Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model

Clay Spence
Vision Technologies
Sarnoff Corporation
Princeton, NJ 08540

Lucas Parra
Vision Technologies
Sarnoff Corporation
Princeton, NJ 08540

Paul Sajda
Biomedical Engineering
Columbia University
New York, NY 10023

Abstract

We develop a probability model over image spaces and demonstrate its broad utility in mammographic image analysis. The model employs a pyramid representation to factor images across scale and a tree-structured set of 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 EM algorithm. The utility of the model is demonstrated for three applications; 1) detection of mammographic masses in computer-aided diagnosis 2) qualitative assessment of model structure through mammographic synthesis and 3) lossless compression of mammographic regions of interest.

1. Introduction

In mammographic computer-assisted diagnosis (CAD) one typically estimates $\Pr(C|I)$, the conditional probability of class C (e.g. benign vs. malignant) given image I or a set of features extracted from I. Previous efforts have concentrated on the development of such *discriminant* models for CAD [1][2][3][4][**?**]. By contrast, a *generative* model, $\Pr(I|C)$, has many attractive features. Classification is possible by training a distribution for each class and using Bayes' rule to obtain $\Pr(C|I) = \Pr(I|C)\Pr(C)/\Pr(I)$. However there are many other benefits of having a model of the distribution of images, since any type of image analysis can be approached using knowledge of the distribution of the data. For example, anomalous images can be detected and rejected, rather than trusting the classifier's output. A generative model can also be used to compress, interpolate, suppress noise, increase or extend resolution, and fuse multiple images.

In the computer vision and pattern recognition community there has been limited work directed at developing probabilities for images. One of the few examples of image distribution models is that constructed by Zhu, Wu and Mumford[5]. In their approach they compute the maximum entropy distribution given a set of statistics across a number of features. Though this approach works well for textures, it is not clear how well it will model the appearance of more structured objects. Several algorithms have investigated modeling the distributions of features extracted from the image, instead of the image itself. The Markov Random Field (*MRF*) models are one such example; see, e.g., References [6, 7]. However, these models tend to be computationally expensive.

Recently, De Bonet and Viola's proposed a flexible histogram approach[8, 9], 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 then 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[10], 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.

All of 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. For

example, in mammography, benign and malignant masses tend to be characterized by a combination of texture and shape features[11] and may also include contextual influences. Therefore local conditioning, like that of the flexible histogram and MSP approaches, is inadequate.

Recently, several groups have developed what are essentially extensions of the MSP models 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 developed their Hidden Markov Tree (*HMT*) models [12] 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 coefficent 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 denoising and texture segmentation. Cheng and Bouman [13] applied another model of this sort for segmentation, in which the observed class labels play the role of hidden variables, and so of course are no longer hidden.

We have independently developed a class of models for probability distributions of images that we call hierarchical image probability (*HIP*) models. These also have tree-structured graph of the dependencies between hidden variables at different scales, and use mixtures of multivariate Gaussians to model the local distributions of vectors of features. In the following we present the basic HIP models, along with EM algorithm for training the models. We show preliminary results of the application of HIP models to mammographic image analysis, including lesion classification, mammographic synthesis and compression of mammographic ROIs.

2 Coarse-To-Fine Factoring Of Image Distributions

Our goal will be to write the image distribution in a form similar to $\Pr(I) \sim \Pr(\mathbf{F}_0 \mid \mathbf{F}_1) \Pr(\mathbf{F}_1 \mid \mathbf{F}_2) \ldots$, where $\mathbf{F}_l$ is the set of feature images at pyramid level l. We expect that the short-range dependencies can be captured by the model's distribution of individual feature vectors, while the long-range dependencies can be captured at low resolution. The large-scale structures affect finer scales by the conditioning.

We first prove that a coarse-to-fine factoring like this is correct. From an image I we build a Gaussian pyramid (repeatedly blur-and-subsample, with a Gaussian filter). Call the l-th level I_l, e.g., the original image is I_0 (Figure 1). From each Gaussian level I_l we extract a set of feature images $\mathbf{F}_l$. Sub-sample these to get feature images $\mathbf{G}_l$. Note that the images in $\mathbf{G}_l$ have the same dimensions as I_{l+1}. We denote by $\tilde{\mathbf{G}}_l$ the set of images containing I_{l+1} and the images in $\mathbf{G}_l$. We further denote the mapping from I_l to $\tilde{\mathbf{G}}_l$ by $\tilde{\mathcal{G}}_l$.

