An Analysis of Perceptual Errors in Reading Mammograms Using Quasi-Local Spatial Frequency Spectra

Claudia Mello-Thoms, Stanley M. Dunn, Calvin F. Nodine, and Harold L. Kundel

In this pilot study the authors examined areas on a mammogram that attracted the visual attention of experienced mammographers and mammography fellows, as well as areas that were reported to contain a malignant lesion, and, based on their spatial frequency spectrum, they characterized these areas by the type of decision outcome that they yielded: true-positives (TP), false-positives (FP), true-negatives (TN), and false-negatives (FN). Five 2-view (craniocaudal and medial-lateral oblique) mammogram cases were examined by 8 experienced observers, and the eye position of the observers was tracked. The observers were asked to report the location and nature of any malignant lesions present in the case. The authors analyzed each area in which either the observer made a decision or in which the observer had prolonged (>1,000 ms) visual dwell using wavelet packets, and characterized these areas in terms of the energy contents of each spatial frequency band. It was shown that each decision outcome is characterized by a specific profile in the spatial frequency domain, and that these profiles are significantly different from one another. As a consequence of these differences, the profiles can be used to determine which type of decision a given observer will make when examining the area. Computer-assisted perception correctly predicted up to 64% of the TPs made by the observers, 77% of the FPs, and 70% of the TNs.

KEY WORDS: computer-assisted perception, frequency spectra, wavelet packets, mammography, breast cancer detection.

BREAST CANCER is one of the leading causes of death among American women, and 46,000 women die from this disease each year. As with other types of cancer, early detection can significantly change the prognosis for a woman with this disease. Thus, renewed efforts have been made to develop accurate imaging techniques that can detect abnormalities of smaller and smaller sizes.

Nonetheless, a problem that usually is overlooked when considering such imaging techniques is the radiologist's ability to correctly interpret what is in the image. Studies have shown that 10% to 30% of all cancers in the breast are missed, and from these, approximately two thirds are seen in retrospect.[1] The missing of these cancers is caused by a variety of factors, such as search errors (43%), interpretation errors (52%), and suboptimal technique (5%).[2] Furthermore, in studies that examined how radiologists' search for cancerous lesions in the breast, using eye position monitoring, it has been shown that approximately 62% of the missed cancers are fixated with the high-resolution fovea.[3]

However, image-derived factors are not the only ones involved in perception, also, factors that come from within the observer have just as much influence.[4] Kundel[5] showed that 3 factors seem to be involved in the decisions made by the radiologists when searching for chest nodules (1) the prevalence of cancer in the population from which the patient is drawn; (2) the cost of making an incorrect decision or incorrectly mistaking normal tissue as being cancerous; and (3) the structure of the image around the nodule and at a distance from it. The third factor, which is directly related to the image, has been shown to have significant impact in image perception. In a study in which nodules in the chest were displayed embedded in an uniformly distributed background and in a real anatomic background,[6] the investigators found that the effects of the anatomic background, called "structured noise,"[7] were about 25 times higher than the effects of the variabilities in the image caused by the fluctuations in the number of x-ray photons reaching the receptors, called "quantum noise."[8] Note that structured noise affects lesion detection either by masking the lesion or by creating artifacts that resemble lesions, thus taking attention away from the real lesion when one is present.

Furthermore, it has been shown that lesions can be classified based on their Fourier spectra, be-

From the Department of Radiology, University of Pennsylvania School of Medicine, Philadelphia, PA, and the Department of Biomedical Engineering, Rutgers University, Piscataway, NJ.

This work was supported partially by the DAMD 17-97-1-7130.

Address reprint requests to Claudia Mello-Thoms, PhD, Department of Radiology, Imaging Research, Suite 4200, University of Pittsburgh, 300 Halket St, Pittsburgh, PA 15213-3180.

