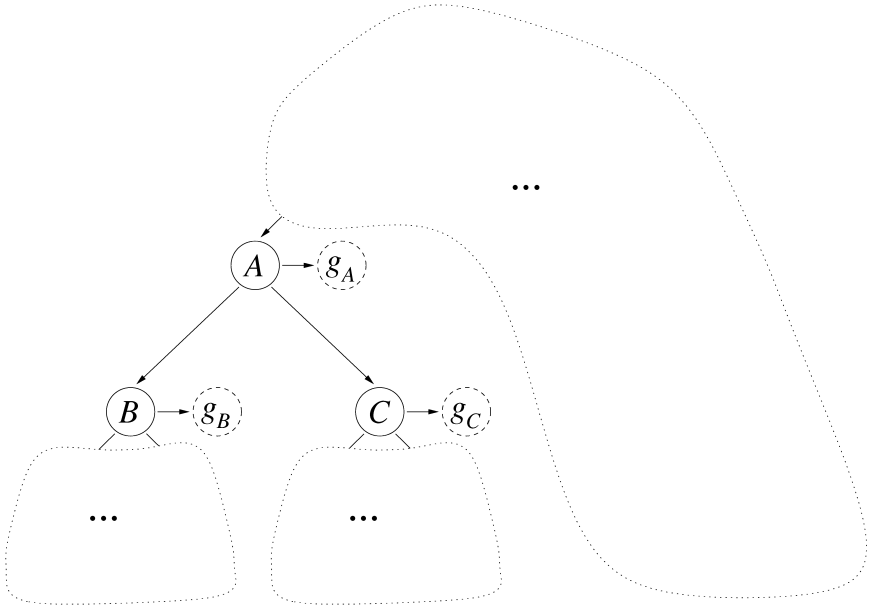


Fig. A.1. Subgraph of the label pyramid. Conditioned on A the variables that are connected to A become independent, such as labels B, C, and the evidence node g_A. These variables are also conditionally independent to the joint variables that can be reached going upwards to the rest of the tree structure.

the probabilistic influences on node B that are down-stream and up-stream from the node. In (A.2) we use the fact that conditioned on A the evidence coming through the children of A is independent from the rest of the tree beyond A. Since the children of A have no other parents, all the probabilistic influence beyond that parent edge can be communicated only through A. Similarly in (A.3) we use the fact that the evidence g_B is independent from the children of B if conditioned on B. Finally, in (A.4) we use the definitions for computing these probabilities recursively in an upwards and downwards probability propagation as follows:

$$u(A) \equiv \Pr(g_A, g_B \ldots, g_C \ldots \mid A), \tag{A.5}$$

$$= \Pr(g_A \mid A)\Pr(g_B \ldots \mid A)\Pr(g_C \ldots \mid A), \tag{A.6}$$

$$= \Pr(g_A \mid A)u_B(A)u_C(A)$$

$$= \Pr(g_A \mid A)\prod_{X \in C_A} u_X(A), \tag{A.7}$$

$$u_B(A) \equiv \Pr(g_B \ldots \mid A), \tag{A.8}$$

$$= \sum_B \Pr(B \mid A)\Pr(g_B \ldots \mid B), \tag{A.9}$$

$$= \sum_B \Pr(B \mid A)u(B). \tag{A.10}$$

We use in (A.6) and (A.9) conditional independence when conditioning on A and B, respectively. In (A.10) we use definition (A.5) for node B and in (A.7) we use definition (A.8) for the children of A. The downward propagating probability is defined and computed as

$$d_B(A) = \Pr(A, g_A \ldots, g_C \ldots), \tag{A.11}$$

$$= \Pr(g_C \ldots \mid A)\Pr(A, g_A \ldots), \tag{A.12}$$

$$= \frac{u(A)}{u_B(A)}d(A), \tag{A.13}$$

$$d(B) \equiv \Pr(B, g_A \ldots, g_C \ldots), \tag{A.14}$$

$$= \sum_A \Pr(B \mid A)\Pr(A, g_A \ldots, g_C \ldots), \tag{A.15}$$

$$= \sum_A \Pr(B \mid A)d_B(A). \tag{A.16}$$

Again, we use the conditional independences when conditioning on A in (A.12), (A.13), and (A.15). One can verify (A.13) by inserting the corresponding definitions and canceling the term $\Pr(g_A \mid A)$ to recover (A.12).