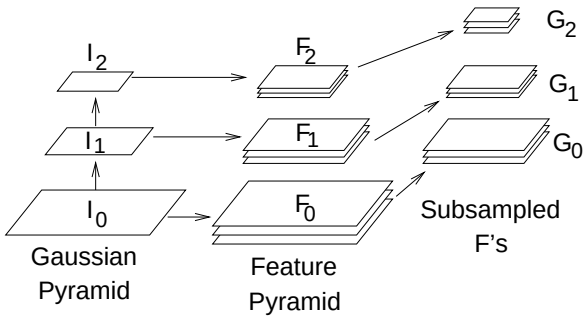

Figure 1: Pyramids and feature notation.

Suppose that $\tilde{\mathcal{G}}_0 : I_0 \mapsto \tilde{\mathbf{G}}_0$ is invertible. Then we can think of $\tilde{\mathcal{G}}_0$ as a change of variables. If we have a distribution on a space, its expressions in two different coordinate systems are related by multiplying by the Jacobian. In this case we get $\Pr(I_0) = |\tilde{\mathcal{G}}_0| \Pr(\tilde{\mathbf{G}}_0)$. Since $\tilde{\mathbf{G}}_0 = (\mathbf{G}_0, I_1)$, we can factor $\Pr(\tilde{\mathbf{G}}_0)$ to get $\Pr(I_0) = |\tilde{\mathcal{G}}_0| \Pr(\mathbf{G}_0 \mid I_1) \Pr(I_1)$. If $\tilde{\mathcal{G}}_l$ is invertible for all $l \in \{0, \ldots, L-1\}$ then we can simply repeat this change of variable and factoring procedure to get

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

This is a very general result, valid for all $\Pr(I)$, with some rather weak restrictions to make the change of variables valid. The restriction that $\tilde{\mathcal{G}}_l$ be invertible is strong, but many such feature sets are known to exist, e.g., most wavelet transforms on images.

3 The Need For Hidden Variables

For the sake of tractability we want to factor $\Pr(\mathbf{G}_l \mid I_{l+1})$ over positions, for example

$$\Pr(I) \sim \prod_l \prod_{x \in I_{l+1}} \Pr\big(\mathbf{g}_l(x) \mid \mathbf{f}_{l+1}(x)\big)$$

where $\mathbf{g}_l(x)$ and $\mathbf{f}_{l+1}(x)$ are the feature vectors at position x. The dependence of $\mathbf{g}_l$ on $\mathbf{f}_{l+1}$ expresses the persistence of image structures across scale, e.g., an edge is usually detectable as such in several neighboring pyramid levels. The flexible histogram and MSP methods share this structure.

While it may be plausible that $\mathbf{f}_{l+1}(x)$ has a strong influence on $\mathbf{g}_l(x)$, a model distribution with this factorization

and conditioning cannot capture some properties of real images. Objects in the world cause correlations and non-local dependencies in images. For example, the presence of a particular object might cause a certain kind of texture to be visible at level l. Usually local features $\mathbf{f}_{l+1}$ by themselves will not contain enough information to infer the object's presence, but the entire image I_{l+1} at that layer might. Thus $\mathbf{g}_l(x)$ is influenced by more of I_{l+1} than the local feature vector.

Similarly, objects create long-range dependencies. For example, an object class might result in a specific kind of texture across a large area of the image (e.g. malignant breast masses tend to have inhomogenous region enhancement). If an object of this class is always present, the distribution may factor, but if such objects are not always present and cannot be inferred from lower-resolution information, the presence of the texture at one location affects the probability of its presence elsewhere.