0897-1889/01/1403-0002$35.00/0
doi:10.1053/jdim.2001.28989

cause different abnormalities will have a different signature in the spatial frequency space.[7]

In this report we will take this argument a step further. Namely, we will consider the effects that different lesion and background spectra have on the observer's perception when reading mammograms. We will attempt to determine if there is a combination of image structures, the quasi-local combination of lesion and background, in the spatial frequency domain, in which lesion detection is facilitated, thus, leading to true-positives (TP) and to true-negatives (TN), or, where it is hindered, thus, yielding false-positives (FP) and false-negatives (FN).

MATERIALS AND METHODS

Eight experienced observers (3 mammographers from the staff of the Hospital of the University of Pennsylvania and 5 fellows undergoing training in Mammography at the same institution) read 5 2-view (cranio-caudal[CC] and medio-lateral oblique, [MLO]) mammogram cases (adding up to a total of 10 images). All cases had a malignant mass present; in 3 cases it was visible in both views, and in 1 case it was only visible in one view (CC). One case contained 2 malignant masses visible in both views.

These cases were obtained from the archives of the Hospital of the University of Pennsylvania. The films were digitized using a Lumiscan Model 100 digitizer (Lumisys Inc, Sunnyvale, CA), using a 100-μm spot size. The 2 views were displayed side-by-side on a single 19-inch, 2,048 × 2,048 gray scale monitor (GMA 201; Tektronix, Beaverton, OR), interfaced to a Sun Sparc computer (Sun Microsystems, Sunnyvale, CA).

The observers were instructed to search for malignancy and freely examine the cases until they felt confident to answer if a malignant lesion was present, and, if so, where it was located. The eye position of the observers was monitored during search using an ASL 4000 SU eye-head-tracking device (Applied Science Laboratories, Bedford, MA), and it was used to determine the areas in the image that attracted the observers' visual attention. Details of this experiment have been published elsewhere.[9]

Eye position was used because it is a good indicator of what in the image attracted the observer's visual attention. It has been shown that the observers dwell as long on lesions that were not reported (that is, FNs) as they do in the ones that were (TPs).[2,10]

At the end of the experiment the eye positions of each observer, for each case, were superimposed to the 2-view mammogram examined as shown in Fig 1.

For each case and each observer, 10 regions were extracted from the 2-view mammogram displayed. These regions were based on the decisions made by the observer when reading the case and the areas where the observer had prolonged (>1,000 ms) visual dwell. In this way, 5 cases × 10 regions × 8 observers generated 400 regions. These regions were rectangular windows of 128 × 128 pixels, which corresponded to 5° of visual angle at 38 cm viewing distance. They were extracted and processed using a program written in IDL (Research Systems, Inc, Boulder, CO).

The extracted regions were labeled with the purpose of generating a truth table that later could be used to compute the