These upwards and downwards propagation equations are the basis for Eqs. (21)–(24).

Appendix B. Wavelets

For the HIP model presented in this paper we use approximately orthogonal wavelets with subsampling by two. The two-dimensional filters are separable, being products of one-dimensional wavelets. For the one-dimensional filters we solve for appropriate tap weights (i.e. filter coefficients) subject to the following constraints:

1. One filter is even-symmetric and low-pass (taps sum to one, zero response at the Nyquist frequency).
2. The second filter is odd-symmetric (high-pass).

Table B.1
Tap weights for 12-tap orthogonal wavelet filters with subsampling by two

Tap index	0	1	2	3	4	5
Tap weights	0.492631	0.060246	−0.060468	0.004588	0.002779	0.000223

3. The first derivatives of the responses of the filters are zero at zero frequency and the Nyquist frequency.
4. The second derivatives of the responses of the filters are zero at zero frequency and the Nyquist frequency.

We adjusted the tap weights to be as close to orthogonal as possible, by minimizing the mean-squared error of analysis and subsequent reconstruction of a noise image. No number of taps gave perfect reconstruction, and we decided that the error with 12 taps was acceptable. Note that the odd-symmetric high-pass filter has the same tap weights as the low-pass filter, except for an alternating sign. Numerical values for the tap weights of the filters are given in Table B.1.

For the HIP model we build a wavelet packet tree using the entropy minimization techniques developed by Saito (1994). We use half the data to compute a wavelet packet with minimal entropy, which is analogous to maximizing the sparsity of the wavelet coefficients. The wavelet packet that is constructed is shown in Fig. A.2. Note that from this representation one can see the dimensionality of $\mathbf{g}_l$ at each scale.

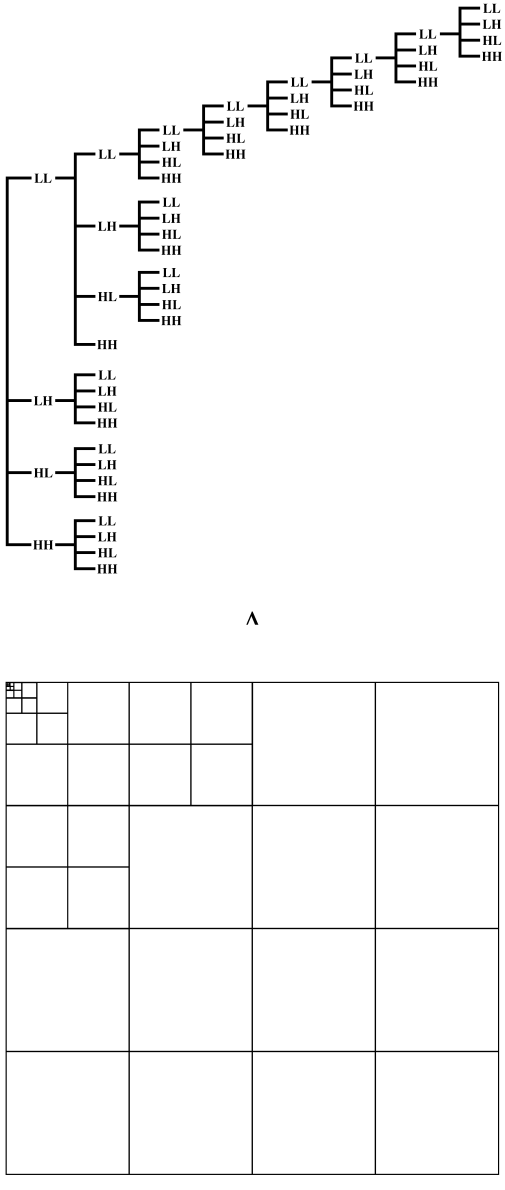


Fig. A.2. Wavelet packet constructed using entropy minimization. (A) Tree structure of wavelet packet with minimum entropy found using mammographic mass data. (B) Corresponding wavelet packet decomposition using conventional representation for images. All blocks of equal size make up the features at a given level, $\mathbf{g}_l$.