To capture these long-range dependencies we introduce hidden variables to represent the non-local information that is not captured by local features. These hidden varibles also constrain the variability of features at the next finer scale. Denoting the hidden variables collectively by A, we assume that conditioning on A allows the distributions over feature vectors to factor. In general, the distribution over images becomes

$$\Pr(I) \propto \sum_A \Big\{ \prod_{l=0}^{L} \prod_{x \in I_{l+1}} \Pr\big(\mathbf{g}_l(x) \,\big|\, \mathbf{f}_{l+1}(x), A\big) \times \Pr(A \,|\, I_{L+1}) \Big\} \Pr(I_{L+1}). \quad (2)$$

This is a very general form for A and we instead would like to be more specific. In particular we would like to preserve the conditioning of higher-resolution information on coarser-resolution information, and the ability to factor over positions. This lead to the following structure for our HIP model:[1]

$$\Pr(I) \propto \sum_{A_0,\ldots,A_L} \prod_{l=0}^{L} \prod_{x \in I_{l+1}} \Big[\Pr(\mathbf{g}_l \,|\, \mathbf{f}_{l+1}, a_l, x) \times \Pr(a_l \,|\, a_{l+1}, x) \Big] \quad (3)$$

To each position x at each level l we attach a hidden discrete index or label $a_l(x)$. The resulting label image A_l for level l has the same dimensions as the images in $\tilde{\mathbf{G}}_l$.

[1] In principle there is also a factor of $\Pr(I_{L+1})$. In many cases I_{L+1} will be a single pixel that is approximately the mean brightness in the image. We ignore this, which is equivalent to assuming that $\Pr(I_{L+1})$ is flat over some range. In this case $\mathbf{f}_{L+1}$ is zero for typical features. In addition, there is no hidden variable a_{L+1}. If we combine these considerations we see that the $l = L$ factor should be read as $\prod_x \Pr(\mathbf{g}_L \,|\, a_L, x) \Pr(a_L, x)$.

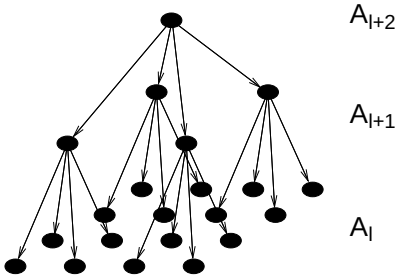

Figure 2: Quadtree structure of the conditional dependency between hidden variables in the HIP model.

Since $a_l(x)$ codes non-local information we can think of the labels A_l as a learned segmentation at the l-th pyramid level. By conditioning $a_l(x)$ on $a_{l+1}(x)$, we mean that $a_l(x)$ is conditioned on a_{l+1} at the *parent* pixel of x. This parent-child relationship follows from the sub-sampling operation. For example, if we sub-sample by two in each direction to get $\mathbf{G}_l$ from $\mathbf{F}_l$, we condition the variable a_l at (x, y) in level l on a_{l+1} at location $(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor)$ in level $l+1$ (Figure 2). This gives a tree structure to the dependency graph of the hidden variables, i.e. a belief network. By conditioning child labels on their parents information propagates though the layers to other areas of the image while accumulating information along the way.

For simplicity we have chosen $\Pr(\mathbf{g}_l \,|\, \mathbf{f}_{l+1}, a_l)$ to be normal with a mean that depends linearly on $\mathbf{f}_{l+1}$,

$$\Pr(\mathbf{g} \,|\, \mathbf{f}, a) = \mathcal{N}(\mathbf{g}, M_a \mathbf{f} + \bar{\mathbf{g}}_a, \Lambda_a) \quad (4)$$

4 EM Algorithm

Due to the tree structure, the belief network for the hidden variables is relatively straightforward to train with an EM algorithm. The expectation step (summing over a_l's) can be performed directly.[2] The expectation is weighted by the probability of a label or a parent-child pair of labels given the image. This can be computed in a fine-to-coarse-to-fine procedure, i.e. working from leaves to the root and then back out to the leaves. The method is based on belief propagation [14].

Once the expectations are computed, the normal distribution makes the M-step tractable; one simply computes the updated $\bar{\mathbf{g}}_{a_l}$, Σ_{a_l}, M_{a_l}, and $\Pr(a_l \,|\, a_{l+1})$ as combinations of various expectation values.

In order to apply the EM algorithm, a parameterization for the model is required. The parameterization of $\Pr(\mathbf{g} \,|\, \mathbf{f}, a)$ is given above in Equation 4. For $\Pr(a_l \,|\, a_{l+1})$

[2] Note that a more densely-connected structure, with each child having several parents, we have required either an approximate algorithm or Monte Carlo techniques.

we use the parameterization

$$\Pr(a_l \mid a_{l+1}) = \frac{\pi_{a_l,a_{l+1}}}{\sum_{a_l} \pi_{a_l,a_{l+1}}} \tag{5}$$

in order to ensure proper normalization.