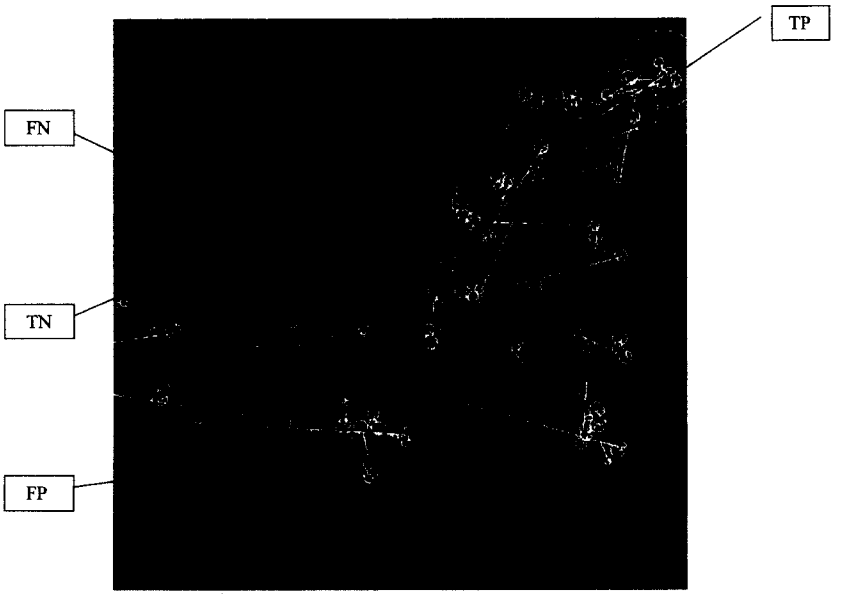


Fig 1. Example of eye-position monitoring when an expert is reading a mammogram case. The 2 larger circles correspond to the known location of a malignant mass. The smaller circles correspond to fixations (sequential dwells on areas of 0.5° of visual angle).

error in the system. To do so, we used the following criteria, which combined perceptual and decision-making elements.

For a True-Positive

To get assigned a TP label a region had to be reported by the observer, at the end of the session (thus, indicating that it had reached an internal suspicion criterion).

For a False-Negative

A region had to contain a malignant lesion that the observer failed to report. In this case an FN label was assigned regardless of the presence of visual dwell in the location of the lesion. It is important to mention that the number of FN regions was very small, but most of such regions received visual dwell for almost as long as the TP regions. Unlike the TP regions, here the observer did not report the lesion. This is in agreement with results previously reported in the literature.[3,10]

For a False-Positive

The observer had to have prolonged (>1,000 ms) visual dwell on a location in which no lesion existed. No differentiation was made between the cases in which the observer reported the presence of a malignant lesion at that location, and the cases in which the observer, after visual scrutiny of the area, decided that no lesion was present. This was because we hypothesized that any lesion-free area that attracted the observer's visual attention for as long as lesion-containing areas deserved further investigation, regardless of the observer reporting the presence of an abnormal finding in that area.

For a True-Negative

In this case the observer did not fixate, and therefore had no visual dwell in a lesion-free area. Thus, this was an area in an image that was free of abnormalities and free of fixations. We hypothesized that such areas did not attract the observer's visual attention, and because it did not contain any lesions, it was a clear TN. We felt that areas that did not attract any visual attention should be examined further to determine what makes them different from the areas that did attract visual attention. These criteria are summarized in Table 1.

To analyze the areas selected we sought inspiration in the human visual system. Visual signals are perceived through spatial frequency channels.[11] The processing carried out by the channels is effectively a filtering mechanism that operates within a specific frequency range. Furthermore, the outputs of the channels correspond to visual stimuli at different spatial scales.[12] In his highly influential book, David Marr[13] pointed out that the nervous system seems to prefer methods that run analysis from coarse to fine, thus, repeating the same process, but obtaining new information at the end of each cycle.

Table 1. Criteria Used to Label Each Image Piece

	Significant Dwell	Decision Criterion	Ground Truth
TP	Yes/no	Reported	Lesion present
TN	No	Not reported	No lesion present
FN	Yes/no	Not reported	Lesion present
FP	Yes	Reported/not reported	No lesion present

These properties of the visual system seem to indicate that an appropriate way to analyze visual data is by using a method capable of taking into consideration scale information, of being recursive, and yet introducing new information at each level. These are exactly the characteristics of a wavelet decomposition, and so we have decided to use wavelet packets to analyze these data.

Wavelet packets are a signal decomposition through a filter bank, in which the low-passed and the high-passed signal from one level serve as the input for the next level. The low-pass and high-pass filters in this filter bank are generated from a single function, called the wavelet prototype, through dilations, contractions, and shifts.[14] Note that this decomposition effectively separates the signal into different spatial frequency bands.

Each of the 400 local image patches extracted and labeled as previously described was analyzed by a wavelet packets tree using a Daubechies filter. These filters were chosen because they have compact support and smooth decay, and thus have good localization properties in space and are computationally efficient.

