Human observer detection experiments with mammograms and power-law noise

Arthur E. Burgess,[a] Francine L. Jacobson, and Philip F. Judy
Radiology Department, Brigham and Womens Hospital, 75 Francis St., Harvard Medical School, Boston, Massachusetts 02115

(Received 28 June 2000; accepted for publication 22 January 2001)

We determined contrast thresholds for lesion detection as a function of lesion size in both mammograms and filtered noise backgrounds with the same average power spectrum, $P(f)=B/f^3$. Experiments were done using hybrid images with digital images of tumors added to digitized normal backgrounds, displayed on a monochrome monitor. Four tumors were extracted from digitized specimen radiographs. The lesion sizes were varied by digital rescaling to cover the range from 0.5 to 16 mm. Amplitudes were varied to determine the value required for 92% correct detection in two-alternative forced-choice (2AFC) and 90% for search experiments. Three observers participated, two physicists and a radiologist. The 2AFC mammographic results demonstrated a novel contrast-detail (CD) diagram with threshold amplitudes that increased steadily (with slope of 0.3) with increasing size for lesions larger than 1 mm. The slopes for prewhitening model observers were about 0.4. Human efficiency relative to these models was as high as 90%. The CD diagram slopes for the 2AFC experiments with filtered noise were 0.44 for humans and 0.5 for models. Human efficiency relative to the ideal observer was about 40%. The difference in efficiencies for the two types of backgrounds indicates that breast structure cannot be considered to be pure random noise for 2AFC experiments. Instead, 2AFC human detection with mammographic backgrounds is limited by a combination of noise and deterministic masking effects. The search experiments also gave thresholds that increased with lesion size. However, there was no difference in human results for mammographic and filtered noise backgrounds, suggesting that breast structure can be considered to be pure random noise for this task. Our conclusion is that, in spite of the fact that mammographic backgrounds have nonstationary statistics, models based on statistical decision theory can still be applied successfully to estimate human performance. [DOI: 10.1118/1.1355308]

Key words: anatomical backgrounds, contrast-detail diagrams, image quality, mammography, observer models

I. INTRODUCTION

Lesion detection in mammography can be very difficult because it must be done in the presence of complex breast structure that is often highly variable over the mammogram and differs from patient to patient. In the context of chest radiography, Revesz *et al.*[1] referred to this as the lesion conspicuity problem. For example, Kundel showed[2] that there was a dramatic difference in lung nodule detectability between radiographs that included only imaging system noise and radiographs that had noise plus chest structure. The rather limited literature on investigation of lesion detectability in anatomical structure is reviewed[3] by Samei. In medical imaging, anatomical variation is often referred to as ''structured noise'' even though it may not be due to a random process. In contrast, the visual psychophysics literature tends to use the term ''masking''[4] to describe all processes that reduce signal detectability and applies the term to both random and deterministic cases. Random ''maskers'' usually have a stronger effect on signal detection then deterministic maskers. In addition humans can learn a random noise image if it is repeatedly used, so that it eventually has a masking effect like a deterministic pattern[5,6] rather than a random noise effect. Some patient structures, such as ribs and hearts, are partially deterministic in that we all have them and they are usually in the same place. So it is unlikely that one could safely consider such structures to be noise even though there is some statistical variation. The question as to whether normal breast structure in mammograms can be considered a form of random noise is not so clear cut.

Bochud *et al.*[7] investigated this ''breast structure as random noise'' question using 2AFC detection experiments. They added a simulated nodule (8 mm diameter) and two simulated microcalcifications (0.22 mm) to both low and medium variability mammographic backgrounds. The ratio of estimated power spectral density at one cycle per cm was about 10 for the two image sets. They also added various amounts of simulated imaging system noise and then compared human results with predictions of a nonprewhitening observer model (NPWE) with an eye filter under two assumptions (A) that patient structure could be treated as either a deterministic masking process (with no effect on detectability) or (B) treated as a pure noise process. Their results depended on the nature of both the signal and the structure. Human results for the nodule fell between the deterministic

masking and pure noise predictions for both background classes. Even so, the effect of the structure on nodule detection was about 30–60 times that of the imaging system noise alone. Results for the microcalcifications depended on background class. For the low variability patient structure images, detection was completely limited by imaging system noise and structure seemed be deterministic. For the medium variability patient structure images, detection was limited by the combination of imaging system noise and patient structure, which appeared to act as a pure noise process. Bochud *et al.*[8] compared 2AFC detectability index (d') results for human and NPWE observers detecting a simulated nodule (10 mm diam) in an collection of digital mammogram backgrounds and collections of filtered noise backgrounds with the same power spectrum. They found that mammographic backgrounds did not degrade 2AFC detection performance as much as would be expected on the assumption that the structure can be treated as a form of pure random noise. Even if mammographic structure were to act like random noise, its statistics appear to be nonstationary.[9] It is important to determine whether, in spite of this fact, detectability calculations for any of a variety of model observers are still in reasonable agreement with human observer performance. This question can only be answered by empirical investigation of performance using a variety of decision tasks.

