AD_____

Award Number: DAMD17-98-1-8061

TITLE: Application of Information Theory to Improve Computer-Aided Diagnosis Systems

PRINCIPAL INVESTIGATOR: Paul Sajda, Ph.D.

CONTRACTING ORGANIZATION: Sarnoff Corporation
Princeton, New Jersey 08543-5300

REPORT DATE: August 2000

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 2000	Annual (1 Jul 99 - 1 Jul 00)

4. TITLE AND SUBTITLE
Application of Information Theory to Improve Computer-Aided Diagnosis Systems

5. FUNDING NUMBERS
DAMD17-98-1-8061

6. AUTHOR(S)
Paul Sajda, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Sarnoff Corporation
Princeton, New Jersey 08543-5300

E-MAIL:
psajda@sarnoff.com

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

Computer-aided diagnosis (CAD) systems for mammography are an approach for low-cost double-reading. Current systems often suffer from unacceptably high false positive rates. Improved methods are needed for optimally setting the system parameters, particularly in the case of statistical models and neural networks which are a common element of most CAD systems. This research project looks to apply principles from information theory to build statistical models for CAD systems. Under the second year of this project we have further evaluated our generative hierarchical image probability (HIP) model, trained using information theoretic model selection. Our results show that HIP can reduce false positive rates by 30% for a data set constructed using The University of Chicago CAD mass detection system. We have also demonstrated the generative utility of our HIP model. We have synthesized regions of interest (ROIs) using the HIP model, enabling one to gain an intuition into the structure the HIP model learns for representing the two classes. Finally we have used the generative structure of the HIP model to detect novel examples-examples that significantly differ from the training data. We have shown how novelty detection can be used to generate confidence measures for improved ROC performance.

14. SUBJECT TERMS
Breast Cancer, Computer-aided Diagnosis, Model Selection, Information Theory, Hierarchical Image Probability

15. NUMBER OF PAGES
21

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

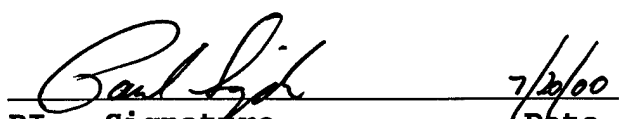
N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

N/A For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.



PI - Signature 7/26/00 Date

Table of Contents

Applications of Information Theory to Improve Computer-Aided Diagnosis Systems

Year 2 Progress Report

Paul Sajda, Clay Spence and Lucas Parra
Sarnoff Corporation
CN5300
Princeton NJ 08543-5300
{psajda, cspence, lparra}@sarnoff.com

Introduction

During the second year of this project we have been further developing and evaluating our hierarchical image probability (HIP) for mammographic computer-aided diagnosis (CAD) applications. In particular our effort has more rigorously evaluated the generative properties of the model for image synthesis and novelty detection. Analysis of the HIP model for synthesizing new mammographic images is important for understanding how the model captures image structure specific to mammographic masses. Novelty detection is particularly relevant since it would enable our system to establish confidence measures on detection, something which most current CAD systems do not offer. Experiments have been done using a mammographic mass dataset from The University of Chicago (UofC) and in all cases performance has been evaluated relative to the UofC CAD system (Nishikawa et al., 1995)—e.g. the HIP model augments the UofC CAD system. Finally, we have investigated two information-based approaches for model selection, the Minimum Description Length principle (MDL) and the Akaike's Information Criterion (AIC) both of which track HIP's generalization performance.

Body

The following are the two primary tasks completed under the second year of this project;

1. Further develop and evaluate the hierarchical image probability model, specifically focusing on the generative aspects of its architecture.
2. Apply and evaluate MDL framework for selecting architecture of hierarchical model. Compare MDL framework with other model selection methods.

In the following sections we describe our progress in accomplishing these tasks. We refer to our year 1 report (Sajda et al, 1999) for a detailed description of the HIP model.

Evaluation of HIP

Most neural networks are discriminant models in that they model P(Class|Image); the probability of a class given an image. HIP is a generative model, instead modeling P(Image|Class). Using Bayes rule one can compute P(Class|Image) for classification. However, one can also use the representation of the image probability to synthesize new images, detect novel examples, remove noise, compress images etc. Thus the generative nature of HIP makes it a more flexible and useful framework compared to conventional neural network approaches. We have evaluated HIP within the context of it generative utility, specifically with regard to 1) mammographic mass classification, 2) mammographic image synthesis and 3) novelty detection/confidence measures.

Mass Classification

Our original HIP architecture used a tree of hidden labels (Sajda et al, 1999; Spence et al, 2000). These labels serve two functions: First, each label determines the distribution of the feature vectors at its pyramid level and spatial position. Second, it determines the distribution of labels below it in the label tree. These two purposes can conflict. For example, at the very top of the tree there are relatively few examples to which the mixture components can be fit, since there are very few pixels per image at that pyramid level. Therefore not many mixture components can be used. However, those labels condition very large parts of the image, which can call for many label values.

To solve this conflict, we separate the two functions by using two sets of labels, which we call *hierarchy* and *mixture* labels. The hierarchy labels have tree-like dependencies as before, but do not directly condition the feature vectors-- they carry only the coarse-to-fine conditioning. The hierarchy labels also condition the mixture labels, which then condition the feature vectors. Each mixture label is only conditioned on the local hierarchy label. This makes it possible to have few mixture components and many hierarchy labels at low-resolution pyramid levels.

Using this new hidden label architecture, the best HIP model pair, as defined by the AIC cost (see below), gives $A_z = 0.78$, and has the ROC curve shown in Figure 1. In this case we have eliminated 30% of the false positives of the UofC CAD system for mass detection, without loss in sensitivity.