Appendix C. Parameters

Each mixture component (label m) has parameters $\bar{\mathbf{g}}$, Λ, and M. If $\mathbf{g}$ has dimension N_g and $\mathbf{f}$ has dimension N_f, then $\bar{\mathbf{g}}$ is N_g parameters, Λ is $N_g(N_g+1)/2$ parameters, and M is $N_g N_f$ parameters. At level 2, for example, $N_g = 13$ and $N_f = 12$, so $\bar{\mathbf{g}}$ is 13 parameters, Λ is 91 parameters, and M is 156 parameters, for a total of 260 parameters per value of the label m. These values for all levels in the MDL

Table C.1
Parameter counts for mixture components in HIP mass model

Level	N_g	N_f	Λ	M	Per comp.	No. comps.	Total
2	13	12	91	156	260	16	4160
3	11	4	66	44	121	16	1936
4	3	4	6	12	21	16	336
5	3	4	6	12	21	16	336
6	3	4	6	12	21	4	84
7	3	4	6	12	21	4	84
8	4	0	10	0	14	1	14
Total							6950

Table C.2
Parameter counts for scale components in HIP mass model

Level	N_z	Total
2	8	7
3	8	7
4	8	7
5	8	7
6	8	7
7	4	3
8	2	1
Total		39

Table C.3
Parameter counts for conditional probability distributions in HIP mass model

Level	N_c	N_m	N_z	$\Pr(c_l \mid c_{l+1})$	$\Pr(m_l \mid c_{l+1})$	$\Pr(z_l \mid z_{l+1}, c_{l+1})$	Total
2	1	16	8	0	2016	1792	3808
3	8	16	8	224	2016	1792	4032
4	8	16	8	224	2016	1792	4032
5	8	16	8	224	2016	1792	4032
6	8	4	8	224	480	896	1600
7	8	4	4	112	240	96	448
8	4	1	2	48	48	0	96
9	4	0	0	4	0	0	4
Total							18,052

optimal positive mass model are shown in Table C.1, along with totals. Table C.2 gives the number of parameters for the scale component label values z. Note that, because of the sum-to-one constraint on z, there is one fewer parameter than N_z.

The number of parameters in conditional probability distribution for labels is less than the product of the numbers of labels appearing in the distribution. This is partly due to the normalization, which reduces the count by one. In addition, for purposes of MDL model selection, we can argue that only non-zero probabilities should be counted. For coding purposes we could code a distribution $\Pr(a \mid b)$ for a given b as the number of values of a for which the distribution is non-zero, followed by the values for which it is non-zero, followed by the corresponding non-zero probabilities (except the last, which is given by the normalization). Asymptotically, as the precision with which we encode the probabilities increases, the cost of coding the integers becomes negligible because they are fixed, whereas the code length for the probabilities increases. In Table C.3 we list the maximum possible number of parameters for each level, that is, assuming no zero probabilities. Note that the number of mixture components is increased by a factor of four compared to Table C.1, due to the rotational symmetry we have imposed. (A set of four components related by rotations are specified by the same parameters, but we do not impose constraints on the label probabilities.) Also, since each non-leaf node in the tree has four children the number of parameters for one of the distributions is four times the number of labels at one level times the number of labels at the parent level. The exception is level 9, where there are four values of the label c_9 that condition the one child at level eight. Of the 18,052 parameters, 12,275 are zero.

We arrive at the total number of parameters in the model as the number of mixture components (6950) added to the number of scale labels (39) added to non-zero parameters for the conditional probabilities ($18{,}052 - 12{,}046 = 6006$) to get 12,995. Since we jackknife our data this represents a rough estimate on the number of parameters. There are a similar number of parameters for the non-mass models.

References

Bird, R., 1990. Professional quality assurance for mammography screening programs. Radiology 177, 8–10.

Buccigrossi, R.W., Simoncelli, E.P., 1998. Image compression via joint statistical characterization in the wavelet domain. Tech. Rep. 414, U. Penn. GRASP Laboratory, available at ftp://ftp.cis.upenn.edu/pub/eero/buccigrossi97.ps.gz.