The work presented in this paper will use hybrid images with both simulated and realistic tumor signals added to mammographic backgrounds. There were two goals. (A) To determine experimentally the variation of lesion detection with lesion size. Our results will be presented in the form of contrast-detail (CD) diagrams—plots of lesion contrast thresholds as a function of lesion size with a maximum range of 0.5–15.6 mm. Thresholds are defined here as the lesion amplitude (contrast) required for 92% correct performance to the 2AFC detection task. CD diagrams for detection in white noise or imaging system noise have been found to have a characteristic form—thresholds decrease as lesion size increases.[10,11] Our experimental results for mammograms were completely different—CD diagram slopes were positive and thresholds decreased as lesion size increased. (B) Our second goal was to determine which of a variety of models based on statistical decision theory can predict these unusual human results. The models included the prewhitening matched filter, the NPWE model, and two versions of a linear discriminator model. We first describe the models, then the experimental methods and finally the detailed results.

II. THEORY

Human observer signal detection performance limited by image noise and background structure fluctuations has been evaluated using a variety of models based on matched filters.[12] They can be divided into three main classes depending on their ability to compensate for spatial correlations in the noise and structure. The classes are referred to as prewhitening (PW), partially prewhitening due to the presence of spatial frequency channels of the type believed to be present in the human visual system,[13] and nonprewhitening (NPW). The models can also be modified to include several other known human limitations such as an eye filter and internal noise. The eye filter equation used here will be $E(f)=f\exp(-bf)$, where f is radial spatial frequency, $f=(u^2+v^2)^{1/2}$, u and v are Cartesian spatial frequencies, and b is selected to give a peak at four cycles per degree of visual angle. There are two forms of internal noise, induced and static.[14] The induced internal noise leads to an apparent spectral density of the combined image noise and background structure fluctuations, $P_A=(1+\varphi)P(u,v)$, where $P(u,v)$ is the measured external spectral noise density and measured values of φ have been in the range from 0.3 to 1.0 for white noise[14] and filtered lowpass noise.[15] Static internal noise is believed to be one of the sources of human contrast sensitivity limitations in noiseless images and is mainly determined by display luminance. A variety of experimental results suggest[14] that the spectral density of static internal noise is small compared to the total spectral densities found in medical images. The issue of how well humans can compensate for spatial correlations (prewhiten) is still an active are of investigation. In general, human experimental results have fallen between the two extremes of complete prewhitening and nonprewhitening.[12,15–20] At present, the best predictions of human performance are made with partially prewhitening models that include arrays of spatial frequency filters (channels).[21,22]

The observer models can be described in several ways. One approach is for continuous images $g(x,y)$ with additive signals $s(x,y)$, noise $n(x,y)$ and statistically defined backgrounds, $b(x,y)$. If the images are discrete then (x,y) can be considered to be pixel row and column addresses. Alternatively, the digital images and components can be described in column vector notation with pixel values in lexicographic order.[23] Starting in the upper left corner and proceeding row by row, the image pixel values are entered as a one-dimensional list of elements in the vector. The vector has length N_p, where N_p is the number of image pixels. In this notation, the image and components are **g**, **s**, **n**, and **b**. The total image is the described by the two alternative forms

$$g(x,y)=s(x,y)+n(x,y)+b(x,y) \quad \text{and} \quad \mathbf{g}=\mathbf{s}+\mathbf{n}+\mathbf{b}. \tag{1}$$

Model observer performance calculations are done using the ensemble statistical properties of the noise and backgrounds—based on collections of images. Analysis is straightforward if both can be treated as wide-sense stationary Gaussian noise—with first- and second-order statistics (mean and covariance matrices or power spectra) independent of position. In this case, model descriptions are most conveniently done in the spatial frequency domain with the signal described by its Fourier transform, $S(u,v)$ and the stochastic components (external noise, backgrounds and internal noise) described by their power spectra, $P_n(u,v)$, $P_b(u,v)$ and $P_{\text{int}}(u,v)$, respectively. For discrete images, the frequencies (u,v) are also discrete. For the case of nonstationary statistics, analysis can be done using discrete models in the spatial domain[12,23] with noise and background fluc-

tuations described by their covariance matrices, $\mathbf{K_n}$ and $\mathbf{K_b}$. A variety of models will be presented in the remainder of this theoretical section.

A. Ideal observer

The ideal observer[24–26] provides an important absolute scale for evaluating human performance. Noise fluctuations set a fundamental limit on how accurately one can detect, locate or identify signals and estimate signal parameters. One can define a hypothetical ideal observer that ''does everything right'' and operates at this performance limit. The ideal observer uses Bayes' theorem to combine *a priori* information (about the signal profile and possible locations for example) with new data optimally extracted from the image to obtain probabilistic estimates of the correctness of various hypotheses about what signals are ''seen'' and where they are located in an image. The optimum strategy for collecting and weighting image data depends on task details, including the statistical properties of the signals, noise and background structure. For example, in the simple SKE case, the optimum strategy for new data collection involves cross correlation. Given a total noise power spectrum $P_t(u,v)$. The ideal observer prewhitens the image (decorrelates the noise) using a filter $F(u,v)=1/(P_t(u,v))^{1/2}$ and uses a detection template with frequency response $T(u,v)=S(u,v)F(u,v)$ to cross correlate with the filtered image data. An alternative (and equivalent) implementation is to eliminate the image prewhitening stage and use a template with frequency response $T_2(u,v)=S(u,v)/P_t(u,v)$ for cross correlation with the original, unfiltered image at potential signal locations. The cross correlation result is then used as the decision variable. For 2AFC tasks and multiple-alternative forced-choice (MAFC) tasks, the alternative with the highest value cross correlation is selected. This template cross-correlation procedure is equivalent to the use of a prewhitening matched filter. The ideal observer detectability index equation for the SKE detection task with Gaussian, stationary noise and backgrounds is given by