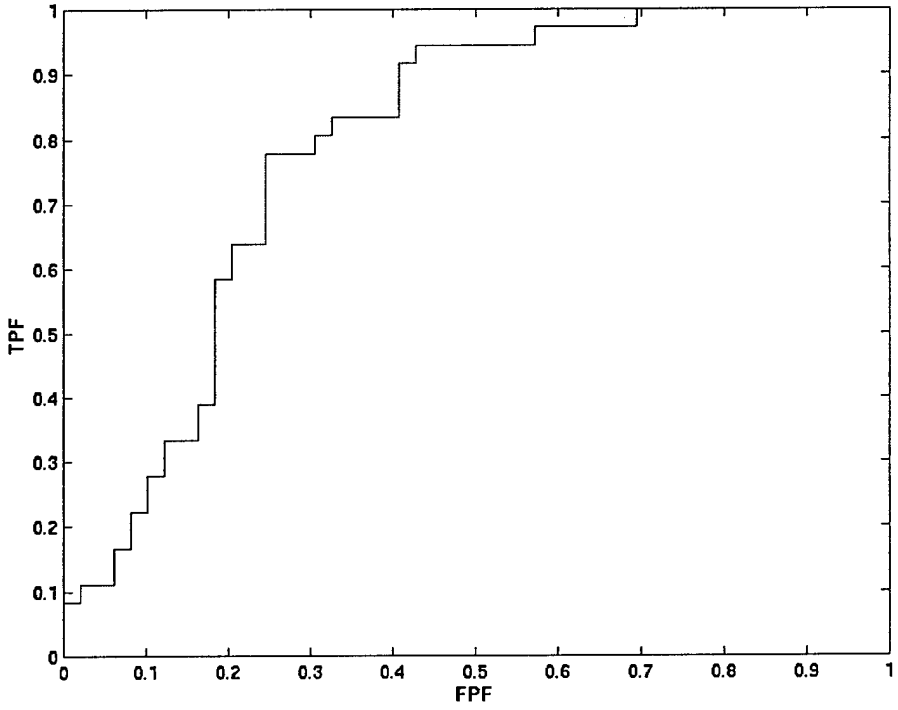
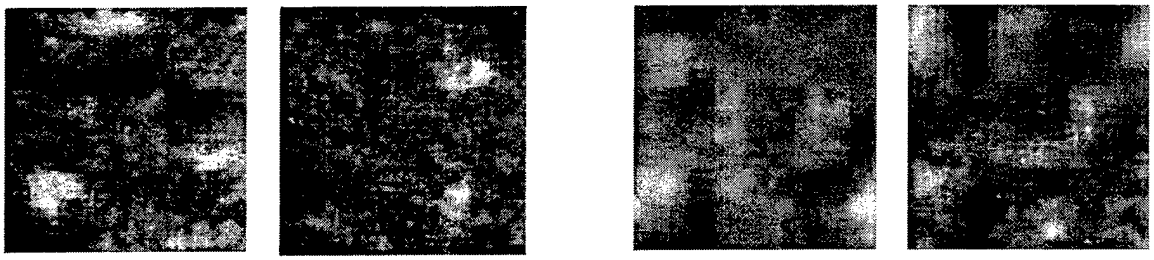


Figure 1. ROC curve of best HIP model, chosen using AIC. Results are relative to UofC CAD system for mass detection.

Mammographic Synthesis

Since the HIP model is a generative model, we can sample the model and synthesize new images. In practice, this property might be best utilized for image compression or noise reduction. Within the context of ROI classification, synthesized images can give us insight into what features the model is extracting and

representing for both positive and negative ROIs. Using the same ROI database used for classification, we constructed HIP models for positives (cancer) and negatives (no cancer). The trained HIP models were sampled to synthesize new ROI images. Figure 2 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.



ROI images synthesized from positive model

ROI images synthesized from negative model

Figure 2: Mammographic ROI images synthesized from positive and negative HIP models. Synthesized positive ROIs tend to have more focal structure, with more defined borders and higher spatial frequency content. Negative ROIs tend to be more amorphous with lower spatial frequency content.

Novelty Detection

Novelty detection identifies examples that are significantly different from the examples on which the model(s) was trained (Bishop, 1994). 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 neural network/statistical model. The HIP model's generative structure enables novel examples to be identified by thresholding the log-likelihood of the models. Figure 3 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 and synthesis). 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 begins to increase. Also shown in Figure 3 are data that are rejected (not classified) because they fall below the threshold at the given rejection rate—these ROIs are novel with respect to the data on which the models were trained. Our current effort is investigating, more thoroughly, the role novelty detection might play in generating new training data for updating a CAD system.

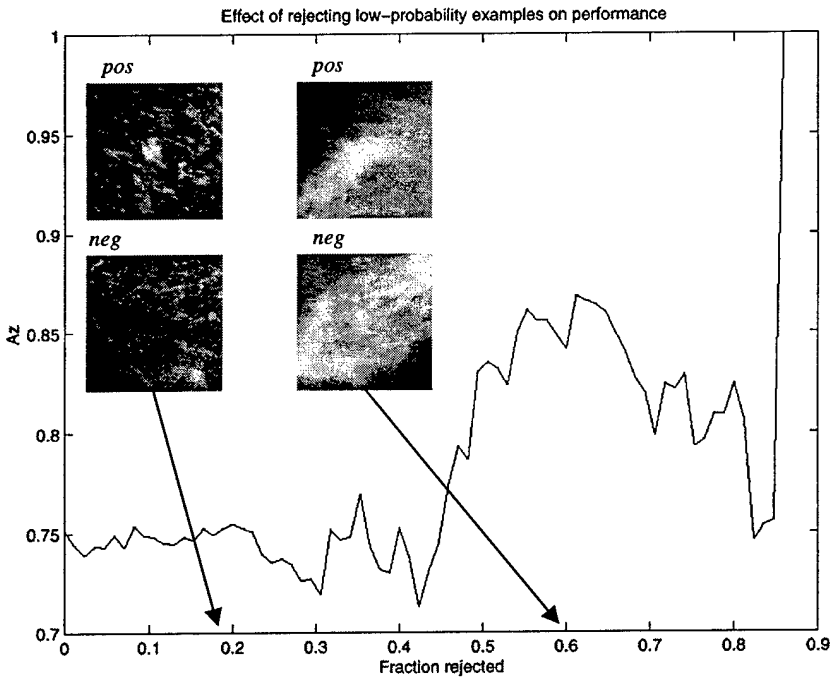


Figure 3: Novelty detection for improving ROC performance. The log-likelihood of the two HIP models (positive and negative) can be thresholded so that we reject (do not classify) a fraction of the test data that is novel, relative to the training examples. Shown is the area under the ROC curve as this novelty/confidence threshold is increased (thus increasing the fraction rejected). Also shown are examples of negative and positive ROIs that would be rejected at different thresholds.

