Report Number: WUCS-98-20

1998-01-01

Diagnostic Screening of Digital Mammograms Using Wavelets and Neural Networks to Extract Structure

Barry L. Kalman, Stan C. Kwasny, and William R. Reinus

As the primary tool for detecting breast carcinoma, mammography provides visual images from which a trained radiologist can identify suspicious areas that suggest the presence of cancer. We describe an approach to image processing that reduces an image to a small number of values based on its structural characteristics using wavelets and neural networks. To illustrate its utility, we apply this methodology to the automatic screening of mammograms for mass lesions. Our results approach performance levels of trained human mammographers.

Recommended Citation

Kalman, Barry L.; Kwasny, Stan C.; and Reinus, William R., "Diagnostic Screening of Digital Mammograms Using Wavelets and Neural Networks to Extract Structure" Report Number: WUCS-98-20 (1998). *All Computer Science and Engineering Research.*
https://openscholarship.wustl.edu/cse_research/472

Diagnostic Screening of Digital Mammograms Using Wavelets and Neural Networks to Extract Structure

Barry L. Kalman, Stan C. Kwasny and
William R. Reinus

WUCS-98-20

July 1998

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
St. Louis MO 63130

Diagnostic Screening of Digital Mammograms Using Wavelets and Neural Networks to Extract Structure

BARRY L. KALMAN AND STAN C. KWASNY
barry@cs.wustl.edu, sck@cs.wustl.edu
Department of Computer Science, Washington University, St. Louis, MO 63130

WILLIAM R. REINUS
reinus@totty.wustl.edu
Mallinckrodt Institute of Radiology, Barnes-Jewish Hospital, St. Louis, MO 63110

Abstract.
As the primary tool for detecting breast carcinoma, mammography provides visual images from which a trained radiologist can identify suspicious areas that suggest the presence of cancer. We describe an approach to image processing that reduces an image to a small number of values based on its structural characteristics using wavelets and neural networks. To illustrate its utility, we apply this methodology to the automatic screening of mammograms for mass lesions. Our results approach performance levels of trained human mammographers.

Keywords: structured knowledge extraction, classification of structured information, unsupervised learning of hierarchical structure, medical diagnosis, linear-output sequential recursive auto-associative memory.

1. Introduction

Increasingly, modern medicine relies on a vast array of imaging studies for diagnosis. Mammography, in particular, supports efforts to screen for and detect breast carcinoma, a disease that will affect one in nine women over their lifetime. Mammograms are two-dimensional images that show the structure of the breast. A trained radiologist can reliably identify most suspicious areas that suggest the presence of cancer, although failure to diagnose cancer occurs an estimated 10% to 31% of the time[5],[7],[19]. At least one study has suggested that double reading of screening mammograms will reduce the cancer miss rate[3]. By virtue of this suggestion, interest has been generated in developing an automatic method of prescreening mammograms as a complement to the radiologist. Neural networks that can extract the structure represented in a mammographic image show promise as a new technique to address this problem.

1.1. Computer Aided Diagnosis and Prescreening

A large body of work has been devoted to using computer screening techniques to help the radiologist recognize areas with possible pathology on an image. The majority of the work done to date has focused on mammography and chest radiography. A number of different algorithmic computer aided diagnosis (CAD) schemes have been investigated. These include subtraction techniques, topographic techniques, filtering techniques and staged expert systems.

The majority of these schemes attempt to identify anomalies using a method that either looks for image differences based on comparison with known normal tissue (subtraction) or by image feature identification and extraction of features that correlate with pathologic anomalies, e.g., changes in density that may indicate a mass on a mammogram. Most systems proceed in stages, first examining the image data and extracting predetermined image features, then localizing regions of interest (ROIs) which can be examined further for potential anomalies. High degrees of sensitivity (85% to 100%) have been achieved using several of these techniques (see, for example, [12],[13],[14],[15],[16],[39],[40],[41],[42],[43],[45]) but many have been hampered by high false-positive rates (1 to 4 false positive identifications per image) and hence low specificity (see, for example, [12],[15],[16],[18],[30],[43],[45]). The problem of false positives is compounded by the fact that false positive rates are reported per image, not per case. Since many radiologic examinations include more than one image, the actual number of false positives per case may be a multiple of those reported.

In an attempt to address the specificity issue, a number of different approaches have been tried to reduce false positive rates. Many of these have focused on the use of artificial neural networks (ANN)[12],[30],[43]. ANNs show great value in analysis of problems that are structural in nature and, as such, they are excellent for problems of pattern recognition.

Inspired by neurophysiology, ANNs are trained in one of two ways:

1. Supervised training: both input data and its corresponding outputs are provided during training. The ANN learns the mapping for which the input-output pairs are an extensional sample.
2. Unsupervised training: input data are provided and a criteria for judging outputs is determined. The ANN learns a mapping that fits the criteria.

Auto-associative learning is based on unsupervised training in which the ANN attempts to learn an identity mapping. More importantly, the patterns of activation that occur internally must encode the data to minimize the error of the identity mapping and this produces potentially efficient and compact encodings of the data.