$$(d'_{\text{ideal}})^2=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\frac{|S(u,v)|^2 du dv}{[P_n(u,v)+P_b(u,v)]}\right]=\mathbf{s}^t(\mathbf{K_n}+\mathbf{K_b})^{-1}\mathbf{s}, \quad \text{where } t \text{ is the transpose.} \tag{2}$$

Evaluation of task performance by other observers, either model or human, can be made by comparing their performance with the ideal observer.[24,26] Efficiency is defined as the square of the detectability index ratios for the observer under test, $d'_{t'}$, and the ideal observer doing the same task

$$\eta=(d'_t/d'_{\text{ideal}})^2. \tag{3}$$

B. NPWE model

The nonprewhitening matched filter (NPW) model was proposed by Wagner *et al.*[27] It uses the expected signal (after imaging system filtering) as an internal cross-correlation template. The model is suboptimal because it cannot compensate for noise correlations (prewhiten) and it has no mechanism for estimating and compensating for local background levels. The model reduces to the ideal observer model for SKE/BKE tasks in white noise. The NPW model fits human results for white noise and noise with CT-like power spectra[16,28,29] on uniform backgrounds but does not agree with detection results in statistically defined (lumpy) backgrounds.[17] This is due to the fact that the NPW cross-correlation procedure includes local variations in background intensity in the decision variable value. However, this model can be modified to estimate and compensate for local background mean levels by either including zero frequency [direct current (dc)] suppression[30] or by adding an eye filter which has no DC response.[18,31] The latter case will be referred to as the NPWE model and its detectability index for the SKE case with Gaussian, stationary noise and backgrounds is given by

$$(d'_{\mathbf{NPWE}})^2=\frac{[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|S(u,v)|^2|E(u,v)|^2 du dv]^2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{|E(u,v)|^4(1+\varphi)[P_n(u,v)+P_b(u,v)]+P_{\text{int}}\}|S(u,v)|^2 du dv}. \tag{4}$$

C. Channelized FH (Fisher–Hotelling) model

Ideal observer performance can only be calculated in a straightforward manner for very simple decision tasks[32] such as the SKE detection task described above. In more complicated cases, detection of signals defined statistically for example, the ideal observer would use nonlinear procedures and analysis can become mathematically intractable. Fiete *et al.*[33] developed a model that obtains the best possible results using linear procedures. They referred to this as the Hotelling observer. It will be referred to here as the Fisher–Hotelling (FH) model since it is very similar to the Fisher linear discriminant model. The distinction is that the Fisher linear discriminant model uses sample statistics while the Hotelling model uses population statistics. For a collection of real images, one only has access to sample statistics—which are estimates of the population statistics. For simulated images, the population statistics are known. One important advantage of the FH model is that it can be used when image noise or background structure statistics are nonstationary. Statistical fluctuations in the image are described by the im-

age covariance matrix, $\mathbf{K_g}$. The SKE detectability index for this observer model for Gaussian noise and backgrounds is given by[12]

$$(d')^2 = \mathbf{s^t K_g^{-1} s}. \tag{5}$$

This model is difficult to use in practice for nonstationary noise because of the large number of matrix elements to be determined: The statistical covariation of every image pixel with every other pixel must be described. For a collection of images with N_p pixels per image, their covariance matrix would have dimension $(N_p)^2$. For example, the lesions used in the experiments described below are as large as 128×128 pixels and the prewhitening templates are twice as large. Therefore, the smallest feasible value of N_p is 65 536 (i.e., 256^2) and there would be 4×10^9 covariance matrix elements. For collections of real images such as mammograms (as opposed to simulated images), the covariance matrix must be estimated in a manner analogous to Eq. (16) below. On the order of 10–100 times N_p independent images are needed to obtain an estimate of $\mathbf{K_g}$ that can be for reliably inverted.[34] Clearly it is not feasible to obtain about 10^6 independent mammographic background regions for analysis.

Barrett *et al.*[34] suggested a method of overcoming the dimensionality problem—by representing the image data in the vicinity of the signal using smooth functions of the pixel values rather than the pixel values themselves. This is similar to the receptive fields (or equivalently spatial frequency channels) used by the human visual system. For the radial component, Barrett *et al.* chose to use Laguerre–Gauss (LG) functions, the product of Laguerre polynomials and Gaussians. They found that six radial functions were adequate. Their basis set also included allowance for angular dependence. Other basis (channel) sets have been used. Myers and Barrett[21] incorporated Rect function spatial frequency channels in the prewhitening observer model since Myers *et al.* had shown[16] that prewhitening models were clearly unable to account for human performance with highpass noise power spectra. Barrett *et al.*[22] showed the FH model with Rect function channels predicted human observer performance within about $\pm 25\%$ accuracy for a variety of tasks. More recent results[19,35] show that the simple Rect channels model is inadequate for some experimental situations. More complex channelized FH models can obtained using physiologically reasonable two-dimensional (2D) filter mechanisms. The filters can be multiresolution hierarchies[4,13] tuned in both radial frequency and angle or more simple filters with no angular dependence for isotropic signals. Each channel filter, $T_k(u,v)$, prevents any subsequent prewhitening (decorrelation) of its contribution to the overall observer decision process, so model efficiency is reduced somewhat. Burgess *et al.*[19] used difference of Mesa filter channels[36] and found good agreement with human results for signal detection in two-component noise. Eckstein *et al.*[37] had success using Gabor function channels.[38] Abbey used[39] difference of Gaussians filter channels.[40] Given the channel output decorrelation procedures used by the FH observer model, performance estimates are not sensitive to the particular choice of channels functions.[39]