Information Theoretic Model Selection

Information theory provides us with at least two criteria for selecting between alternative models for a probability: the Minimum Description Length (*MDL*) criterion and the Akaike's Information Criterion (*AIC*). We have investigated the usefulness of these criteria for choosing Hierarchical Image Probability (*HIP*) models for classifying mammographic mass Regions Of Interest (*ROIs*). A typical result is shown in is Figure 4. Both MDL and AIC track test A_z performance—MDL and AIC cost decrease as A_z performance on the test set increases. In the following we describe the two criteria and then suggest a methods to further improve the information theoretic selection criteria.

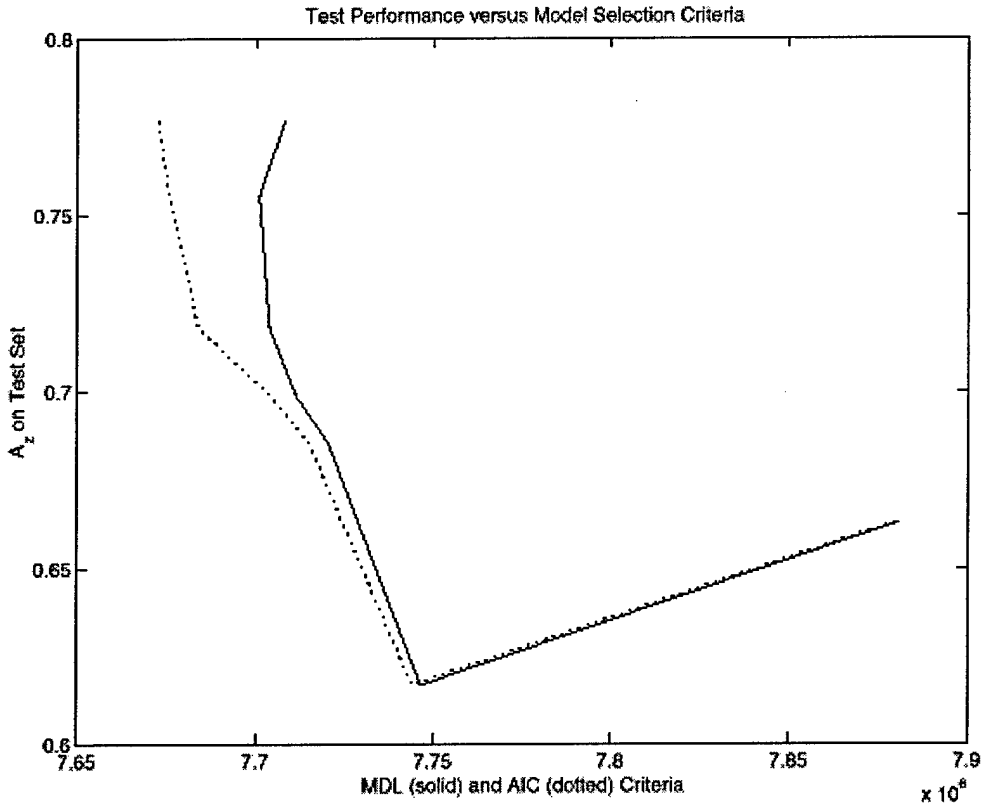


Figure 4. Information theoretic model selection using AIC (red) and MDL (blue). Plotted is model cost vs A_z on the test data. MDL would choose a model with test $A_z = 0.75$ while AIC would choose a model with A_z=0.78.

MDL

The minimum description length of a set of data is the length of the data encoded according to some probability model, which is the model we are trying to fit to the data, plus the length of the description of the model (Rissanen, 1983; Rissanen, 1996). The length of the encoding of the data is the negative log probability density of the data according to the model, plus a constant representing the precision with which the data must be specified. We ignore this constant when doing model selection, since it is the same for all models.

The code length of the model has two components, a term for coding the architecture, and a term for encoding the parameters. Suppose we are comparing models with different structures. For example, we may be comparing mixture density models with different numbers of mixture components. We will call the different models *architectures*. In this example, the number of mixture components needs to be encoded, and in general the specific architecture must be encoded. In practice this is often ignored, since it is a small contribution to the total description length.

Given an architecture, we need to encode the parameter. The Cramer-Rao bound gives a lower limit on the variance of the parameters about their true value, assuming that the true probability is equal to our architecture with some values for the parameters. This limit is the inverse $\mathbf{M}^{-1}$ of the Fisher information matrix **M**,which is the negative expected value of the second derivative (or Hessian) with respect to the parameters of the log probability of the data according to the model, evaluated at the true value of the parameters. The precision with which we encode the parameters need not be greater than the precision with which we know them, i.e., it need not be greater than the standard deviations given by

$\mathbf{M}^{-1}$. Thus we would compute the components of the parameter vector along the eigenvectors of $\mathbf{M}^{-1}$, and the precision of these components are given by the square roots of the eigenvalues. The total code length

of the parameters is the sum of the logarithms of these precisions, which is the log of the square root of the determinant of $\mathbf{M}^{-1}$. The code length for the parameters θ is the negative log of the square root of this volume, or

$$-\log(|\mathbf{M}^{-1}|^{1/2}) = \frac{1}{2}\log(|\mathbf{M}|). \tag{1}$$

Since it involves the probability of all of the data, $\mathbf{M}$ is proportional to the number of examples N, at least when there are enough examples. Because of this we can pull out the dependence on N. If there are d parameters, this gives

$$\begin{aligned}\frac{1}{2}\log(|\mathbf{M}|) &= \frac{1}{2}\log\left(\left|\frac{N\mathbf{M}}{N}\right|\right) = \frac{1}{2}\log\left(N^d\left|\frac{\mathbf{M}}{N}\right|\right) \\ &= \frac{d}{2}\log(N) + \frac{1}{2}\log\left(\left|\frac{\mathbf{M}}{N}\right|\right)\end{aligned} \tag{2}$$