Neural Networks are limited in at least four related ways:

1. The quantity of data required for each case determines the size of the input layer which partially determines the size of the network in terms of adjustable weights. Aside from the time required to compute and apply weight adjustments during each training cycle, a large number of inputs may increase the complexity of the learning task and may require more training cycles.
2. The size of the data set from which training is performed determines the number of presentations required during each training cycle. It may also reflect an underlying complexity of the learning task if there is sufficient diversity among members of the data set.
3. Fixed data set size and increased number of inputs lead to an excess of connections and most likely poor generalization.
4. The training dataset must be adequate in representation and depth.

Nonetheless appropriately configured ANNs have proven to be useful for many problems in medical diagnosis[4],[9],[22],[34],[35]. In mammography, ANNs have been studied primarily as a method to reduce the rate of false-positive anomaly detection arising from other computer-aided diagnosis techniques. Such studies have shown that ANNs can significantly reduce the rate of detection of false-positive anomalies on mammograms by as much as 50% to 62%[40],[43],[44].

1.2. Full Image Prescreening

To our knowledge, no group has developed a technique that uses ANNs to directly analyze and detect anomalies from entire digitized medical images without first using extraction techniques. When other researchers have applied neural network technology to mammographic anomaly detection, they have used the ANNs to examine either small ROIs taken from the image or symbolic information extracted from the image, but not the entire image itself (see, for example, [43],[20],[40],[12],[30],[44],[39]). Most researchers require that the input data for the ANNs undergo preliminary enhancement of specific features, density or edges, for example.

2. Approach

It has been suggested that mammograms actually contain significantly more information than is transmitted to the human eye[10]. Our structural approach is based on the premise that mammographic images reflect an internal structure of higher dimensionality and that this structure can be extracted automatically using wavelet and ANN technology. Unlike other approaches, no predetermined image features are extracted.

Instead, using our methodology, feature-like values that reflect the image structure dynamically emerge from the data during training. Clustering techniques help in extracting those values which are synthesized into two small sets of values for each image. Subsequently, these data are supplied to a set of feed forward neural networks (FFNN) for analysis and anomaly identification. Voting among the FFNNs leverages their individual performance to enhance the screening process.

2.1. Processing Steps

As shown in Figure 1, processing full mammograms for screening involves several steps. First, cases with the proper pathology (masses) are selected for inclusion in the dataset. The images from these cases are scanned, digitized, and resized as necessary to match the demands of our software. Collectively, the dataset of images undergoes wavelet processing, structure discovery, and finally evaluation for mass lesion screening. Once training has been performed, images can be efficiently digitized, wavelet transformed, processed in a forward direction by a recurrent network, formed into two small feature vectors and classified by the collection of FFNNs to determine a screening outcome.

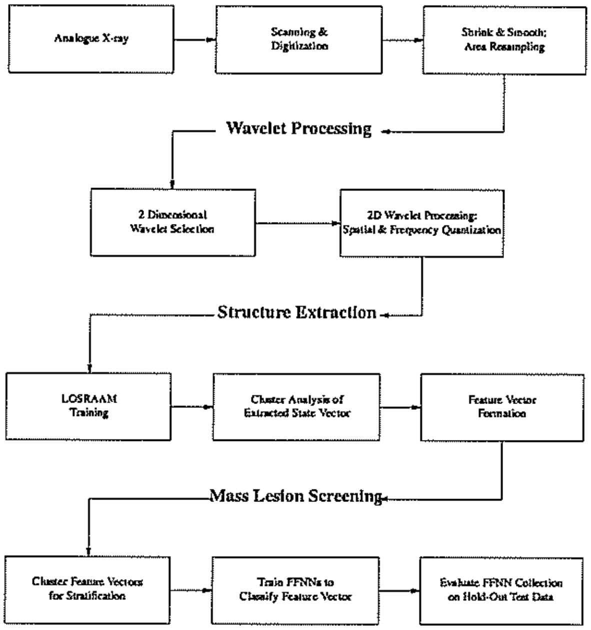
Analogue X-ray
Scanning & Digitization
Shrink & Smooth; Area Resampling
Wavelet Processing
2 Dimensional Wavelet Selection
2D Wavelet Processing: Spatial & Frequency Quantization
Structure Extraction
LOSRAAM Training
Cluster Analysis of Extracted State Vector
Feature Vector Formation
Mass Lesion Screening
Cluster Feature Vectors for Stratification
Train FFNNs to Classify Feature Vector
Evaluate FFNN Collection on Hold-Out Test Data

Fig. 1. Schematic diagram of image processing using wavelets and neural networks.

2.1.1. Wavelet Processing. Multiresolution (five-level) and multidirection (two-dimensional) wavelet analysis with quadratic spline wavelets is applied to each square image[28]. These wavelets are equivalent to the first-order derivative of a smoothing function, and so they are specifically designed to enhance the edges of image objects. Laine[27] has used a similar wavelet design to show that high-quality mammogram reconstruction can be performed with a truncated wavelet hierarchy wherein coefficients below a certain threshold are discarded. Truncation eliminates noise and insignificant features from the data while aiding in data reduction[11]. Empirically, a hard [1] threshold of 0.25 was determined to produce the best results for our data.