We have used the following convention to label the spatial frequency bands. The first number (0,1,2,3,4) refers to the position in the tree at which the band is located. The second number (0,1,2,3) can be interpreted as: 0, a low-pass filter was applied to the low-passed signal from the previous step; 1, a high-pass filter was applied to the low-passed signal from the previous step; 2, a low-pass filter was applied to the high-passed signal from the previous step; 3, a high pass filter was applied to the high-passed signal from the previous step.

The energy of the representation in each spatial frequency band was calculated as follows

$$Ei = (1/Ni) \sum x^2$$

where x is a vector containing the elements in each band, and Ni is the number of elements in the band.

To automatically characterize the type of decision outcome made by the observer when examining specific areas in the mammogram, we used a pattern classifier, an Artificial Neural Network (ANN). We chose ANNs because they are unbiased estimators, that is, they do not need a mathematical model that relates input to output to be known beforehand. In other words, they "learn from experience."[15] This enables them to learn adaptively, and sometimes intelligently, the patterns present in the data set.

An unsupervised learning clustering algorithm, the Adaptive Resonance theory (ART) algorithm, was chosen to train the ANN.[16] The main advantage of this type of algorithm is that the network itself decides how many classes are present in the data set, and how the classes are partitioned. One of the main disadvantages of this type of algorithm is that labeling has to be done in a separate step, after the network has converged to a state of equilibrium.

To generate the features for the ANN we run multiple analyses of variance (ANOVA), aiming at determining which spatial frequency bands contributed for the differentiation of the decision outcomes. This is in agreement with the notion that different spatial scales carry different information about the same input; thus, it may be possible to define an optimal

Table 2. Mean ± SD for the Energy Per Spatial Frequency Band

Band Name	Mean	SD	Band Name	Mean	SD
00	11,327.45	8,792.16	22	2.08	1.67
01	4.29	3.44	23	1.99	1.62
02	20.17	17.01	30	16.11	13.65
03	25.01	23.88	31	2.08	1.67
10	42,502.70	33,870.77	32	59.05	51.34
11	0.44	0.36	33	0.44	0.36
12	16.11	13.65	40	19.86	19.41
13	19.86	19.41	41	1.99	1.62
20	0.44	0.36	42	0.44	0.36
21	11.38	9.33	43	73.77	71.69

encoding for a particular input by carefully selecting the scales used.[11]

RESULTS

After processing each image patch as described previously, the values for the energy in each of the spatial frequency bands were calculated. This yielded the following values for the mean energy per spatial frequency band.

As Table 2 shows, there is a wide variability in the contribution of each of the bands. Some carry a great amount of information, which is reflected in the mean value for the energy in those bands, and some barely carry any information at all. Thus, the next step was to divide the bands into 3 types, by establishing thresholds between the spatial frequency bands: low energy ($x < 10$): 01, 11, 20, 22, 23, 31, 33, 41, 42; medium energy ($10 \leq x < 50$): 02, 03, 12, 13, 21, 30, 40; high energy ($x \geq 50$): 00, 10, 32, 43.

To assess the contribution of each band on the decision outcomes (ie, TP, FP, TN, FN) multiple ANOVA analyses were run, using the energy values as the dependent variable, and decision outcome as the independent variables, splitted by the spatial frequency bands.

For the high-energy bands, it was found that the band 00 ($F[3,382] = 6.206$; $P = .004$) and band 10 significantly contributed to differentiate TPs from FPs ($F[3,382] = 5.145$; $P = .017$).

For the low-energy bands, 6 bands marginally contributed to differentiate FPs from TNs: band 11 ($F[3,382] = 3.22$; $P = .023$), band 20 ($F[3,382] = 3.22$; $P = .023$), band 23 ($F[3,382] = 2.24$; $P = .033$), band 33 ($F[3,382] = 3.22$; $P = .023$), band 41 ($F[3,382] = 2.94$; $P = .033$), and band 42 ($F[3,382] = 3.22$; $P = .023$). Note that none of the medium-energy bands contributed to differentiate any decision outcomes.