The second term is constant in the limit of large N, so in that limit we can ignore it. The remaining term is straighforward to compute, since we only need to know the number of parameters and the number of training examples. The total code length for MDL is therefore,

$$\text{MDL} = -\sum_{i=1}^{N}\log P(x_i \mid \theta) + \frac{d}{2}\log N \tag{3}$$

AIC

Akaike's Information Criterion is the expected Kullback-Leibler distance between the true model and the best model of the current architecture, given the data set. It is assumed that the architectures form nested sets with the true distribution being a member of one of these sets, that the number of examples N is sufficiently large, and that the current model is not too far from the true distribution. The resulting criterion is

$$\text{AIC} = -2\sum_{i=1}^{N}\log P(x_i \mid \theta) + 2d \tag{4}$$

Deficiencies of the information criteria

Both MDL and AIC assume that there are sufficent examples, N. Treatments of MDL, for example, sometimes use the term "asymptotically", which implies the number of examples goes to infinity *for a fixed model* (Rissanen, 1996). Thus we should only expect to get good results from these criteria if we have enough examples and we find a best model before trying models that are too large for the amount of data. In our current experiments we are not obviously in this situation. We have a fixed number of examples, and we are varying the model complexity. There is no criterion for deciding whether we have enough examples, or, alternately, when we have too complex a model for the criteria to be valid.

One possible method to address this "asymptotic" issue is to add corrections to the criteria For MDL, the correction is clear: include the second term from Equation (2). This is certainly more complex, but it is feasible. For the HIP model it should be possible to estimate the Hessian numerically. We are currently investigating this approach.

Current difficulties with the HIP model

While investigating image synthesis with HIP models we noticed that our filter sets tended to suppress high frequencies. This means that the inverse transformation (reconstructing an image from the filtered and sub-sampled images) must boost these frequencies. In extreme cases this will result in ringing. In less severe

cases there is a tendency toward "blockiness". This appears if we generate a set of white noise images, and then construct the original image that would have given these white noise images as feature images. That is, we assume the white noise images are feature images and reconstruct the corresponding original image. With our previous features this tended to give sharp horizontal and vertical edges that nearly group into squares (Figure 5).

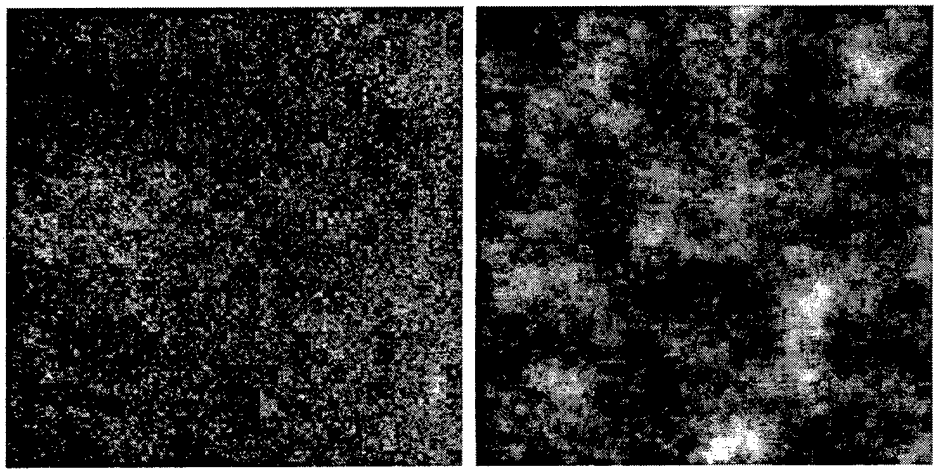


Figure 5. Two random Gaussian images. Both were produced by generating random images that were treated as transform or feature images, i.e., as if they were generated by filtering and subsampling some original image with a complete set of filters. The transform was then inverted to obtain the original image that would produce the random transform images. The left image was generated by setting the transform images to unit variance Gaussian white noise. The right image was generated by a HIP model with only one component per level. The HIP model was fit to mammographic mass ROIs. The decreased blockiness is due to the correlation between features at successive pyramid levels, which is captured by the HIP model.

As shown in Figure 5, the HIP model can partially eliminate this blockiness because it captures correlations between features at neighboring levels. Ignoring these correlations gives increased blockiness. While it is good that the HIP model can learn to eliminate artifacts such as blockiness, it is not a good use of the HIP model's resources since these artifacts are introduced because of the choice of features. We would prefer to have features that do not have such artifacts, so the resources of the HIP model can be devoted to learning other structures.

To address these issues we have developed a new set of features. Our design is based on the need for white noise in the features to imply white noise in the original image. This condition implies that the transform be orthogonal, so our new features are constructed using orthogonal filters suitable for sub-sampling by three. When we synthesize an image by setting transform images to white noise, the resulting image is significantly less blocky. The remaining blockiness can be understood by considering distortions introduced in the power spectrum. With the orthogonal transform, images at several pyramid levels will combine to give an image with a stepped power spectrum, i.e., it is piece-wise flat with larger powers at lower frequencies. This approximates a 1/f power spectrum, but the square shape and sharp steps in the spectrum tend to generate blobs near certain scales in a square arrangement. We are continuing to analyze these filters to see if they can be modified to further reduce blockiness artifacts.

Key Research Accomplishments

1. We have demonstrated that the HIP model, trained using information theoretic model selection, can eliminate 30% of the false positives mass ROIs, without loss in sensitivity, using a database generated from The University of Chicago CAD mass detection system.
2. We have demonstrated the generative utility of the HIP architecture for identifying novel ROIs. This novelty detection is useful for defining confidence measures for the classifier.
3. We have demonstrated the generative utility of the HIP architecture for synthesizing new positive and negative mammographic ROIs. We have discussed how synthesis can be used to gain an intuitive understanding of the structure that is captured by the model.
4. We have shown that different information theoretic measures track the HIP generalization performance and thus offer good criteria for model selection..