Below, we denote the new parameter values computed during the t-th maximization step as θ^{t+1} and the old values as θ^t.

4.1 Maximization

Maximizing the expectation of the likelihood over the hidden variables with respect to the model parameters gives the following update formulae:

$$\pi^{t+1}_{a_l,a_{l+1}} = \sum_x \Pr(a_l, a_{l+1}, x|I, \theta^t), \tag{6}$$

$$\begin{aligned} M^{t+1}_{a_l} =& \left(\left\langle \mathbf{g}_l \mathbf{f}^T_{l+1} \right\rangle_{t,a_l} - \left\langle \mathbf{g}_l \right\rangle_{t,a_l} \left\langle \mathbf{f}^T_{l+1} \right\rangle_{t,a_l} \right) \\ & \times \left(\left\langle \mathbf{f}_{l+1} \mathbf{f}^T_{l+1} \right\rangle_{t,a_l} - \left\langle \mathbf{f}_{l+1} \right\rangle_{t,a_l} \left\langle \mathbf{f}^T_{l+1} \right\rangle_{t,a_l} \right)^{-1}, \end{aligned} \tag{7}$$

$$\bar{\mathbf{g}}^{t+1}_{a_l} = \left\langle \mathbf{g}_l \right\rangle_{t,a_l} - M^{t+1}_{a_l} \left\langle \mathbf{f}_{l+1} \right\rangle_{t,a_l}, \tag{8}$$

and

$$\begin{aligned} \Lambda^{t+1}_{a_l} = & \left\langle \left(\mathbf{g}_l - M^{t+1}_{a_l} \mathbf{f}_{l+1} \right) \left(\mathbf{g}_l - M^{t+1}_{a_l} \mathbf{f}_{l+1} \right)^T \right\rangle_{t,a_l} \\ & - \bar{\mathbf{g}}^{t+1}_{a_l} \bar{\mathbf{g}}^{t+1\,T}_{a_l}. \end{aligned} \tag{9}$$

Here the brackets $\langle . \rangle_{t,a_l}$ denote the expectation value

$$\left\langle X \right\rangle_{t,a_l} = \frac{\sum_x \Pr(a_l, x \mid I, \theta^t) X(x)}{\sum_x \Pr(a_l, x \mid I, \theta^t)}. \tag{10}$$

4.2 Expectation

In the E-step we need to compute the probabilities of pairs of labels from neighboring layers $\Pr(a_l, a_{l+1}, x_l \mid I, \theta^t)$ for given image data. Note that in all occurrences of the reestimation equations, i.e. (5,6) and (10), we require that quantity only up to an overall factor. We can choose that factor to be $\Pr(I|\theta^t)$ and can compute $\Pr(a_l, a_{l+1}, x_l, I|\theta^t)$ instead using

$$\begin{aligned} \Pr(a_l, a_{l+1}, x \mid I, \theta^t) \Pr(I \mid \theta^t) &= \Pr(a_l, a_{l+1}, x, I \mid \theta^t) \\ &= \sum_{A \setminus a_l(x), a_{l+1}(x)} \Pr(I, A|\theta^t) \end{aligned} \tag{11}$$

The computation of these quantities can be cast as recursion formulae, defined in terms of quantities u and d, which approximately represent upwards and downwards propagating probabilities. The recursion formulae are

$$\begin{aligned} u_l(a_l, x) = & \Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l, x) \\ & \times \prod_{x' \in \mathsf{Ch}(x)} \tilde{u}_{l-1}(a_l, x') \end{aligned} \tag{12}$$

$$\tilde{u}_l(a_{l+1}, x) = \sum_{a_l} \Pr(a_l \mid a_{l+1}) u_l(a_l, x) \tag{13}$$

$$d_l(a_l, x) = \sum_{a_{l+1}} \Pr(a_l \mid a_{l+1}) \tilde{d}_l(a_{l+1}, x) \tag{14}$$

$$\begin{aligned} \tilde{d}_l(a_{l+1}, x) = & \frac{u_{l+1}(a_{l+1}, \mathsf{Par}(x))}{\tilde{u}_l(a_{l+1}, x)} \\ & \times d_{l+1}(a_{l+1}, \mathsf{Par}(x)) \end{aligned} \tag{15}$$