Typically a total number, N_c, of six to eight radial channels are adequate for the simple isotropic signals and power spectra used in medical imaging signal detection investigations. Experiments on orientation effects in human visual experiments indicate that channel angular bandwidths are about 30 degrees. This suggests that about 40–50 two-dimensional basis functions should be adequate to model human performance for nonisotropic signals. Each channel is described in the spatial domain by a basis function $t_k(x,y)$ or alternatively by a vector, $\mathbf{t_k}$, of dimension N_p. The channel set $\{\mathbf{t_k}, k=1,2,...N_c\}$ is combined to form an $N_p \times N_c$ dimensional matrix, $\mathbf{T}$, whose columns are the individual channel vectors. The signal response vector after the channels is $\mathbf{r_s} = \mathbf{T^t s}$, with dimension N_c. The covariance matrix of the image data after the channels becomes $\mathbf{K_C} = \mathbf{T^t K_g T}$, with dimension $N_c \times N_c$. The important point is that one can determine the elements of $\mathbf{K_C}$ directly from the image collection. This will be discussed in the methods section. The reduction in dimensionally allows one to obtain a reliable estimate of the covariance matrix with a reasonable number of images. For N_c equal to 6, $\mathbf{K_C}$ is a 6 by 6 matrix which can be reliably estimated with a few hundred images. Human observer induced internal noise can be included by multiplying $\mathbf{K_C}$ by the scalar $(1+\varphi)$. Observer static internal noise is assumed to be zero mean and uncorrelated between channels so it is represented by a diagonal matrix, $\mathbf{K_{int}}$, with elements equal to a common variance. The static internal noise contribution is usually very small and for purposes of model analysis, it is sometimes useful to neglect it. If $\mathbf{K_{int}}$ is set to zero, then the induced internal noise effect can be interpreted as defining the human observer statistical efficiency, $\eta = (1+\varphi)^{-1}$. The detectability index for this model is given by[23]

$$(d'_{\text{chan}})^2 = \mathbf{r_s^t}[(1+\varphi)\mathbf{K_C} + \mathbf{K_{int}}]^{-1}\mathbf{r_s}. \tag{6}$$

D. CD diagram estimation

The purpose of the present experiments was to measure the CD diagrams for simulated and realistic lesions in simulated and realistic mammographic backgrounds. The equation describing the CD diagram for the ideal observer will be calculated for a special case to illustrate the general form of the expected result. The signal will be assumed to be isotropic with spatial domain polar equation, $s_R(r) = \alpha s(r/R)$, where r is radial distance from the center and R is a positive real-valued size scaling factor. The profile $s(r/R)$ has unit amplitude and α is an amplitude scaling factor. The 2D Fourier transform of this scalable signal is[41]

$$S_R(f) = \alpha R^2 S(Rf). \tag{7}$$

Normal breast structure will be assumed, for illustration purposes, to be described by isotropic, stationary Gaussian noise with a two-dimensional power spectrum slice described by the equation $P_b(f) = B/f^\beta$. Power spectrum estimates[7,42–44] for mammograms gave average estimates of β of about three and indicate that this power law holds up to frequencies of

about 1 cycle/mm, indicating that image quantum noise is unimportant for detecting signals larger than about 1 mm in diameter. Therefore the imaging system noise power spectrum, $P_n(f)$, will be neglected. One open question is whether the detectability index integral will converge as the integration limit goes to infinity, so a high frequency cutoff, f_2, will be used. Substituting in the polar equivalent of Eq. (2), the signal detectability index for the ideal observer under SKE conditions becomes

$$(d')^2 = 2\pi \int_0^{f_2} \frac{|S_R(f)|^2 f\,df}{P(f)} = \frac{2\pi\alpha^2 R^4}{B} \int_0^{f_2} |S(Rf)|^2 f^{(1+\beta)}\,df. \tag{8}$$

This equation can be transformed to a new coordinate system with radial frequency $\phi = Rf$ to give

$$(d')^2 = \frac{2\pi\alpha^2 R^{(2-\beta)} I_S}{B} \quad \text{where} \quad I_S = \int_0^{\phi_2/R} |S(\phi)|^2 \phi^{(1+\beta)}\,d\phi. \tag{9}$$