The remaining wavelet coefficients can be viewed structurally, descending along the two dimensions (x followed by y) at each of the five

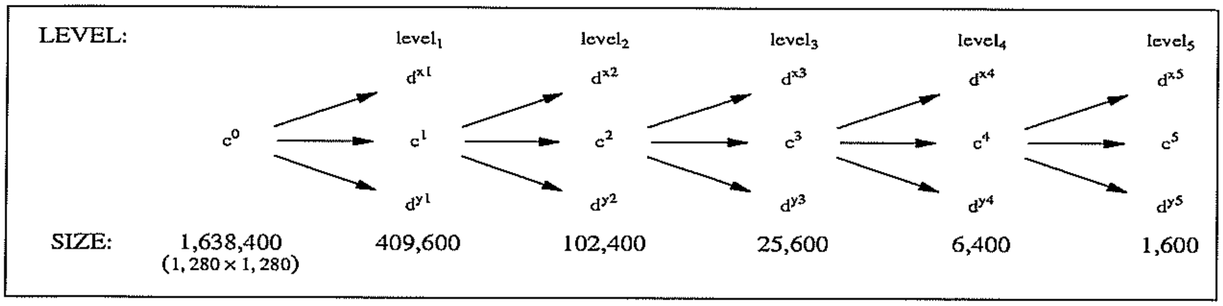

Fig. 2. Diagram of wavelet structure shows number of coefficients resulting at each level.

levels of resolution, terminating with the remaining (non-decomposed) coefficients at level five, as shown in Figure 2. Each coefficient, together with its level and position in the hierarchy, forms a triplet of values. These can be linearly sequenced by a traversal to produce a canonical linearized representation. Note that every step of this process, except the elimination of noise, is reversible.

2.1.2. Finding Structure in Linearized Wavelet Data. The linearized wavelet data is transformed into fixed-length fuzzy feature vectors[2] (FFVs) and fuzzy feature transition matrices (FFTM) using a form of recursive auto-associative memory (RAAM)[31]. The RAAM architecture has been modified into a recurrent network called an SRAAM[8][37] which accepts sequences of inputs after a method by Kwasny and Kalman[26]. The output units of the SRAAM have been further modified by eliminating the sigmoidal function on outputs to allow linear-valued outputs as shown in Figure 3. We call this variation a linear-output SRAAM (LOSRAAM). In this form, the network is a recurrent network connected by layers (i.e., no shortcut connections) which is trained auto-associatively.

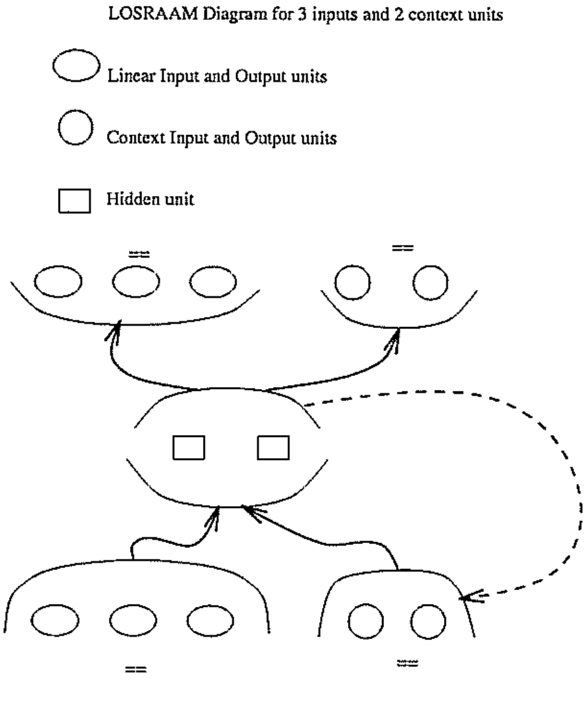
LOSRAAM Diagram for 3 inputs and 2 context units
Linear Input and Output units
Context Input and Output units
Hidden unit
==
==
==
==

Fig. 3. LOSRAAM neural network.

The LOSRAAM, upon processing the wavelet triplets from each image, develops a sequence of activation patterns on the N-unit hidden layer of the LOSRAAM. This sequence of N-dimensional vectors traces the trajectory of the image as a series of hyperspace points with the path of the trajectory determined by the structure of the image. Clustering these points identifies K centers of attraction[3] and exposes a type of iterated function system (IFS)[6],[32].

An image traverses the IFS space in a particular way leaving its unique "fingerprint" which can be captured by aggregating within clusters to produce a single K-dimensional vector aggregate as a summarization of the image. Aggregation includes a fuzzying[4] process which first weights each hyperspace point according to its distance from the K centers of attraction and then normalizes the weighted values so they sum to one. An FFV is thus a K-dimensional vector aggregate formed by taking the sum of all vectors derived from a given image.

A related view of the trajectories traced by each image gives rise to an alternate form of aggregation. Similar to bi-gram processing often applied to speech and natural language processing[2], the sequence of points in the trajectory can be paired according to the transitions they make out-of and into each cluster. This can be summarized in a $K \times K$ matrix of fuzzy values by taking the outer product of each pair of endpoint vectors and summing the matrices to form a single $K \times K$ FFTM.