The upward recursion relations (12–13) are initialized at $l = 0$ with $u_0(a_0, x) = \Pr(\mathbf{g} \mid \mathbf{f}_1, a_0, x)$ and end at $l = L$. At level L Equation 13 reduces to $\tilde{u}_L(a_{L+1}, x) = \tilde{u}_L(x)$.[3] Since we do not model any further dependencies beyond layer L, the pixels at layer L are assumed independent. Considering the definition of u, it is evident that the product of all $\tilde{u}_L(x)$ coincides with the total image probability,

$$\Pr(I|\theta^t) = \prod_{x \in I_L} \tilde{u}_L(x) = u_{L+1}. \tag{16}$$

The downward recursion (14 - 15) can be executed, starting with equation (15) at $l = L$ with $d_{L+1}(a_{L+1}, x) = d_{L+1}(x) = 1$.[3] The downwards recursion ends at $l = 0$ with equation (14).

We can now compute (11) as

$$\begin{aligned} \Pr(a_l, a_{l+1}, x, I \mid \theta^t) = & u_l(a_l, x) \tilde{d}_l(a_{l+1}, x) \\ & \times \Pr(a_l | a_{l+1}) \end{aligned} \tag{17}$$

$$\Pr(a_l, x, I \mid \theta^t) = u_l(a_l, x) d_l(a_l, x) \tag{18}$$

Computations (12–18) in the E-step at iteration t are done with fixed parameters θ^t.

5 Experimental Results

In this section we report some of our preliminary results for applying the HIP model to mammographic image analysis.

5.1 Mass Detection

To demonstrate utility, we use HIP as a post-processor (i.e. adjunct) to the University of Chicago's (UofC) CAD system[15]. False positive and true positive regions of interest (ROIs) were output from the UofC CAD system and

[3] The (non-existent) label a_{L+1} can be thought of as a label with a single possible value, which is always set. The conditional $\Pr(a_L \mid a_{L+1})$ turns then into a prior $\Pr(a_L)$

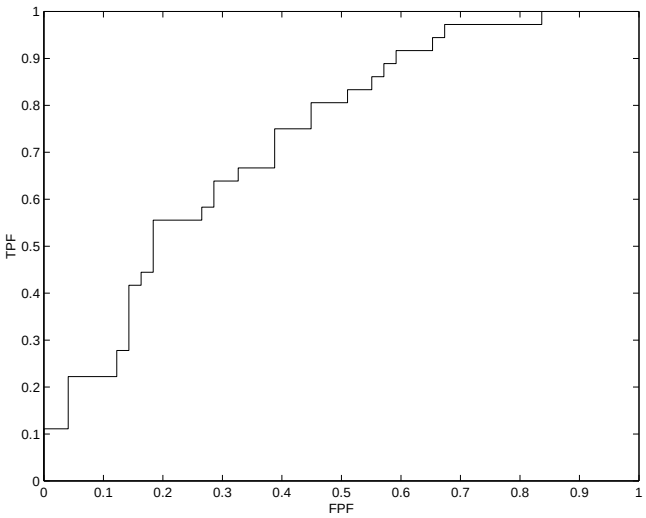

Figure 3: ROC curve for results of HIP models used as a post-processor for mass detection in the University of Chicago's mammographic CAD system.

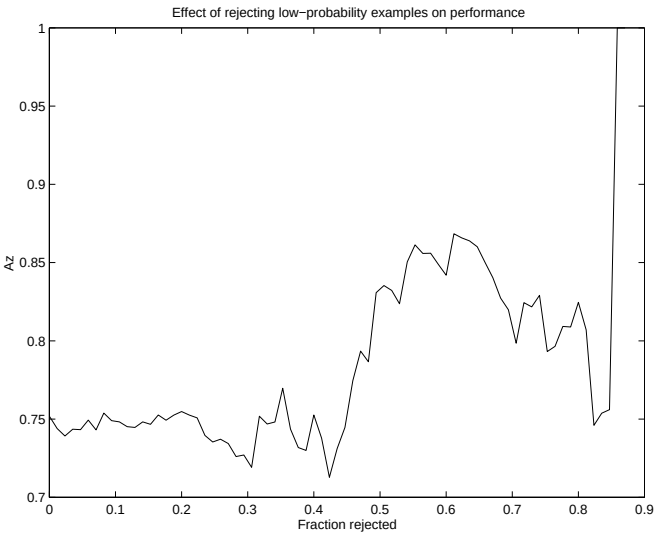

Figure 4: Using the HIP model for novelty detection and generating confidence measures. Thresholding the absolute value of the likelihood (abscissa) enables rejection of a fraction of the data that is novel, relative to the data on which the models were trained. This acts as a confidence measure, which can improve the performance of the model (Az values on ordinate axis).