In this way, TPs had greater energy than FPs on band 00 (Scheffe's test; $P < .018$) and 10 (Scheffe's test; $P < .011$). However, TNs had greater energy than FPs on the bands 11 (Scheffe's; $P < .085$), 20 (Scheffe's; $P < .085$), 23 (Scheffe's; $P < .105$), 33 (Scheffe's; $P < .085$), 41 (Scheffe's; $P < .105$), and 42 (Scheffe's; $P < .085$).

The importance of these results is that they seem to indicate differences in the energy contents, per spatial frequency band, of the regions in the image that yield different decision outcomes. The next step was to look at the energy breakdown for each of the decision outcomes to assess which spatial frequency bands carry the most information. The ranking information was obtained by deriving a table, like Table 2, for each decision outcome, and then hierarchically placing the energy contents from the band with the highest mean energy value to the 10th highest mean energy value. The final ranking is shown in the Table 3.

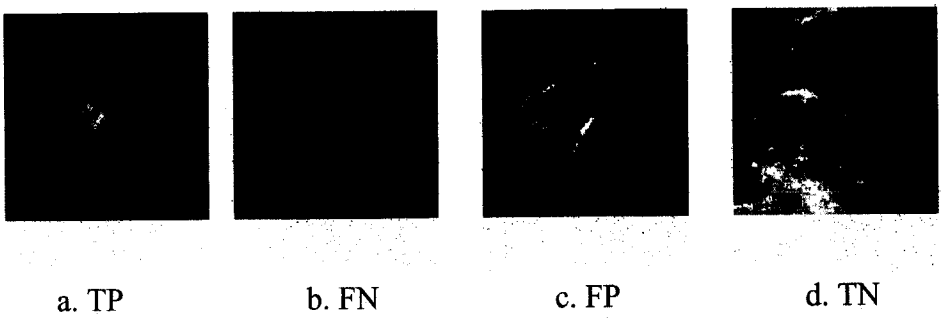
Note that for all of the decision outcomes except for the TPs, the energy mosaic is the same (albeit the mean values in each band may be quite different, which is critical for the Artificial Neural Network (ANN) to be able to predict the different decision outcomes). Moreover, the mosaic for the TPs is exactly the same as it is for the other decision outcomes, except for an inversion of information contents between the sixth and the seventh highest spatial frequency bands.

An example of the 4 decision outcomes and their energy profiles, in the spatial frequency domain, is shown in Fig 2.

The TP indicated in Fig 2 was found by all 8 observers; the FN was missed by all observers.

Table 3. Rank of Energy Bands, From Highest to Tenth Lowest, Per Decision Outcome

Rank	TP	FP	TN	FN
First	10	10	10	10
Second	00	00	00	00
Third	43	43	43	43
Fourth	32	32	32	32
Fifth	03	03	03	03
Sixth	13-40	02	02	02
Seventh	02	13-40	13-40	13-40
Eighth	12-30	12-30	12-30	12-30
Ninth	21	21	21	21
Tenth	01	01	01	01