Burhenne, L., Wood, S., D'Orsi, C., Feig, S., Kopans, D., O'Shaughnessy, K., Sickles, E., Tabar, L., Vyborny, C., Castellino, R., 2000. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 215, 554–562.

Burt, P.J., Adelson, E.H., 1983. The Laplacian pyramid as a compact image code. IEEE Trans. Commun. COM-31 (4), 532–540.

Chan, H., Sahiner, B., Lam, K.L., Petrick, N., Helvie, M.A., Goodsitt, M., Adler, D., 1998. Computerized analysis of mammographic microcalcifications in morphological and feature spaces. Med. Phys. 25, 2007–2019.

Chellappa, R., Chatterjee, S., 1985. Classification of textures using Gaussian Markov random fields. IEEE Trans. ASSP 33, 959–963.

Cheng, H., Bouman, C.A., 2001. Multiscale Bayesian segmentation using a trainable context model. IEEE Trans. Image Process. 10 (4), 511–525.

Coi, H., Baraniuk, R.G., 2001. Multiscale image segmentation using wavelet-domain hidden Markov models. IEEE Trans. Image Process. 10 (9), 1309–1321.

Cootes, T., Hill, A., Taylor, C., Haslam, J., 1994. The use of active shape models for locating structure in medical images. Image Vis. Comput. 12 (6), 355–366.

Cootes, T., Taylor, C., 2001. Statistical models of appearance for medical image analysis and computer vision. In: Sonka, M., Hanson, K. (Eds.). Medical Imaging 2001, Vol. 4322. SPIE Press, pp. 236–248.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley, New York.

Crouse, M.S., Nowak, R.D., Baraniuk, R.G., 1998. Wavelet-based statistical signal processing using hidden Markov models. IEEE Trans. Signal Process. 46 (4), 886–902.

Dayan, P., Abbott, L., 2002. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. MIT Press, Cambridge, MA.

De Bonet, J.S., Viola, P., 1998. Texture recognition using a non-parametric multi-scale statistical model. In: Conference on Computer Vision and Pattern Recognition. IEEE, pp. 641–647.

De Bonet, J.S., Viola, P., Fisher, J.W., 1998. Flexible histograms: a multiresolution target discrimination model. In: Zelnio, E.G. (Ed.). Proceedings of SPIE, Vol. 3371, pp. 519–530.

Deco, G., Obradovic, D., 1996. An Information-Theoretic Approach to Neural Computing. Springer, New York.

Dempster, N.M., Laird, A., Rubin, D.B., 1977. Maximum likelihood from

incomplete data via the EM algorithm. J. R. Statist. Soc. B 39, 185–197.

Doi, K., Giger, M., Nishikawa, R., Hoffmann, K., MacMahon, H., Schmidt, R., Chua, K., 1993. Digital radiography: a useful clinical tool for computer-aided diagnosis by quantitative analysis of radiographic images. Acta Radiol. 34, 426–439.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification, 2nd Edition. Wiley, New York.

Floyd, C., Lo, J., Yun, A., Sullivan, D., Kornguth, P., 1994. Prediction of breast cancer malignancy using an artificial neural network. Cancer 74, 2944–2948.

Freeman, W.T., Jones, T.R., Pasztor, E.C., 2002. Example-based super-resolution. IEEE Comput. Graph. Applic. 22 (2), 56–65.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. PAMI 6 (6), 194–207.

Giger, M., Huo, Z., Kupinski, M., Vyborny, C., 2000. Computer-aided diagnosis in mammography. In: Sonka, M., Fitzpatrick, J. (Eds.), Medical Image Processing and Analysis. Handbook of Medical Imaging, Vol. 2. SPIE Press, pp. 917–986.

Grenander, U., 1983. Tutorials in Pattern Synthesis. Brown University, Providence, RI.

Grenander, U., Chow, Y., Keenan, D., 1991. Hands: A Pattern Theoretic Study of Biological Shapes. Springer, New York.

Huo, Z., Giger, M., Vyborny, C., Wolverton, D., Schmidt, R., Doi, K., 1998. Automated computerized classification of malignant and benign mass lesions on digital mammograms. Acad. Radiol. 5, 155–168.