The value of the integral I_S depends on the signal spectrum and the exponent of the power-law noise but does not depend on either the signal spatial or amplitude scaling factors as long as the signal bandwidth is small compared to the cutoff frequency (more on this point below). The integral can be solved exactly[45] for the special case of a Gaussian signal profile with the upper limit of integration equal to infinity. Using $S(v) = \exp(-\phi^2)$, the result is

$$I_S = \frac{1}{2^{(\beta+4)/2}} \Gamma\left(\frac{\beta+2}{2}\right), \tag{10}$$

with a range of validity $\beta > -2$. One can define a detection threshold amplitude, A_t, using the criterion of d' equal to some arbitrarily selected constant, δ. Then the signal amplitude required to reach this amplitude as a function of signal scale factor R is given by

$$A_t = \delta \sqrt{B/2\pi I_S} R^{(\beta-2)/2}. \tag{11}$$

For comparison purposes, the conventional logarithmic representation is more convenient. The relationship between log(amplitude threshold) and log(signal size scale) then becomes

$$\log(A_t) = C + m[\log(R)], \quad \text{where} \quad C = \log(\delta\sqrt{B/2\pi I_S}) \quad \text{and} \quad m = (\beta-2)/2. \tag{12}$$

Since C is a constant for a given signal and power spectrum exponent, the slope of the constant threshold line on CD plot is $m = (\beta-2)/2$. As an example, for the average mammographic background power spectrum, $1/f^3$, the predicted slope is $+0.5$. We evaluated Eq. (8) by numerical integration for the simulated nodule signal used in some of our experiments for β from 0 to 4 and found the same linear relationship between power-law exponent and CD diagram slope.

There are constraints on application of the above heuristic argument about CD diagram slopes. First, it is necessary to observe the assumption of profile scaling. If the (radius normalized) signal profile changes as radius changes, then the analysis is no longer valid. This could occur when a scalable signal is smoothed by a filter of fixed bandwidth. A more serious limitation is that the energy spectra of many signals do not decay fast enough with frequency to allow convergence of the integral I_S. This is particularly important for analysis of digital signals, where I_S should approach its limiting value by the Nyquist frequency (0.5 cycles/pixel). This can be illustrated using the designer nodule signal,[20] which is radially symmetric (isotropic) with spatial domain form, $s(\rho) = A^*\text{Rect}(\rho/2)(1-\rho^2)^v$, where ρ is a normalized distance (r/R), R is the nodule radius and A is signal amplitude. The nodule profile can be changed (designed) by changing the value of v. For example, the result is a sharp-edged, flat-topped disc for v equal to zero and the projection of a sphere for v equal to 0.5. The nodule profile for v equal to 1.5 gave the best fit to lung nodule measurements by Samei *et al.*[46] and was used in some of our experiments. The designer nodule has a convenient Hankel transform, $S(v,R,f)$, proportional to $J_{v+1}(2\pi Rf)/(2\pi Rf)^{v+1}$. Envelopes of Bessel functions, $J_n(x)$, decay as approximately $x^{-0.5}$, independent of the order n. This means that the envelope decay of $S(v,R,f)^2$ is approximately proportional to $f^{-(3+2v)}$. More importantly, the envelope of the integrand of I_S is proportional to $f^{(\beta-2-2v)}$. For example, with a sharp-edged disc signal $(v=0)$ the integral, I_S, does not converge for power-law spectrum exponent β greater than 2. Recall that the integral is due to a convenient approximation that neglects imaging system noise. When this noise is included in the analysis then convergence is ensured, but CD diagrams will be more complicated.

III. MATERIALS AND METHODS

The experimental methods will first be summarized and details will follow. The human observer experiments were done using hybrid digital images with four combinations of conditions: realistic or simulated mass lesions digitally added to normal mammographic structure or simulated ('$1/f^3$' filtered noise) backgrounds. The realistic mass lesions were extracted from carefully selected, digitized radiographs of breast biopsy tissue. The mammographic backgrounds were square regions (61×61 mm on the film) selected from the constant breast thickness region of digitized mammograms. A log exposure data scale was used. The regions were displayed as 512×512 pixel arrays on a monochrome cathode ray tube (CRT). An example 2AFC image is shown in Fig. 1. Observer signal detection performance was measured under SKE conditions using 2AFC and search methods. Three observers took part in the experiments, two physicists and a board certified radiologist. There was no statistically significant difference in performance between observers. Five experiments were done, the first four using the 2AFC method. The first was designed to determine the CD diagram using four different extracted masses added to mammographic

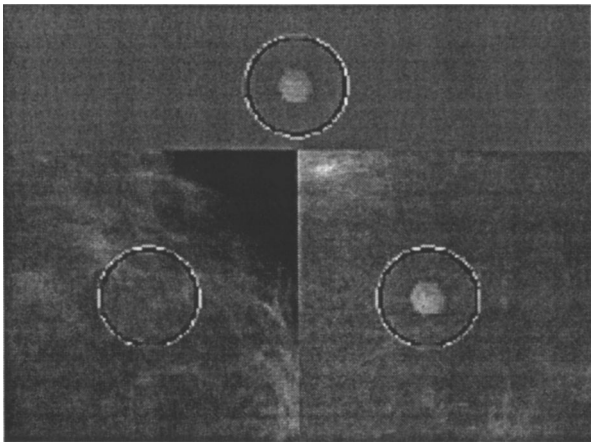

FIG. 1. Example display for the 2AFC observer experiments done using a monochrome monitor. A high contrast reference copy of the lesion is shown above the mammographic backgrounds, each 512×512 (61×61 mm with 0.12 mm pixels on the mammogram, 0.29 mm pixels on the monitor). The two possible lesion locations in the fields are surrounded by unobtrusive circle cues (Ref. 56). The lesion is added to the background on the right and the circles are exaggerated for publication.