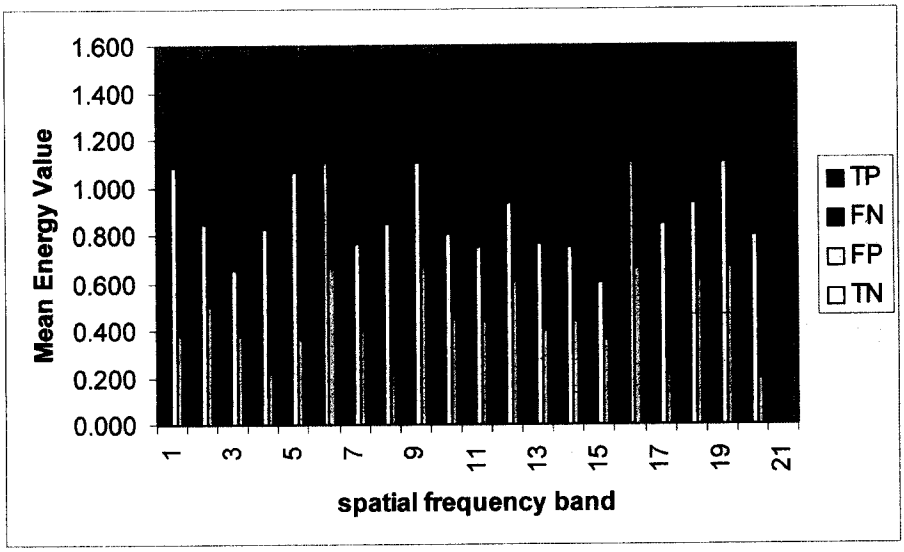


Fig 2. Example of the energy profiles yielded by the different decision outcomes. The lesion shown on A was reported by all observers; the one shown on B was missed by all observers. The artifact shown on C was mistaken as malignant by 6 observers, and the region shown in D was not scrutinized by any of the observers.

Additionally, 6 observers indicated incorrectly a lesion in the location shown as an FP, and none of the observers indicated a lesion on the location shown as a TN.

To select the parameters used for the Artificial Neural Network, we used the results of the ANOVA analysis, which showed that the energy values in the high- and low-energy bands significantly contributed to differentiate the decision outcomes. Additionally, a number from 1 to 8 was used to separate the observers. This was done because we reasoned that different observers could react differently to the same elements in the image. For example, an experienced observer may point out a subtle malignant lesion, whereas a less experienced observer may not see anything in the same location. In this way, the mapping performed by the network was from a 14-dimensional input space onto a 1-dimensional output space (TP, FP, FN, or TN).

Using an ART network with 14 neurons in the input layer, the results shown in Table 4 were obtained in terms of percentage of correct and incorrect responses per category.

The above results suggest that the energy profiles for the TPs, TNs, and FPs are unique and can be used to predict correctly each of these decision outcomes. However, the results for the predictions of the FNs were low. There are a variety of factors at play; for example, there was a very

Table 4. Percentage of Correct and Incorrect Decisions as Yielded by the ANN

Predicted Decision Outcome	Percentage Correct
TP	64
FP	77
TN	70
FN	28

limited number of such samples (18 *v.* 69 of the TPs, 60 of the FPs and 253 of the TNs), which hindered the network learning. Additionally, most FNs were mistaken as TPs by the network, which might indicate that, even though these lesions did not raise the observer's suspicious threshold high enough to make a decision in the location, their profile was similar enough to the profile of the TPs to make the network report the presence of a finding in the location (which is an underlying condition to a TP label). Currently, studies are being conducted with the purpose of improving the classification of the FN decision outcomes.

Regarding the correct prediction rates for TPs, FPs, and TNs, it is important to keep in mind that they depend strongly on the observer's self-consistency. The network cannot account for a clearly visible lesion that is not reported because the observer was distracted, as well as it cannot account for artifacts that are a product of the observer's imagination, rather than of image elements. In this way, we can say that the achieved prediction rates were appropriate, given these elements that cannot be taken into account.

DISCUSSION

The results obtained herein indicate that there is a particular configuration of energy, in the spatial frequency domain, that is associated with the detection of true lesions. This configuration is unique for the TPs, and it is hierarchically similar, albeit with different mean values, for the other decision outcomes. Furthermore, when deriving the energy ranking in the spatial frequency domain for areas of the image that led to the different decision outcomes, it becomes clear that for the TPs there is an inversion in the order of 2 spatial frequency bands. Thus, an interesting hypothesis rises, namely, that part of the FNs are missed because they are too "subtle," that is, the features that define the lesion are not strong enough yet to cause the shift in the information contents of these 2 bands, which characterized the TPs.