Reportable Outcomes

1. C. Spence, L. Parra, and P. Sajda, "Mammographic mass detection with a hierarchical image probability (HIP) model," in *Medical Imaging 2000: Image Processing*, Kenneth M. Hanson, Editor, Proceedings of SPIE Vol. 3979, 990-997 (2000)
2. Invited talk at Columbia University Medical School "Hierarchical Neural Networks for Object Recognition: Applications to Mammographic Computer-aided Diagnosis", June 2000
3. DoD Era of Hope meeting (poster), "A Hierarchical Image Probability Model for Mammographic Mass Detection", June 2000
4. Invited lecture at The University Of Pennsylvania, Department of Bioengineering "Computer Assisted Diagnosis for Mammography", November 1999
5. NIMA/DARPA Medical Dual-use project ($1.8M). Focus on developing dual-use technology for medical and military applications. Medical areas include breast cancer, lung cancer, retinal disease and neurological disease.

Conclusions

Under the second year of this effort we have applied information theoretic criteria for selecting HIP models for mass classification in a mammographic CAD system. Our results show that HIP models selected using this criteria can reduce false positive rates by 30% for a data set constructed using The University of Chicago CAD mass detection system. We have also demonstrated the generative utility of our HIP model. We have sampled positive and negative HIP models for synthesizing ROIs, enabling us to gain an intuition into the structure the HIP model learns for representing the two classes. Finally we have used the generative structure of the HIP model to detect novel examples—examples that significantly differ from the training data. Novelty detection can be used to generate confidence measures and we have shown how these confidence measures can be used to improve ROC performance.

"so what" section

Statistical pattern recognition is a key element in any mammographic computer-aided diagnosis system. Hierarchical pattern recognizers are particularly useful since they are capable of exploiting contextual and multi-resolution information for detecting clinically significant objects. Most statistical pattern recognizers that have been previously developed for mammographic CAD have been trained to estimate P(Class|Image). By contrast, a HIP model, trained to estimate P(Image|Class), has many attractive features. One could use HIP for detection/classification in the usual way by training a distribution for each object class and using Bayes' rule to get P(Class|Image). We have reported results for the application of HIP for reducing false positives generated by The University of Chicago CAD system for mass detection. There are other attractive features of the HIP framework, which could have a major impact on the design, and development of mammographic CAD systems. Since HIP computes P(Image|Class), we can detect unusual images and reject them rather than trust the classifier; something that is not possible with models of P(Class|Image). We have shown results illustrating how novelty detection can be used to improve the ROC

performance of CAD systems. Building confidence measures into CAD systems is an open area of research and the HIP model provides a mechanism by which to generate these measures.

The HIP model has applications other than detection/classification. Since the HIP model is a generative model, one can use it to compress data, given the probability distribution of the objects of interest. If one wants lossless compression of a digital mammogram one need only train a HIP model for a set of mammographic images and then use the probability model to compress the data. More interesting is the application of HIP for lossy compression. In that case, one might train a HIP model on clinically significant objects, such as mammographic masses, since those are the parts of the image one would like to preserve– i.e. have minimal distortion and compression artifacts. The entire image can then be compressed using this model. Though there will be loss over regions of the mammogram which do not fit the model, those regions of clinical significance will be preserved since they will have a good fit to the probability model and require very few bits for compression.

References

C. M. Bishop, Novelty detection and neural network validation. *IEE Proceedings: Vision, Image and Signal Processing*, 141 (4), 217-222, 1994.

R. M. Nishikawa, R. C. Haldemann, J. Papaioannou, M. L. Giger, P. Lu, R.A. Schmidt, D. E. Wolverton, U. Bick, and K. Doi. Initial experience with a proto-type clinical intelligent mammography workstation for computer-aided diagnosis. In *Medical Imaging 1995*, Murray H. Loew and Kenneth M. Hanson, editors, Proceedings of SPIE Vol. 2434, 65-71, 1995.

J. A. Rissanen. A universal prior for integers and estimation of minimum description length. *Annals of Statistics*, 11(2):416–431, 1983.

J. A. Rissanen. Information theory and neural nets. In Smolensky, Mozer, and Rumelhart, editors, *Mathematical Perspectives on Neural Networks*, pages 567–602, 1996.

P. Sajda, C. Spence and L. Parra, Applications of information theory to improve computer-aided diagnosis Systems, Year 1 Report, *DoD Breast Cancer Program*, DAMD17-98-1-8061, July 1999

C. Spence, L. Parra, and P. Sajda, Mammographic mass detection with a hierarchical image probability (HIP) model, in *Medical Imaging 2000: Image Processing*, Kenneth M. Hanson, Editor, Proceedings of SPIE Vol. 3979, 990-997, 2000.

Appendices

Attached paper: C. Spence, L. Parra, and P. Sajda, "Mammographic mass detection with a hierarchical image probability (HIP) model," in *Medical Imaging 2000: Image Processing*, Kenneth M. Hanson, Editor, Proceedings of SPIE Vol. 3979, 990-997 (2000)

Mammographic mass detection with a hierarchical image probability (HIP) model

Clay Spence, Lucas Parra, and Paul Sajda

Sarnoff Corporation CN5300 Princeton, NJ 08543-5300

ABSTRACT

We formulate a model for probability distributions on image spaces. We show that any distribution of images can be factored exactly into conditional distributions of feature vectors at one resolution (pyramid level) conditioned on the image information at lower resolutions. We would like to factor this over positions in the pyramid levels to make it tractable, but such factoring may miss long-range dependencies. To fix this, we introduce hidden class labels at each pixel in the pyramid. The result is a hierarchical mixture of conditional probabilities, similar to a hidden Markov model on a tree. The model parameters can be found with maximum likelihood estimation using the EM algorithm. We have obtained encouraging preliminary results on the problems of detecting masses in mammograms.