2.1.3. Mass Lesion Screening. Both FFVs and FFTMs are used as inputs to FFNNs designed to classify mammographic images for screening of suspicious masses indicative of cancer. These are presented to a set of FFNNs, each of which is trained to decide whether a mass is present or not. A number of similarly trained FFNNs each render votes on a particular image view for a particular case and the votes are collected and used to decide the screening outcome.

The success of the voting process hinges on the degree to which each vote (i.e., each network) is independent of the others. To some degree, each breast view provides correlated, but independent, information. Similarly, each FFNN converges to a solution which may or may not differ from that of other FFNNs trained from similar datasets. The succes of voting depends on the degree to which these solutions differ.

2.2. Training LOSRAAMs and FFNNs

All neural network training, including both LOSRAAM and FFNN training, is based on techniques reported in Kalman and Kwasny[25]. Reviewed here are some of the more important techniques.

2.2.1. Training RAAM-like Networks. LOSRAAMs can be trained as recurrent networks except for the portion of the output layer which represents the previous RAAM pattern. Since autoassociative training involves targeting the outputs to be the same as the inputs, and since a portion of the input is the activation pattern from the hidden layer of the previous iteration, part of the target is evolving as training proceeds. Error must be reduced against these moving targets in order for training to succeed. The derivative which controls learning must be adjusted by a term which represents the derivative for these units and which takes into consideration changes in these targets. See Appendix A for more details.

2.2.2. Singular-Valued Decomposition. Singular-Valued Decomposition[21] (SVD) is an important mathematical technique for transforming data to meet orthogonality constraints. Applying SVD to FFNN inputs was first described by Kalman[23][25] as a tool for conditioning inputs and reducing their numbers.

In this application, before presentation to the FFNN, the $K+K^2$ inputs of the FFV and FFTM are transformed using singular value decomposition (SVD) to determine the relevance of each input, a process which often reduces the number of inputs and, therefore, the size of the network.

It has proven invaluable in the mammography work in two distinct ways:

1. Reducing the number of inputs from $K + K^2$ to approximately K.
2. Pre-conditioning the inputs to facilitate faster training and in many cases, based on our experience, making training possible that otherwise would not have been so.

Our use of SVD permits a FFNN trained on transformed inputs to be reverse-transformed into an FFNN that performs identically on the original inputs. Thus, the use of SVD is invisible on the surface, but invaluable as a technique for enhancing training.

2.2.3. Hints. Hints are realized as additional output units whose only role is to direct and push the training toward a helpful direction. In this work, they are provided during the training of FFNNs for screening. As demonstrated by Abu-Mostafa[1], poor or misleading hints may actually damage training performance while good hints can greatly improve it.

For this work, an important hint points out the location of a mass in the breast. This is realized by an additional three output units indicating upper vs. lower half of the breast, outer vs. inner half of the breast, and subareolar vs. central location in the breast[5]. We have determined through numerous training runs that this hint has a very positive affect on training.

Note that the hint itself is based on position within the breast structure and not within the image. This suggests that the higher dimensional structure of the breast is being extracted from the two-dimensional image by our techniques.

Once training is complete, the hint units, because they are extra output units, may be removed from the network without changing the activations on the other output units. Alternatively, they have the potential to provide additional information about location which would be helpful in identifying the lesion.

2.2.4. Other Training Techniques. Summarized below are other training techniques that contributed to the success of this work.

1. *Use of skip (shortcut) connections to completely connect each layer with every preceding layer in FFNNs*[6]. Without skip connections, the FFNN is required to learn both linear and non-linear features of the data, wherein the direct connections from input to output support the linear part of the mapping. This typically reduces the number of hidden units required and helps to avoid overspecification.
2. *Superlinear convergence of the conjugate gradient method for training.* This method has been enhanced using an update strategy due to Powell[33], a derivative-free line-search technique, and an adaptive step size control.
3. *Use of a self-scaling error function.* This error function punishes output values at the opposite end of the interval from the target value.

3. Results

These results represent a major expansion and redesign of a pilot experiment reported by Kalman[24]. In that study, only 55 cases were screened with 79% specificity and 50% sensitivity.

3.1. Dataset

A dataset[7] of 350 mammograms, consisting of two images each, were digitized with a 100-micron focal spot film digitizer. Each image measured 9.375×6.825 inches (23.43×17.06 cm). After digitization the images were on the order of $2,400 \times 1,800X12$ bits and then area-resampled[29] to $1,280 \times 1,280 \times 1$ byte.

The dataset contained 221 cases with masses and 129 cases without. Each of the patients had undergone breast biopsy because of a radiographically detected mass or microcalcifications or because of a palpable abnormality not detected on a mammogram. All mammograms were pathologically correlated with the results of biopsy.

For experimental purposes, the 350 cases were randomly divided into two disjoint sets: 87 cases were held out to be used for final testing and evaluation of the method; and 263 cases were designated for training purposes. The training cases were further randomly subdivided into 154 cases presented during training and 109 PAC[8] cases used to determine when to stop training. Note that the PAC set of cases does not participate in training directly and therefore constitutes a good, low-bias estimator of performance on the unbiased held-out set of cases. Table 1 shows the sizes and distribution of cases across these sets.