This topic needs further investigation, because of the small amount of data available for this experiment. Additionally, no strict restrictions were made when labeling the image pieces to be further processed between areas that had been fixated from the ones that had only been reported. For example, to assign a TP label the area had to be reported; for an FN it only needed not to be reported, whereas for a FP it could either be reported or not reported. Moreover, all TN labels were assigned to areas that did not attract any visual attention, and that were not reported as containing a lesion.

The image areas selected represent quasi-local features, because they are not produced by fixations, but rather by fixation clusters. Furthermore, in the wavelet analysis of the local image patches, no consideration was made regarding the spatial frequency composition of the whole image. It is clear that if such composition had been taken into account, by either working as an additional set of features or as a "modifier" of the features that were described previously, the results could have been quite different. For example, the energy configuration of the whole image could have been used to normalize the energy in the labeled areas. However, by doing so, the small local differences were masked by the distribution of the data in the entire image. Another possibility to take into account information regarding the whole image would be to add the image information as a set of additional features, but in this case one is doubling the number of dimensions in the problem space but adding fixed coordinates in these dimensions, because the image energy decomposition is the same, regardless of local decision outcomes. This biases the results of the pattern classifier artificially, because these new dimensions do not add any new information to the problem, they just introduce an artificial stability.

Another possible outcome of this work is that, because the energy configuration of the FPs is different from the one of the TPs, then perhaps it is possible to screen out some of the FPs generated by systems that automatically detect suspicious findings in the breast, based on this difference in energy. This might have a considerable application in an analysis of the outputs of computer-assisted diagnosis systems, because these systems in general produce a large number of FPs.[17]

REFERENCES

1. Martin JE, Moskowitz M, Milbrath JR: Breast cancer missed by mammography. Am J Roentgenol 132:737-739, 1979

2. Krupinski EA, Nodine CF: Gaze duration predicts the location of missed lesions in mammography, in Gale AG (ed): Digital Mammography. Amsterdam, The Netherlands, Elsevier Science, B.V, 1994

3. Krupinski EA, Nodine CF, Kunde HL: Enhancing recognition of lesions in radiographic images using perceptual feedback. Optical Engineering 37:813-818, 1998

4. Burgess AE: Image quality, the ideal observer, and human performance of radiologic decision tasks. Acad Radiol 2:522-526, 1995

5. Kunde HL: Predictive value and threshold detectability of lung tumors. Radiology 139:25-29, 1981

6. Kundel HL, Nodine CF, Thickman D, et al: Nodule detection with and without a chest image. Invest Radiol 20:94-99, 1985

7. Revesz G, Kundel HL: Feasibility of classifying disseminated pulmonary diseases based on their Fourier spectra. Invest Radiol 8:345-349, 1973

8. Eckstein MP, Whiting JS: Why do anatomic backgrounds reduce lesion detectability? Invest Radiol 33:203-208, 1998

9. Nodine CF, Kundel HL, Lauver SC: Nature of expertise in searching mammograms for breast masses. Acad Radiol 6:575-585, 1996

10. Kundel HL, Nodine CF, Krupinski EA: Searching for lung nodules—Visual dwell indicates locations of false-positive and false-negative decisions. Invest Radiol 24:472-478, 1989

11. Oliva A, Schyns PG: Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. Cognitive Psychology 34:72-107, 1997

12. Olds ES, Engel SA: Linearity across spatial frequency in object recognition. Vision Research 38:2109-2118, 1998

13. Marr D, Vision. New York, NY, W.H. Freeman, 1982

14. Wickerhauser MW: Adapted Wavelet Analysis From Theory to Software. Natick, MA, A.K. Peters, 1994

15. Kosko B: Neural Networks and Fuzzy Systems—A Dynamical Approach to Machine Intelligence. Upper Saddle River, NJ, Prentice-Hall, Inc, 1992

16. Freeman J: Implementing Neural Networks with Mathematica. New York, NY, Academic Press, 1994

17. Doi K, MacMahon H, Katsuragawa S, et al: Computer-aided diagnosis in radiology: Potential and pitfalls. Eur J Radiol 31:97-109, 1997