Keywords: Mammography, CAD, Image Probability

1. INTRODUCTION

Many approaches to object recognition in images estimate Pr(class | image). By contrast, a model of the probability distribution of images, Pr(image), has many attractive features. We could use this for object recognition in the usual way by training a distribution for each object class and using Bayes' rule to get Pr(class | image) = Pr(image | class) Pr(class)/ Pr(image). Clearly there are many other benefits of having a model of the distribution of images, since any kind of data analysis task can be approached using knowledge of the distribution of the data. For classification we could attempt to detect unusual examples and reject them, rather than trusting the classifier's output. We could also compress, interpolate, suppress noise, extend resolution, fuse multiple images, etc.

Many image analysis algorithms use probability concepts, but few treat the distribution of images. One of the few examples of image distribution models was constructed by Zhu, Wu and Mumford.[1] They compute the maximum entropy distribution given a set of statistics for some features, which seems to work well for textures but it is not clear how well it will model the appearance of more structured objects.

There are several algorithms for modeling the distributions of features extracted from the image, instead of the image itself. The Markov Random Field (*MRF*) models are an example of this line of development; see, e.g., References 2,3. However, they tend to be very computationally expensive.

In De Bonet and Viola's flexible histogram approach,[4,5] features are extracted at multiple image scales, and the resulting feature vectors are treated as a set of independent samples drawn from a distribution. The distribution of feature vectors is then modeled using Parzen windows. This has given good results, but the feature vectors from neighboring pixels are treated as independent when in fact they share exactly the same components from lower-resolutions. To fix this one might 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,[6] 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. The Gaussian distributions are a limitation of MSP models. The result is also a model of the probability of the observations on the tree, not of the image.

All of these methods seem well-suited for modeling texture, but it is unclear how one might build models to capture the appearance of more structured objects. We will argue below that the presence of objects in images can make local conditioning like that of the flexible histogram and MSP approaches inappropriate. In the following we

E-mail: {cspence, lparra, psajda}@sarnoff.com

In *Medical Imaging 2000: Image Processing*, Kenneth M. Hanson, Editor, Proceedings of SPIE Vol. 3979 (2000) • 1605-7422/00/$15.00

Mammographic mass detection with a hierarchical image probability (HIP) model

Clay Spence, Lucas Parra, and Paul Sajda

Sarnoff Corporation CN5300 Princeton, NJ 08543-5300

ABSTRACT

We formulate a model for probability distributions on image spaces. We show that any distribution of images can be factored exactly into conditional distributions of feature vectors at one resolution (pyramid level) conditioned on the image information at lower resolutions. We would like to factor this over positions in the pyramid levels to make it tractable, but such factoring may miss long-range dependencies. To fix this, we introduce hidden class labels at each pixel in the pyramid. The result is a hierarchical mixture of conditional probabilities, similar to a hidden Markov model on a tree. The model parameters can be found with maximum likelihood estimation using the EM algorithm. We have obtained encouraging preliminary results on the problems of detecting masses in mammograms.

Keywords: Mammography, CAD, Image Probability

1. INTRODUCTION

Many approaches to object recognition in images estimate Pr(class | image). By contrast, a model of the probability distribution of images, Pr(image), has many attractive features. We could use this for object recognition in the usual way by training a distribution for each object class and using Bayes' rule to get Pr(class | image) = Pr(image | class) Pr(class)/ Pr(image). Clearly there are many other benefits of having a model of the distribution of images, since any kind of data analysis task can be approached using knowledge of the distribution of the data. For classification we could attempt to detect unusual examples and reject them, rather than trusting the classifier's output. We could also compress, interpolate, suppress noise, extend resolution, fuse multiple images, etc.

Many image analysis algorithms use probability concepts, but few treat the distribution of images. One of the few examples of image distribution models was constructed by Zhu, Wu and Mumford.[1] They compute the maximum entropy distribution given a set of statistics for some features, which seems to work well for textures but it is not clear how well it will model the appearance of more structured objects.

There are several algorithms for modeling the distributions of features extracted from the image, instead of the image itself. The Markov Random Field (*MRF*) models are an example of this line of development; see, e.g., References 2,3. However, they tend to be very computationally expensive.

In De Bonet and Viola's flexible histogram approach,[4,5] features are extracted at multiple image scales, and the resulting feature vectors are treated as a set of independent samples drawn from a distribution. The distribution of feature vectors is then modeled using Parzen windows. This has given good results, but the feature vectors from neighboring pixels are treated as independent when in fact they share exactly the same components from lower-resolutions. To fix this one might 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,[6] 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. The Gaussian distributions are a limitation of MSP models. The result is also a model of the probability of the observations on the tree, not of the image.

All of these methods seem well-suited for modeling texture, but it is unclear how one might build models to capture the appearance of more structured objects. We will argue below that the presence of objects in images can make local conditioning like that of the flexible histogram and MSP approaches inappropriate. In the following we

E-mail: {cspence, lparra, psajda}@sarnoff.com

In *Medical Imaging 2000: Image Processing*, Kenneth M. Hanson, Editor, Proceedings of SPIE Vol. 3979 (2000) • 1605-7422/00/$15.00

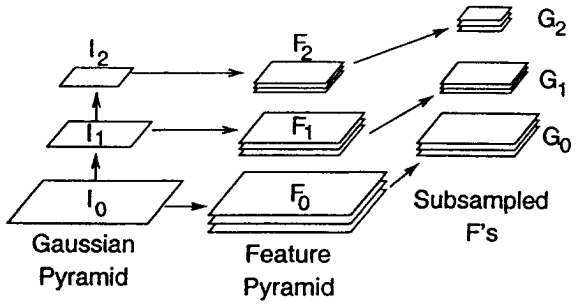


Figure 1. Pyramids and feature notation.

present a model for probability distributions of images, in which we try to move beyond texture modeling. This hierarchical image probability (*HIP*) model is similar to a hidden Markov model on a tree, and can be learned with the EM algorithm. In preliminary tests of the model on classification tasks the performance was comparable to that of other algorithms.

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 somehow at low resolution. The large-scale structures affect finer scales by the conditioning.

In fact we can 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 some 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 $\tilde{\mathcal{G}}_l$.