3.2. Data Reduction from Wavelets

Wavelet processing with hard thresholding achieved a 42.5:1 reduction in the size of the data. The 700 images have a total size of $700 \times 1,280 \times 1,280 \times 1$ $= 1.14688 \times 10^9$ bytes. This is reduced to an average of 4,824 triplets (three 8-byte values) for each image after thresholding or a total for all 700 images of 2.70144×10^7 bytes which accounts for the reduction.

3.3. Data Reduction from LOSRAAMs

The LOSRAAM network required 4 hidden units for sufficient training making the LOSRAAM network a 7-4-7 network with 67 adjustable weights. This part of the training took the most time. Us-

Table 1. Distribution of Cases in Dataset

Sets	No. of Cases	Mass Cases	Non-Mass Cases
Entire DataSet	350	221	129
Training Set	154	97	69
PAC Set	109	57	40
Hold Out Set	87	55	32

ing 8 dedicated processors on a Sun SparcCenter 2000 machine, the required 250 conjugate gradient iterations took approximately 14 days to complete.

Clustering produced $K = 7$ clusters. Therefore, each FFV contained 7 double-precision values and so each FFTM contained 49. SVD determined that these 56 values could be transformed into 5 values which account for 99.97% of the variance. The resulting 5 values complete the reduction process to $700 \times 5 \times 8$ bytes or 2.8×10^4 bytes, a reduction of 4.096×10^4:1.

3.4. Screening Processing

The mass lesion screening step involves training a FFNN to decide if there is a mass or not. To facilitate voting, 24 networks were trained using different random starting weights. Each network has 5 inputs, 5 outputs (2 for mass vs. no-mass and 3 for positional hints), and 1 hidden unit. We carefully tested FFNNs using more and fewer hidden units, but one hidden unit gave the best generalization as measured on the PAC set. With skip connections, there are 41 adjustable weights.

Networks are harvested when the performance reached 58% on the worst of the two screening outcomes. This turned out to always be the non-mass outcome. The harvesting criteria was determined empirically to provide several networks that performed well.

3.5. Voting

The final step in the process combines the outputs of the 24 networks into a single decision. Since each case consists of two image views, there are a total of 48 votes for each case. While we tested many voting methods, the following criteria proved best: If a simple majority of votes (25 or more) are tallied for an outcome, determine that it is that outcome (either mass or no-mass). If there is a draw (exactly 24 votes are tallied for each outcome), then the case is classified as indeterminate. Voting results under this criteria are given in Table 2.

In a practical situation, the tie in voting could be utilized to re-take the mammogram session. Here, the automated screening has a clear advantage since the repeat session can be done immediately and not require a second visit to the radiologist. Similarly, detection of a mass could indicate immediate referral for a more detailed diagnostic mammogram.

Erring in favor of declaring a mass when in doubt, however, gives a clear decision. In Table 3, these results are re-tabulated with ties declared to be masses. Under these conditions, sensitivity is 75% and specificity is 56%. This compares favorably with the 10% to 31% error rate for human mammographers stated earlier. It also demonstrates that our earlier preliminary study in which only 55 cases were used scales to the larger dataset and was not the result of sampling bias.

4. Summary and Future Work

4.1. Summary

The wavelet processing provides a lossless way of reducing a very large image into a canonical sequence of data suitable for filtering. The LOS-RAAM provides a partially unsupervised learning technique for discovering the structure hypothesized to exist within the data set and clustering extracts the attractors that give shape to the IFS-like space. The high-dimensional internal structure is reflected in the structure of the IFS produced and consequently also within the features of the FFVs and FFTMs. Supervised training leads to votes from independent images and from inde-

Table 2. Voting Outcome with Indeterminate Cases

Sets	Mass			Non-Mass		
	Correct	Incorrect	Draw	Correct	Incorrect	Draw
Training/ PAC	118	42	6	58	33	6
Hold Out	36	14	5	18	13	1

Table 3. Voting Outcome without Indeterminate Cases

Sets	Mass		Non-Mass	
	Correct	Incorrect	Correct	Incorrect
Training/PAC	124	42	58	39
Hold Out	41	14	18	14

pendent training runs which often combine productively to increase performance.

We hypothesize that FFVs and FFTMs, as introduced here, preserve important features of the image and provide an important abstraction in much the same way statistical physics permits description of chaotic situations from a small number of control variables.

As applied to mammography, our technique approaches human-level performance. More importantly, it demonstrates a viable technique for reducing arbitrary images, according to their structure, into a small set of control-like values that realistically can be used in neural network experiments. As each step is applied, the objective is to preserve some degree of reversibility so that analyzing FFVs and FFTMs remains very close to analyzing the structure represented in the original image.

The location hint, as an indication of where a mass can be located relative to human physiology, worked extremely well as an aid to producing a trained network capable of good generalization. This supports the notion that the underlying structure, and not simply regions of the image, are being extracted by our techniques.

4.2. Future Work

Since a large amount of data is required to train and prove our techniques, we continue to collect and digitize cases.

We are evaluating other wavelet transforms by visually evaluating combinations of transforms and thresholds in the hope that a better method of data reduction for that part of the process can be found.

Finding good hints to use during classification training can be challenging. In addition to location, we have recently added hints based on maximal mass diameter with normals having a diameter of zero. Other hints we are considering require additional analysis by the radiologist to determine degree of difficulty or the presence of additional mass-like structures that might confuse an automatic screening system.