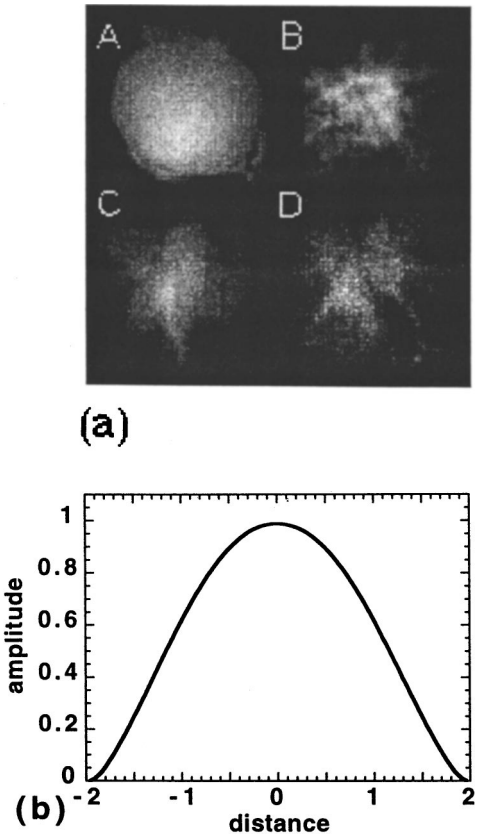

FIG. 2. (a) Images of the four extracted mass lesions used in the experiments, fibroadenoma (*A*) and invasive ductal carcinomas (*B,C,D*). Original sizes were from 8 to 18 mm. (b) The profile of the simulated nodule signal used in some experiments, based on a diameter of 4 mm. The profile equation form is $s(\rho)=A^{*}\mathrm{Rect}(\rho/2)(1-\rho^2)^{1.5}$, where ρ is a normalized distance (r/R), R is the nodule radius and A is an amplitude scaling parameter.

backgrounds. An adaptive display window and level procedure was used to maximize displayed image contrast while avoiding data clipping. The second experiment was done using white noise backgrounds as a control to determine whether we would obtain the traditional CD diagram form with our display conditions. The third experiment was designed to investigate the effects of CRT display contrast variation on the CD diagram for realistic mass detection in mammographic backgrounds. The fourth experiment was done using three hybrid image combinations: A simulated mass in mammographic backgrounds, a simulated mass in simulated mammographic backgrounds, and a realistic mass in simulated backgrounds, The final experiment was done under search conditions with one signal always present within a defined area. The observer used a trackball and cursor to select the most probable location. The goal of these experiments was to compare human and observer model performance over a wide variety of conditions.

A. Materials

Four realistic breast lesions (1 fibroadenoma and 3 ductal carcinomas) were selected from a data base of 25 digitized specimen radiographs. In all cases, the lesions were clearly visible, there were no microcalcifications, surrounding tissue was very uniform with a minimum of complex tissue overlapping the masses and complete biopsy records were available. The four selected masses had particularly well defined boundaries, which simplified the segmentation and extraction process. The four selected lesion images are shown in Fig. 2(a), items *A* to *D*. Their original sizes were 11, 18, 8, and 14 mm, respectively. The specimen radiographs were digitized by Dr. J. Beutel of Sterling Diagnostic Imaging using a Lumisys digitizer with 50 micron sampling and 12 bits/pixel. We then estimated the background in the vicinity of the tumor and subtracted it to obtain a digital image of the isolated tumor. The background was estimated using a quadratic surface method[47] as follows. A 3×3 square grid of points was centered on the lesion with the midpoints of the grid edges close to, but not overlapping, the nearest lesion point. The image data values for the eight grid points on the boundary were recorded and the value of the central grid point was then estimated. The nine values were used to calculate a quadratic surface. This surface was subtracted from the specimen image and negative values were set to zero. The resulting first estimate of the ''lesion only'' image was compared to the specimen image. The visual disparity was used to estimate perturbations to the nine grid values and the revised values were used to calculate a second quadratic surface. This interactive procedure was repeated a number of times until there was a satisfactory agreement between the appearance and boundaries of the extracted lesion and the lesion in the specimen image. Since the masses had different sizes, they were all dimensionally rescaled using bilinear interpolation so that their boundaries just fitted within a 256×256 array. Their amplitudes were also scaled so all peak values (amplitudes) were equal.

The mammographic backgrounds were obtained from 210 digitized normal craniocaudal mammograms provided by Dr. Maria Kallergi of the University of South Florida. The criterion for selecting the normal cases were: No artifacts, no

microcalcification clusters, absence of visible disease (BI-RADS category one) and a follow up examination two years later that also was disease free. The films had been digitized using a DBA model R3000[48] with 30 micron sampling (14 bits per pixel) and then reduced to 120 micron sampling for experiments. We wanted to investigate the effect of normal breast parenchymal structure on lesion detection and, therefore, needed to eliminate the confounding effect of the large, systematic thickness variations at the periphery of the breast. Several methods were developed[43] to estimate the boundary of the constant breast thickness region between the compression plates. It was necessary to select a data scale for this work. The possibilities were x-ray exposure (E), log exposure ($\log E$) or optical density (OD). We concluded that the best choice[49] for our purposes is $\log E$ units because, to a good approximation, $\log E$ differences are proportional to physical attenuation differences in the patient.