Suppose now 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

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

This is a very general result, valid for all $\Pr(I)$, no doubt with some rather mild 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, something like

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

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 kind of texture across a large area of the image. If an object of this class is always present, the distribution may factor, but if such objects aren't always present and can't be inferred from lower-resolution information, the presence of the texture at one location affects the probability of its presence elsewhere.

We introduce hidden variables to represent the non-local information that is not captured by local features. They should also constrain the variability of features at the next finer scale. Denoting them collectively by A, we assume that conditioning on A allows the distributions over feature vectors to factor. In general, the distribution over images becomes

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

As written this is absolutely general, so we need 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.

As a first model we have chosen the following structure for our HIP model:*

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

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 $\bar{\mathbf{G}}_l$.

Since $a_l(x)$ codes non-local information we can think of the labels A_l as a segmentation or classification 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 the dependency graph of the hidden variables a tree structure. Such a probabilistic tree of discrete variables is sometimes referred to as 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 the sake of simplicity we've chosen $\Pr(\mathbf{g}_l \,|\, \mathbf{f}_{l+1}, a_l)$ to be normal with mean $\bar{\mathbf{g}}_{l,a_l} + M_{a_l}\mathbf{f}_{l+1}$ and covariance Σ_{a_l}, that is,

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

4. EM ALGORITHM

Due to the tree structure, the belief network for the hidden variables is relatively easy to train with an EM algorithm. The expectation step (summing over a_l's) can be performed directly. If we had chosen a more densely-connected structure with each child having several parents, we would need either an approximate algorithm or Monte Carlo techniques. 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.[7]

*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 $\prod_x \Pr(\mathbf{g}_L \,|\, a_L, x) \Pr(a_L, x)$.

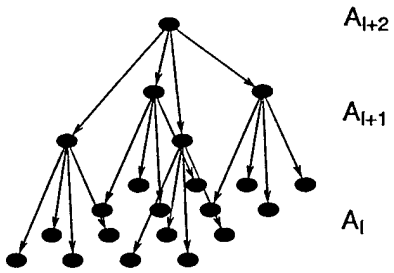


Figure 2. Tree structure of the conditional dependency between hidden variables in the HIP model. With subsampling by two, this is sometimes called a quadtree structure.

Once we can compute the expectations, the normal distribution makes the M-step tractable; we simply compute 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, we need to choose a parameterization for the model. The parameterization of $\Pr(\mathbf{g} \,|\, \mathbf{f}, a)$ is given above in Equation 4. For $\Pr(a_l \,|\, a_{l+1})$ we use the parameterization

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

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:

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

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

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

and

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

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

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

4.2. EXPECTATION

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

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

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

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

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

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

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

The upward recursion relations (12–13) are initialized at $l = 0$ with $u_0(a_0, x) = \Pr(\mathbf{g} \,|\, \mathbf{f}_1, a_0, x)$ and end at $l = L$. At layer L Equation 13 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. 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 $d_{L+1}(a_{L+1}, x) = d_{L+1}(x) = 1$.† The downwards recursion ends at $l = 0$ with equation (14).

We can now compute (11) as

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

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

Obviously computations (12–18) in the E-step at iteration t need to be completed with fixed parameters θ^t.

Because of the dependence of $\mathbf{g}_l$ on $\mathbf{f}_{l+1}$, these u's and d's are not, in general, actual probabilities. In spite of this it can be shown that these recursion relations are correct.

5. EXPERIMENTS

5.1. CLASSIFICATION OF VEHICLES IN SAR IMAGERY

Though not a medical imaging problem, we first present the results of our experiments on synthetic aperture radar (*SAR*) imagery, since SAR imagery is noisy and involves detecting an extended textured object, much like a breast mass and many other medical imaging problems. The problem was to discriminate between three target classes in the MSTAR public targets data set, to compare with the results of the flexible histogram approach of De Bonet, et al.[5] We trained three HIP models, one for each of the target vehicles BMP-2, BTR-70 and T-72 (Figure 3). As in Reference 5 we trained each model on ten images of its class, one image for each of ten aspect angles, spaced approximately 36° apart. We trained one model for all ten images of a target, whereas De Bonet et al trained one model per image.

We first tried discriminating between vehicles of one class and other objects by thresholding $\log \Pr(I \,|\, \text{class})$, i.e., no model of other objects is used. In essence this discriminates simply by judging whether an image looks sufficiently similar to the training examples. For the tests, the other objects were taken from the test data for the two other vehicle classes, plus seven other vehicle classes. There were 1,838 image from these seven other classes, 391 BMP2 test images, 196 BTR70 test images, and 386 T72 test images. The resulting ROC curves are shown in Figure 4a.

We then tried discriminating between pairs of target classes using HIP model likelihood ratios, i.e., $\log \Pr(I \,|\, \text{class1}) - \log \Pr(I \,|\, \text{class2})$. Here we could not use the extra seven vehicle classes. The resulting ROC curves are shown in Figure 4b. The performance is comparable to that of the flexible histogram approach.

†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 $\Pr(a_L | \, a_{L+1})$ turns then into a prior $\Pr(a_L)$

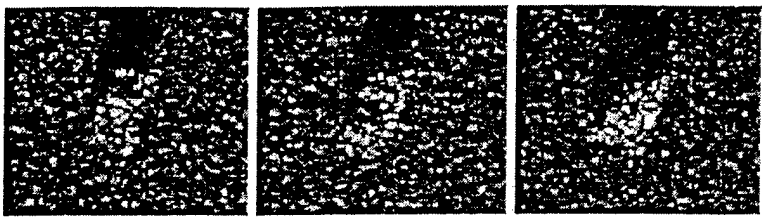


Figure 3. SAR images of three types of vehicles to be detected.