Automating mammographic screening is an important goal. Even with recent advances toward developing blood tests to screen for cancers, breast cancer will still require images for full evaluation. If every woman received a mammogram according to the recommended schedule, there would be more images to interpret than there are radiologists capable of reading them.

Acknowledgements

The authors express their appreciation to Andrew Laine for providing the edge-enhancing wavelet transform from which this work began. We are also grateful to David Gur for his help in digitizing images.

Appendix A

A.1. Details of LOSRAAM Training

In this appendix we highlight the details of training a LOSRAAM network. For elaboration refer to Kalman and Kwasny[25].

A.1.1. The Error Term

As mentioned in the text we use a self-scaling error function. The error is definded by:

$$e_{pk} = (t_{pk} - a_{pk}) \qquad \text{(A1)}$$

and the error function is given by:

$$\Phi = \sum_k \sum_p g_{pk} \qquad \text{(A2)}$$

where p is a pattern, k identifies an output unit, g_{pk} is a function of the target value t_{pk} and the activation value a_{pk}. In Kalman and Kwasny[25], we give criteria that g_{pk} and its derivatives should meet. In short for linear and sigmoidal outputs g_{pk} is defined by:

$$g_{pk} = \begin{cases} \left(e_{pk}^2\right) & \text{Linear output} \\ \frac{e_{pk}^2}{1-a_{pk}^2} & \text{Sigmoidal} \end{cases} \qquad \text{(A3)}$$

Our sigmoidal function is:

$$a_{pk}(x) = \tanh(1.5x) \qquad \text{(A4)}$$

where x is the excitation computed from connections terminating at the output unit.

A.1.2. The Derivatives

For our conjugate gradient back-propagation training we use the generalized delta rule of Rumelhart and McClelland[36] to derive our derivative equations.

A.1.2.1. Simple Recurrent Networks. For simple recurrent networks the delta terms are given here. For the output units the delta terms are:

$$\Delta_{pk} = \begin{cases} (-2e_{pk}) & \text{Linear output} \\ 3\,(g_{pk}a_{pk} - e_{pk}) & \text{Sigmoidal} \end{cases} \qquad \text{(A5)}$$

The delta term for the hidden units is:

$$\Delta_{pj} = 1.5\left(1 - a_{pj}^2\right)\sum_k \Delta_{pk}w_{kj} \qquad \text{(A6)}$$

where j ranges over the hidden units, k is as in Eq. A6 and w_{kj} is the weight on the connection between units k and j.

For the feedback units the delta term is:

$$\Delta_{pj} = \sum_k \Delta_{pk}w_{kj} \qquad \text{(A7)}$$

where j ranges over the feedback units and k ranges over the hidden and output units.

The only derivative equation that affects the gradient and which involves feedback units is:

$$\frac{\partial \Phi}{\partial w_{ji}} = \sum_p \left[\Delta_{pj}a_{pi} + \sum_l \Delta_{pl}\frac{\partial a_{pl}}{\partial w_{ji}}\right] \qquad \text{(A8)}$$

where j ranges over the hidden units, i ranges over the input and feedback units and l ranges over the feedback units.

The remainder of the terms in the gradient are given by the generalized delta rule:

$$\frac{\partial \Phi}{\partial w_{ki}} = \sum_p \Delta_{pk}a_{pi} \qquad \text{(A9)}$$

where k ranges over output units and i ranges over input, feedback and hidden units. Equations for the bias derivatives are obtained by replacing the appropriate a_{pi} by 1.

Because of the recurrence in the network we must still consider the derivatives of the activations of the hidden units which become derivatives of the activations of feedback units for the next pattern in a sequence. Here is the equation for this case:

$$\frac{\partial a_{p+1,l}}{\partial w_{ji}} = \delta_{li}a_{pi} + \sum_s w_{js}\frac{\partial a_{ps}}{\partial w_{ji}} \qquad \text{(A10)}$$

where j ranges over the hidden units, i, l, and s all range over the feedback units, δ_{li} is the Kronecker delta function which is one if $l = i$ and zero otherwise.

A.1.2.2. LOSRAAM Derivative. Here we show the additional derivatives required for LOSRAAM training. As mentioned in the text the major difference between LOSRAAMs and simple recurrent networks is that there are output units which correspond to the distributed representation of the target. The self-scaling error function for that part of the output is:

$$\Phi_R = \sum_T \sum_k \frac{\left(R_{Tk} - R'_{Tk}\right)^2}{1 - {R'_{Tk}}^2} \qquad \text{(A11)}$$

where T ranges over the recursive auto associative memories in a sequence, k ranges over the recursive auto associative memory units, R_{Tk} is the target and input value(see figure 3) and R'_{Tk} is the output value. Part of the purpose of training is to reduce the difference between R_{Tk} and R'_{Tk}.

Since R_{Tk} depends on the network parameters we add a correction term to the generalized delta rule Eq. A9 to get:

$$\frac{\partial \Phi_R}{\partial w_{ki}} = S' + DR \qquad \text{(A12)}$$

where

$$DR = \frac{2\left(R_{Tk} - R'_{Tk}\right)}{\left(1 - {R'_{Tk}}^2\right)}\frac{\partial R_{Tk}}{\partial w_{ki}} \qquad \text{(A13)}$$

where S' is the derivative term which corresponds to Eq. A9.