The $\log E$ scale has the advantage of being independent of display method. Within the constant thickness region, $\log E(x,y) \approx \log E_0 - \mu_z(x,y)T$, where $\mu_z(x,y)$ is the average linear attenuation variation along the x-ray projection ray path from the image position to the x-ray focus, T is the tissue thickness and E_0 is image plane exposure with no attenuation. Therefore, image amplitude variations, when expressed in $\log E$ units, are directly proportional to variations in average linear attenuation variations. This simple model ignores a number of complicating factors such the anode heel-effect, beam hardening, spatially varying secondary radiation, and variation in the H&D curve over time. The mammogram data were first converted to optical density (OD) scale using the known calibration curve of the digitizer and then to a relative $\log E$ scale using H&D curve calibration data for the Min-R/Microvision film/screen system. Most of the mammogram data fell within the optical density range from 0.3 to 4.0 corresponding to a $\log E$ range of 1.4 units (log base 10). As a practical matter we stored image data in files as 16 bit integers corresponding to $10^4 * \log E$. Finally, potential lesion locations in each mammogram were identified on a square lattice within the constant thickness region with centers 7.7 mm apart. As an added precaution, the centers of potential lesion locations were selected so that no center position was closer than 23 mm from the nearest constant thickness region boundary. This gave 10 750 possible lesion locations for the collection of mammograms. A total of 2048 lesion locations were then randomly selected for the detection experiments.

Previous spectral analysis of mammograms had shown power-law results, $P(f) = B/f^{\beta}$, with exponents near three. We did measurements in 384×384 pixel (46×46 mm) squares in the constant thickness region of 213 images to estimate β. Three methods[43] were used to check for consistency, two were spectral and the other was done in the spatial domain. Power-law spectra are very sharply peaked at low frequency so spectral analysis is challenging[50] and considerable care was needed to reduce systematic error. Two spectral analysis approaches were used: The discrete Fourier transform (DFT) method with a radial Hanning window[51] and the maximum entropy method (MEM).[52] These methods

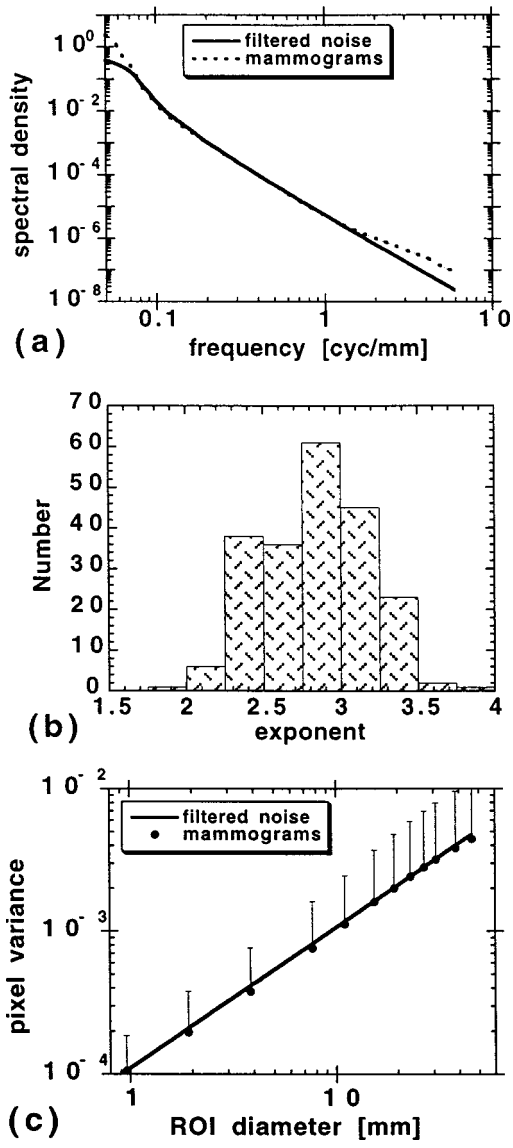

FIG. 3. (a) The estimated power spectra for the collections of mammographic backgrounds and the matching filtered noise. The units of $P(f)$ are $(\log E^* \text{ mm})^2$. The slope of the solid (noise) line is -3.00 for frequencies greater than 0.2 cyc/mm. (b) The distribution of power spectrum exponents (averaged over angle) for 213 individual mammographic backgrounds, with a mean of 2.83 and collection standard deviation of 0.35. (c) Pixel variance as a function of circular measurement region (ROI) diameter for the mammographic backgrounds and the $1/f^3$ filtered noise using in observer experiments. The error bars show the standard deviations of the distributions of measurements for the mammographic backgrounds, not standard error of the estimates of measurement mean values. Distribution standard deviations for the filtered noise measurements were about somewhat smaller.