used for training and testing. The goal was to determine if the HIP model could be used to reduce false positives without significant loss in sensitivity.

Two HIP models were trained; one using 36 randomly-chosen ROIs that contained masses, and a second trained on 48 randomly-chosen ROIs without masses. 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 rates as a function of the threshold were measured on a novel test set consisting of 36 mass and 49 non-mass ROIs.

A search was performed over the number of hidden labels values at each level. The search criterion used the negative log-likelihood on the training data plus the minimum-description-length penalty term, $d \log(N)/2$, where d is the number of model parameters and N is the the number of training examples [16]. The maximum number of labels in a level was bounded at 17.

The best perfroming model had an architecture of 17, 17, 11, 2, and 1 hidden label in levels 0–4, respectively. The receiver operatin characteristic (ROC) curve [17] for the test images is shown in Figure 3. For this architecture the area under the curve (A_z) was 0.75. For this architecture and set of parameter the HIP model is able to eliminate 17% of the false positives generated by the UofC CAD system, without loss in sensitivity.

5.2 Novelty Detection

Novelty detection identifies examples that are significantly different from the examples on which the model(s) was trained [18]. Detecting novel examples can be useful in a CAD system for generating confidence measures on the CAD output and identifying data that could be used in future training of the model. The HIP model's generative structure enables novel examples to be identified by thresholding the log-likelihood of the models. Figure 4 illustrates how ROC performance improves if novelty detection is used to generate a confidence measure for rejecting low-confidence examples. In this example, two HIP models were trained, one for positive ROIs and one for negatives ROIs (same ROI database as for classification). Test data was evaluated by computing the likelihood ratio of the models as well as the absolute value of the log-likelihoods. The absolute value of the log-likelihoods are thresholded such that low values are considered low confidence and therefore rejected (not classified). As the threshold on the log-likelihood is increased, more ROIs are rejected because of low confidence and the area under the ROC curve increases.

5.3 Mammographic Synthesis

Since the HIP model is a generative model, we can sample the model and synthesize new images. In the context of ROI classification, synthesized images can provide qualitative insight into what features the model is extracting and representing for both positive and negative ROIs. Using the same

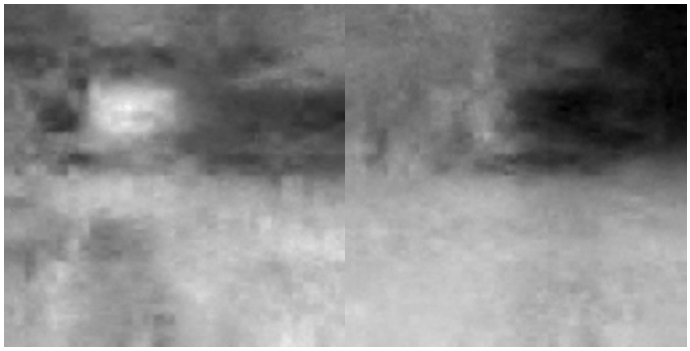

Figure 5: Mammographic ROI images synthesized from positive and negative HIP models. Synthesized positive ROIs (left) tend to have more focal structure, with more defined borders and higher spatial frequency content. Negative ROIs (right) tend to be more amorphous with lower spatial frequency content.