Notes

1. We evaluated the use of both soft and hard thresholding, as defined by Burrus[11], and discovered that hard thresholding performed best for our purposes. We believe the importance of sharp, edge-like changes in intensity within the image accounts for this observation.

2. The notion of aggregating wavelet coefficients into vector form can be attributed to Dai[17].
3. The choice of K is determined empirically according to how many coherent clusters seem to be present. This part of the training is performed several times within a small range of K values. The best measure of coherence determines the one to be used.
4. Note that we also tried a discrete approach in which simple tallies were kept as each point was classified as belonging to a cluster. This provided low generalization indicating that the fuzzy process is a necessary part of finding structure in this way.
5. To avoid a misleading hint, these three units are managed during FFNN training so as not to contribute to the error function for non-mass (negative) cases.
6. As mentioned earlier, skip connections are not used for LOSRAAM networks since that would destroy their use for encoding data.
7. Although there are several standard databases of mammograms, we investigated these and determined that they were inappropriate for this work:

- University of South Florida database. 117 cases (61 with masses) digitized at 100 microns only. Too small/inappropriate digitization for our use.
- Nijmegen database. 40 cases (microcalcifications) Too small/wrong pathology/no technical specifications.
- Lawrence Livermore National Laboratories and UCSF. 50 patients (20 masses) digitized to 35 microns. Too small/inappropriate digitization for our use.
- Mammographic Image Analysis Society database. Size unknown, digitization to 50 microns but 8 bit depth. Inappropriate digitization for our use.
- Washington University digital mammography database. Digitally acquired 512 × 512 × 12 bits ROIs only. Inappropriate for our use.

8. Named after PAC (Probably Approximately Correct) learning, popularized by Valiant[38]. The main principle is that any hypothesis or theory that consistently makes correct predictions over a sufficiently large set of examples is not likely to have any serious flaws. Our PAC set is used for exactly that purpose.

References

1. Abu-Mostafa, Y.S., "Machines that learn from hints," Scientific American 1995; 272:64-69.
2. Allen, J., *Natural Language Understanding*, Benjamin/Cummings, 1995.
3. Anttinen, I., Pamilo, M., Soiva, M., Roiha, M., "Double reading of mammography screening films – on radiologist or two?" Clinical Radiology 1993, 48:414-421.
4. Asada N., Doi K., MacMahon H., Montner S.M., Giger M.L., Abe C., Wu Y., "Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases: pilot study," Radiology 1990; 177:857-860.
5. Baines, C.J., Miller, A.B., Wall, C., McFarlane, D.V., Simor, I.S., Jong, R., Shapiro, B.J., Audet, L., Petitclerc, M., Ouimet-Oliva, D., Ladouceur, J., Hebert, G., Minuk, T., Hardy, G. Standing, H.K., "Sensitivity and specificity of first screen mammography in the Canadian national breast screening study: a preliminary report from five centers," Radiology 1986, 160:295-298.
6. Barnsley, M.F., *Fractals Everywhere*, Academic Press, 1988.
7. Bird, R.E., Wallace, T.W., Yankaskas, B.C., "Analysis of cancers missed at screening mammography," Radiology 1992, 184:613-617.
8. Blank D.S., Meeden L.A., Marshall J.B., "Exploring the symbolic/subsymbolic continuum: a case study of RAAM," in *Closing the Gap: Symbolism vs. Connectionism*, Lawrence Erlbaum Associates, 1992.
9. Boone J.M., Gross G.W., Greco-Hunt V., "Neural networks in radiologic diagnosis: I. Introduction and illustration," Invest Radiol 1990; 25: 1012-1016.
10. Brodie I., Gutcheck R.A., "Radiographic information theory and application to mammography," Medical Physics 1982; 9: 79.
11. Burrus C.S., Gopinath R.A., Guo H., *Introduction to Wavelets and Wavelet Transforms, A Primer*, Prentice-Hall, 1998.
12. Chan H.P., Lo S.C.B., Sahiner B., Lam K.L., Helvie M.A., "Computer-aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network," Med Phys 1995; 22:1555-1567.
13. Chan H.P., Sahiner B., Petrick N., Helvie M.A., Lam K.L., Adler D.D., Goodsitt M.M., "Computerized classification for malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," Phys Med Biol 1997; 42: 549-567.
14. Chang Y.H., Zheng B., Gur D., "Computer-aided detection of clustered microcalcifications on digitized mammograms: a robustness experiment," Acad Radiol 1997; 4:415-418.
15. Chang Y.H., Zheng B., Gur D., "Robustness of computerized identification of masses in digitized mammograms. A preliminary assessment," Invest Radiol 1996; 31:563-568.
16. Clarke L.P., Kallergi M., Qian W., LI H.D., Clark R.A., Silbiger M.L., "Tree-structured non-linear filter and wavelet transform for microcalcification segmentation in digital mammography," Cancer Lett 1994; 77:173-181.
17. Dai X., "Wavelet applications in process sensor data analysis," doctoral dissertation, Washington University, 1996.
18. Ema T., Doi K., Nishikawa R.M., Jiang Y., Papaioannou J., "Image feature analysis and computer-

aided diagnosis in mammography: reduction of false-positive clustered microcalcifications using local edge-gradient analysis," Med Phys 1995; 22:161-169.