Jiang, Y., Nishikawa, R., Wolverton, D., Metz, C., Giger, M.L., Schmidt, R., Doi, K., 1996. Automated feature analysis and classification of malignant and benign microcalcifications. Radiology 198, 671–678.

Jordan, M.I. (Ed.), 1998. Learning in Graphical Models. NATO Science Series D: Behavioral and Brain Sciences, Vol. 89. Kluwer Academic.

Kopans, D., 1989. Breast Imaging. Lippincott, Philadelphia, PA.

Lo, J., Kim, J., Baker, J., Floyd, C., 1996. Computer-aided diagnosis of mammography using an artificial neural network: predicting the invasiveness of breast cancers from image features. In: Giger, M.L. (Ed.). Medical Imaging 1996: Image Processing, Vol. 2710. SPIE Press, pp. 725–732.

Luettgen, M.R., Willsky, A.S., 1995. Likelihood calculation for a class of multiscale stochastic models, with application to texture discrimination. IEEE Trans. Image Proc. 4 (2), 194–207.

Metz, C., 1988. Current problems in ROC analysis. In: Proceedings of the Chest Imaging Conference, Madison, WI, pp. 315–333.

Metz, C., Shen, J., 1992. Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis. Med. Decis. Making 12, 60–75.

Nishikawa, R., Schmidt, R., Osnis, R., Giger, M., Doi, K., Wolverton, D., 1996. Two-year evaluation of a prototype clinical mammographic workstation for computer-aided diagnosis. Radiology 201, 256.

Nishikawa, R.M., Haldemann, R.C., Papaioannou, J., Giger, M.L., Lu, P., Schmidt, R.A., Wolverton, D.E., Bick, U., Doi, K., 1995. Initial experience with a prototype clinical intelligent mammography workstation for computer-aided diagnosis. In: Loew, M.H., Hanson, K.M. (Eds.). Medical Imaging 1995, Vol. 2434. SPIE, Bellingham, WA, pp. 65–71.

Portilla, J., Simoncelli, E., 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. Int. J. Comput. Vis. 40 (1), 49–71.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–285.

Rissanen, J., 1996. Information theory and neural nets. In: Smolensky, Mozer, Rumelhart (Eds.), Mathematical Perspectives on Neural Networks. pp. 567–602.

Romberg, J.K., Coi, H., Baraniuk, R.G., 2001. Bayesian tree-structured image modeling using wavelet domain hidden Markov models. IEEE Trans. Image Process. 10 (7), 1056–1068.

Saito, N., 1994. Local feature extraction and its applications using a library of bases. Tech. Rep., Ph.D.Thesis (Advisor: Prof. R.R. Coifman), Department of Mathematics, Yale University.

Sajda, P., Spence, C., Pearson, J., 2002. Learning contextual relationship in mammograms using a hierarchical pyramid neural network. IEEE Trans. Med. Imaging 21 (3), 239–250.

Thurfjell, E., Lernevall, K., Taube, A., 1994. Benefit of independent double reading in a population-based mammography screening program. Radiology 191, 241–244.

Wainwright, M., Simoncelli, E., Willsky, A., 2001. Random cascades on wavelet trees and their use in analyzing and modeling natural images. Appl. Comput. Harmonic Anal. 11, 89–123.

Wainwright, M.J., Simoncelli, E.P., 1999. Scale mixtures of Gaussians and the statistics of natural images. In: Solla, S.A., Leen, T., Müller, K.-R. (Eds.). Advances in Neural Information Processing Systems, Vol. 12. MIT Press, Cambridge, MA, pp. 855–861.

Wells, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multi-modal volume registration by maximization of mutual information. Med. Image Anal. 1 (1), 35–51.

Zhang, W., Doi, K., Giger, M.L., Wu, Y., Nishikawa, R.M., Schmidt, R., 1994. Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. Med. Phys. 21 (4), 517–524.

Zhu, S.C., Wu, Y.N., Mumford, D., 1997. Minimax entropy principle and its application to texture modeling. Neural Comput. 9 (8), 1627–1660.