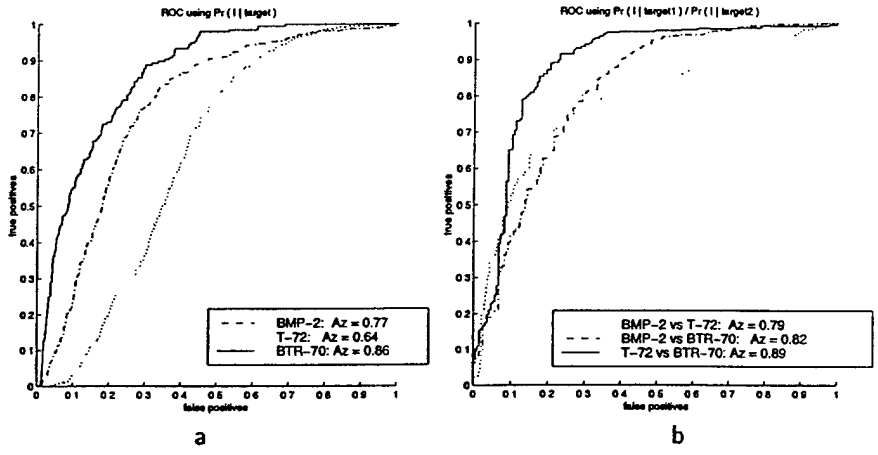


Figure 4. ROC curves for vehicle detection in SAR imagery. (a) ROC curves by thresholding HIP likelihood of desired class. (b) ROC curves for inter-class discrimination using ratios of likelihoods as given by HIP models.

5.2. MASS DETECTION

We applied HIP to the problem of detecting masses in ROIs taken from mammograms, as detected by a CAD system at the University of Chicago. We trained a HIP model of the distribution of positive images on 36 randomly-chosen ROIs that contained masses, and a second HIP model on 48 randomly-chosen ROIs without masses. The likelihood ratio was then used as the test criterion, i.e., a threshold on this ratio is used to decide which ROIs will be called masses. The true and false positive rates as a function of the threshold were measured on a test set with 36 mass and 49 non-mass ROIs.

A search was performed over the number of hidden labels values at each level. The search criterion was 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. The maximum number of labels in a level was bounded (somewhat arbitrarily) at 17, since doubling the number of components in a level at this point was observed to decrease the MDL criterion, but very little, and the computation time would approximately double.

The best architecture had 17, 17, 11, 2, and 1 hidden label in levels 0–4, respectively. For this architecture, A_z was 0.73. This detector had a specificity of 33 % at a sensitivity of 95 %. The ROC curve is shown in Figure 5. While this performance is not as good as we might hope, being worse than our own HPNN classifier,[8] for instance, it demonstrates that the model captures relevant information for classification. We hope that further work, particularly in model and feature selection, will improve on these results.

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

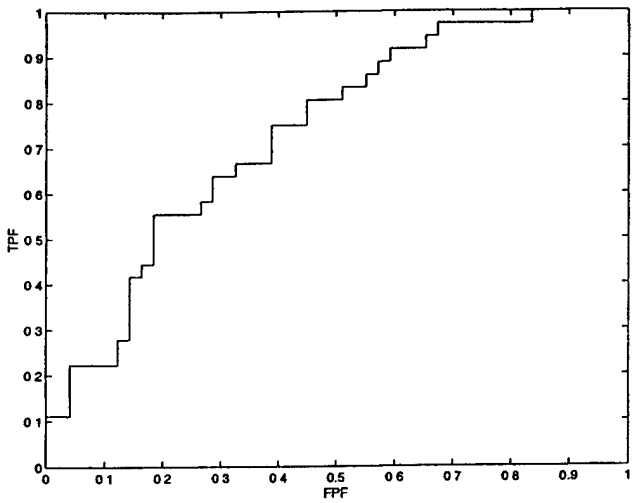


Figure 5. ROC curve for HIP detector of Mass ROIs generated by U. Chicago CAD.

somewhat like a hidden Markov model on a tree. The HIP model can be used for a wide range of image processing tasks besides classification, e.g., compression, noise-suppression, up-sampling, error correction, etc.

There is much room for further work on variations of the specific HIP model presented here. The tree-structured discrete hidden variables lend themselves well to exact marginalization, but they fail to capture certain image properties. For example, contrast level and orientation could be given continuous parameterizations. See, for example, the work of Simoncelli and Wainwright, who developed a very similar model to capture the statistics of contrast level (which they refer to as "scale"), though they did not formulate their model as an image probability.[9] Furthermore, as is well known, the tree structure of the hidden variable dependencies will tend to artificially suppress the statistical dependence between some neighboring pixels, but not others. Allowing multiple parents would alleviate this. Unfortunately, either of these modifications would make it impractical to marginalize over the hidden variables, which is the proper probabilistic procedure. There are approximate alternatives to exact marginalization, which should allow a far wider variety of hidden variable structures.

ACKNOWLEDGEMENTS

We thank Drs. Robert Nishikawa and Maryellen Giger of The University of Chicago for useful discussions and providing the data. This work was supported by the US Department of the Army under grant number 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. S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural Computation* **9**(8), pp. 1627–1660, 1997.
2. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. PAMI* **PAMI-6**, pp. 194–207, Nov. 1984.
3. R. Chellappa and S. Chatterjee, "Classification of textures using Gaussian Markov random fields," *IEEE Trans. ASSP* **33**, pp. 959–963, 1985.
4. J. S. D. Bonet and P. Viola, "Texture recognition using a non-parametric multi-scale statistical model," in *Conference on Computer Vision and Pattern Recognition*, IEEE, 1998.
5. J. S. D. Bonet, P. Viola, and J. W. F. III, "Flexible histograms: A multiresolution target discrimination model," in *Proceedings of SPIE*, E. G. Zelnio, ed., vol. 3370, 1998.
6. 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.* 4(2), pp. 194–207, 1995.

7. M. I. Jordan, ed., *Learning in Graphical Models*, vol. 89 of *NATO Science Series D: Behavioral and Brain Sciences*, Kluwer Academic, 1998.
8. C. D. Spence and P. Sajda, "Applications of multi-resolution neural networks to mammography," in *Advances in Neural Information Processing Systems 11*, M. S. Kearns, S. A. Solla, and D. A. Cohn, eds., pp. 981–988, MIT Press, (Cambridge, MA), 1998.
9. M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. Leen, and K.-R. Müller, eds., MIT Press, (Cambridge, MA), 1999.