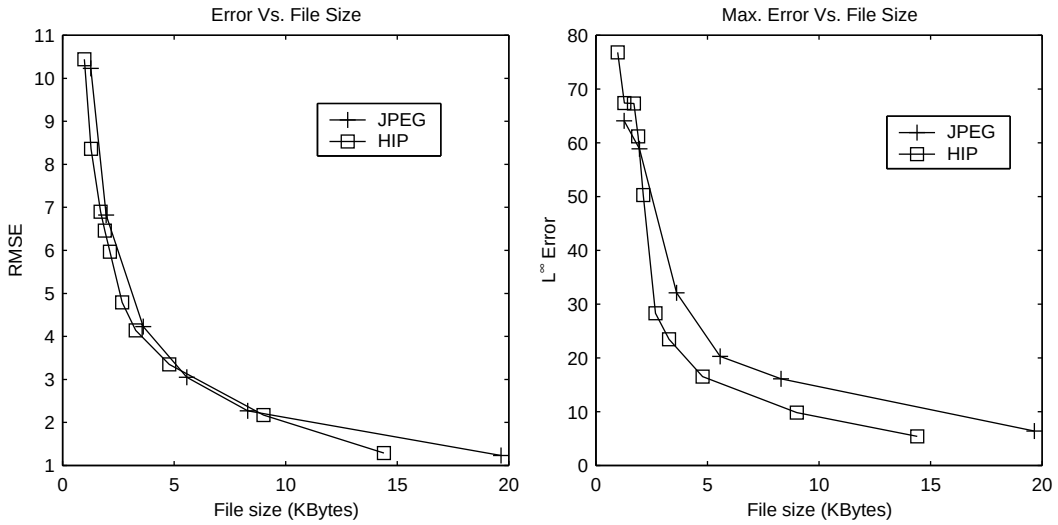

Figure 6: (Left)Root mean-squared error vs. size of compressed file, JPEG and HIP. (Right) Maximum error (L^∞ norm) vs. size of compressed file, JPEG and HIP.

ROI database used for classification, we constructed HIP models for positives (masses) and negatives (no masses). The trained HIP models were sampled to synthesize new ROI images. The sampling procedure begins at the coarsest resolution, where the hidden labels are randomly sampled from the distribution $\Pr(A_L)$. The feature images $\mathbf{G}_L$ are then sampled from $\Pr(\mathbf{G}_L \mid A_L)$. The $\mathbf{G}_L$ are used to construct I_{L-1}, from which the $\mathbf{F}_L$ are constructed. We then sample A_{L-1} from $\Pr(A_{L-1} \mid A_L)$, and then $\mathbf{G}_{L-1}$ from $\Pr(\mathbf{G}_{L-1} \mid \mathbf{F}_L, A_{L-1})$. This is repeated until the finest resolution is reached and I_0 is constructed.

Figure 5 shows examples of these images. Inspection of the synthesized positive ROIs shows more focal structure, with more well-defined borders and higher spatial frequency content than the negative ROIs.

5.4 Mammographic Image Compression

A stream of random variables can be optimally compressed if we know their distribution, and so having a HIP model of a source of images should allow us to compress examples of those images with high efficiency. Here we demonstrate compression with HIP models using a simple technique.

Given an image and a HIP model, we compute the most likely value of each hidden label, $a_l^*(x) = \arg\max_{a_l(x)} \Pr(a_l, x, I \mid \theta^t)$ using Equation 18, and code each 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 of $\Pr(\mathbf{g}_l \mid \mathbf{f}_{l+1}, a_l^*, x)$, $\Sigma_{a_l^*}$, and coding those components with a specified precision using Huffman encoders for the Gaussian distributions with variances given by the eigenvalues of $\Sigma_{a_l^*}$. The resulting bitstream was stored in a file that was susbequently

Figure 7: Compression artifacts of JPEG and HIP. Left: Original image, center: JPEG, right: HIP.

compressed with gzip to reduce the redundancy in the many short identical bit patterns. This procedure is currently very computationally expensive, and is not necessarily optimal even if the HIP model exactly matches the image distribution, but it is straightforward to code and serves to demonstrate the capability.

Figure 6 shows the root-mean-squared and maximum errors versus the size of the reulting compressed file, respectively. This is for one randomly-chosen mass ROI image, which was not part of the training set of the HIP model. The HIP algorithm gives mean errors that are comparable to JPEG, and suggests that its maximum errors are a little lower. It is perhaps not surprising, since the HIP model was fit to similar data while JPEG is intended to be general, but it demonstrates the potential. Compressed and uncompressed images are shown in Figure 7.