19. Giger, M.L., Yin, F., Doi, K., Metz, C.E., Schmidt, R.A., Byborny, C.J., "Investigation of methods for the computerized detection and analysis of mammographic masses," SPIE Med Imaging 1990, 1233:183-184.
20. Goldberg V., Manduca A., Ewert D.L., Giscoid J.J., Greenleaf J.F., "Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," Med Phys 1992; 19: 1475-1481.
21. Golub G.H., van Loan C.F., *Matrix Computations*, 2nd Ed, Johns Hopkins University Press, 1989.
22. Gross G.W, Boone J.M., Greco-Hunt V., "Neural networks in radiologic diagnosis: II. Interpretation of neonatal chest radiographs," Invest Radiol 1990; 25: 1017-1023.
23. Kalman B.L., Kwasny S.C., Abella A., "Decomposing Input Patterns to Facilitate Training," Proceedings of the World Congress on Neural Networks 1993;3:503-506.
24. Kalman B.L., Reinus W.R., Kwasny S.C., Laine A., Kotner L., "Screening entire mammograms for masses using artificial neural networks: preliminary results" Academic Radiology 1997; 4:405-414.
25. Kalman B.L., Kwasny S.C., "High performance training of feedforward and simple recurrent networks," Neurocomputing 1997; 14:63-84.
26. Kwasny S.C. and Kalman B.L., "Tail-Recursive Distributed Representations and Simple Recurrent Networks," Connection Science 1995; 7:61-80.
27. Laine A, Song S., "Wavelet processing techniques for digital mammography, presented at the Conference on Visualization in Biomedical Computing, Chapel Hill, NC, October 13-16, 1992.
28. Mallat S., Zhang S., "Characterization of signals from multiscale edges," IEEE Trans Pattern Anal Machine Inell 1992; 14: 710-732.
29. Newman W.F., Sproull R., *Principles of interactive computer graphics*, McGraw-Hill, 1979, 402-404.
30. Nishikawa R.M., Giger M.L., Doi K., Vyborny C.J., Schmidt R.A., Metz C.E., Wu Y., Yin F.F., Huo Z., Lu P., Zhang W., Ema T., Bick U., Papaioannou J., Nagel R.H., "Computer-aided detection and diagnosis of masses and clustered microcalcifications from digital mammograms," SPIE 1993; 1905:422-432.
31. Pollack J., "Recursive Distributed Representations," Artificial Intelligence 1990; 46:77-105.
32. Pollack J., "The induction of dynamical recognizers," Machine Learning 1991; 7:227-252.
33. Powell M.J.D., "Restart procedures for the conjugate gradient method," Mathematical Programming 1977; 12:241-254.
34. Reinus W.R., Wilson A., Kwasny S.C., Kalman B.L., "Parallel distributed processing systems as a means of diagnosing solitary lesions of bone," Investigative Radiology 1994; 29:606-611.
35. Reinus W.R., Kalman B.L., Kwasny S.C., "Artificial neural networks as a device to select patients for emergent cranial CT scans in emergency departments," Academic Radiology 1995; 2: 193-198.
36. Rumelhart, D.E., McClelland, J.L., and the PDP Research Group, *Parallel Distributed Processing, Volume 1*, MIT Pres, 1986.
37. Servan-Schreiber D., Cleeremans A., McClelland, J.L., "Encoding sequential structure in simple recurrent networks," Technical Report CMU-CS-88-183, Carnegie Mellon University, 1988.
38. Valiant L.G., "A theory of the learnable," Comm of ACM 1984; 27(11):1134-1142.
39. Wu Y., Giger M.L., Vyborny C.J., et al., "Application of neural networks to mammographic diagnosis of breast cancer (abstr)," Radiology 1990; 177(P):149.
40. Wu Y., Doi K., Giger M.L., Nishikawa R.M., "Computerized detection of clustered microcalcifications in digital mammograms: applications of artificial neural networks," Med Phys 1992; 19: 555-560.
41. Yin, F.F., Giger, M.L., Doi, K., Metz, C.E., Vyborny, C.J., Schmidt, R.A., "Computerized detection of masses in digital mammograms: analysis of bilateral subtraction images," Med Phys 1991; 18: 955-963.
42. Yin, F.F., Giger, M.L., Vyborny, C.J., Doi, K., Schmidt, R.A., "Comparison of bilateral-subtraction and single-image processing techniques int the computerized detection of mammographic masses," Invest Radiol 1993; 28: 473-481.
43. Zhang W., Doi K., Giger M.L., Wu Y., Nishikawa R.M., Schmidt R.A., "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," Med Phys 1994; 21:517-524.
44. Zhang W., Doi K., Giger M.L., Nishikawa R.M., Schmidt R.A., "An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms," Med Phys 1996; 23:595-601.
45. Zheng B., Chang Y.H., Gur D., "Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis," Acad Radiol 1995; 2:959-966.