6 Conclusion

We have developed a class of image probability models we call hierarchical image probability or HIP models. To justify these, we showed 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 argued 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 act as indices of mixture components. The resulting model is very similar to the Hidden Markov Tree models, but allows modelling somewhat more general image structures. Because they are models of 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. Here we have presented results for mammographic image analysis. However there are obviously other modalities and medical application areas were 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 [19]. 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.

Acknowledgements

We thank Drs. Robert Nishikawa and Maryellen Giger of The University of Chicago for useful discussions and providing the data. This work was funded by the U.S. Army Medical Research and Material Command (DAMD17-98-1-8061). This paper does not necessarily reflect the position or the policy of the US government, and no official endorsement should be inferred.

References

[1] C.E. Floyd, J.Y. Lo, A.J. Yun, D.C. Sullivan, and P.J. Kornguth, "Prediction of breast cancer malignancy using an artificial neural network," *Cancer*, vol. 74, pp. 2944–2948, 1994.

[2] Y. Jiang, R.M. Nishikawa, D.E. Wolverton, C.E. Metz, M. L. Giger, R.A. Schmidt, and K. Doi, "Automated feature analysis and classification of malignant and benign microcalcifications," *Radiology*, vol. 198, pp. 671–678, 1996.

[3] W. Zhang, K. Doi, M. L. Giger, Y. Wu, R. M. Nishikawa, and R. Schmidt, "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," *Medical Physics*, vol. 21, no. 4, pp. 517–524, 1994.

[4] S.C. Lo, H.P. Chan, J.S. Lin, H. Li, M.T. Freedman, and S.K. Mun, "Artificial convolution neural network for medical image pattern recognition," *Neural Networks*, vol. 8, no. (7/8), pp. 1201–1214, 1995.

[5] C. D. Spence and P. Sajda, "Applications of multi-resolution neural networks to mammography," in *Advances in Neural Information Processing Systems 11*,

Michael S. Kearns, Sara A. Solla, and David A. Cohn, Eds., Massachusetts Institute of Technology, Cambridge, MA 02142, 1999, pp. 938–944, MIT Press.

[6] S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural Computation*, vol. 9, no. 8, pp. 1627–1660, 1997.

[7] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. PAMI*, vol. PAMI-6, no. 6, pp. 194–207, Nov. 1984.

[8] R. Chellappa and S. Chatterjee, "Classification of textures using Gaussian Markov random fields," *IEEE Trans. ASSP*, vol. 33, pp. 959–963, 1985.

[9] J. S. De Bonet and P. Viola, "Texture recognition using a non-parametric multi-scale statistical model," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 1998.

[10] J. S. De Bonet, P. Viola, and J. W. Fisher III, "Flexible histograms: A multiresolution target discrimination model," in *Proceedings of SPIE*, E. G. Zelnio, Ed., 1998, vol. 3370.

[11] M. R. Luettgen and A. S. Willsky, "Likelihood calculation for a class of multiscale stochastic models, with application to texture discrimination," *IEEE Trans. Image Proc.*, vol. 4, no. 2, pp. 194–207, 1995.

[12] D. Kopans, *Breast Imaging*, Lippincott, Philadelphia, PA, 1989.

[13] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 886–902, 1998.

[14] H. Cheng and C. A. Bouman, "Multiscale bayesian segmentation using a trainable context model," *IEEE Transactions on Image Processing*, vol. 10, no. 4, pp. 511–525, Apr. 2001.

[15] M. I. Jordan, Ed., *Learning in Graphical Models*, vol. 89 of *NATO Science Series D: Behavioral and Brain Sciences*, Kluwer Academic, 1998.

[16] R.M. Nishikawa, R.A. Schmidt, R.B. Osnis, M.L. Giger, K. Doi, and D.E. Wolverton, "Two-year evaluation of a prototype clinical mammographic workstation for computer-aided diagnosis," *Radiology*, vol. 201, no. (P), pp. 256, 1996.

[17] J.A. Rissanen, "Information theory and neural nets," in *Mathematical Perspectives on Neural Networks*, Smolensky, Mozer, and Rumelhart, Eds., 1996, pp. 567–602.

[18] C. Metz, "Current problems in ROC analysis," in *Proceedings of the Chest Imaging Conference*, Madison, WI, Nov. 1988, pp. 315–33.

[19] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, New York, 1995.

[20] W. M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, no. 1, 1996.