were also used to analyze 256 filtered ($1/f^3$) noise images produced for observer experiments with simulated backgrounds. The two-dimensional DFT spectra were calculated and then averaged over angle to obtain estimates of the average radial dependence of slices through the spectra. The results are shown in Fig. 3(a). Results are unreliable for frequencies lower than 0.2 cycles/mm due to the window artifact effect—convolution of the Fourier transform of the window function with the ''true'' spectral values that are rapidly changing with frequency. The estimated exponents were 3.0

for the both mammograms and the noise. When the mammogram spectrum was corrected for the film/screen and digitizer MTFs, it became flat above 1 cycle/mm,[43] suggesting dominance by image noise at higher frequencies. The distribution of exponents (radial average of slices) for the periodograms of individual mammogram regions are shown in Fig. 3(b). The mean value is 2.8 with a standard deviation of 0.35. The similar noise periodogram distribution had a mean of 3.00 and standard deviation of 0.03.

Power-law processes have an unusual relationship between spatial domain variance values and measurement region of interest (ROI) size. We used a 2D adaption of a 1D method suggested[53] by Mandelbrot for analysis of stock market data, measuring pixel variance for a nested sequence of circular ROIs with a common center. This is also referred to as the root-mean-square (rms) method in analysis of fractal surfaces.[54] Results are shown in Fig. 3(c) for the mammograms and filtered noise images. One would expect a slope of unity for '$1/f^3$' power-law processes. The regression line fit is 0.99 for both image collections. The variance measurement values for power-law processes do not change in a consistent way as ROI size is increased—results are very different from one image to another. The standard deviations of the distributions of variance measurements for mammographic background images are indicated by the bars in Fig. 3(c). Note that they do not indicate the standard errors of the estimates of distribution mean values. The standard deviations are approximately equal to mean values. The standard deviations for filtered noise variance measurements are smaller. Pixel variance measurement results for white noise are very different. The mean value is independent of ROI diameter and the standard deviation of a collection of measurements is inversely proportional to diameter.

The digital image display system included a Macintosh computer, a DOME model MD2 display board (Dome Imaging Systems, Waltham, MA 02451), and two different monitors. For the first two experiments we used an Image Systems model M24L 24 inch diagonal gray-scale monitor (Image Systems Corporation, Hopkins, MN 55343) with a maximum luminance of 75 cd/m^2 and a 1200(V) by 1600(H) pixel format. For the remaining experiments we used an Clinton model DS2000HB 20 inch diagonal, high-brightness gray-scale monitor (Clinton Electronics, Rockford, IL, 61111) and a 1024(V) by 1280(H) pixel format. The Clinton monitor contrast and brightness were adjusted to give a peak luminance of 210 cd/m^2 and to approximate the Barten tone-scale transformation.[55] Luminance was measured using a Minolta LS-110 telephotometer (Minolta Corp., Ramsey, NJ 07446) with range from 0.01 to 10^6 cd/m^2. We used an inverted display scale so that high x-ray exposure corresponded to a dark image region just as it does on mammograms.

B. Methods

The experiments were done using the 2AFC/SKE technique to allow precise control of the experimental conditions so that human results could be compared to model observer performance predictions for the SKE detection task. Careful experimental design is necessary to try to provide human observers with the same prior information available to the model observers. During each decision trial, two randomly selected background regions were displayed side-by-side with the lesion added to one randomly selected region. Observers selected the region that they believed to contain the lesion. The same lesion was used for each block of 128 decision trials, a high contrast reference copy of the lesion was shown above the two backgrounds and the two alternative lesion locations were identified using circle cues[56] that were large enough that they did not interfere with the decision task. The two background regions included in each 2AFC image trial were randomly selected based on the 2048 possible lesion locations and were always chosen from different mammograms. The search experiment was done using a collection of 213 mammographic background images that were large enough to include a 384×384 pixel square within the constant thickness region. We also produced 1024 white noise and 1024 filtered ($1/f^3$) noise backgrounds.

It was desirable to have a wide range of lesion sizes to facilitate CD diagram slope estimates. We selected 0.5 mm (four pixels) as the minimum to include the size region where mammographic imaging system noise becomes important. Selection of the upper size limit was based on two considerations. One was practical. In previous observer experiments,[57] we have found that lesion detection performance was reduced if lesion size was larger than about 25% of the width of the background. The other consideration was statistical data on tumor sizes detected in mammograms. Reintgen *et al.* reported[58] the size distributions of 215 nonpalpable tumors with cumulative percentages of 18%, 59%, and 85% for 5, 10, and 15 mm sizes, respectively. Zheng *et al.* described[42] the distribution of 220 tumors detected in mammograms, none were smaller than 5 mm and 81% were smaller than 15 mm. Therefore, we selected 15.4 mm (128 pixels) as an upper size limit. During threshold determination experiments, the lesions were minified to smaller array dimensions to obtain reduced nominal lesion sizes. At the smallest size, the minified lesions did not resemble tumors and might be regarded as simulations of compact calcifications. The minified lesions were subjectively realistic for the 4 mm scale and larger. Note that sizes are one-dimensional. Based on the 0.12 mm sampling used for the mammographic backgrounds, example array sizes of 128×128 and 64×64 corresponded to nominal lesion sizes of 15.4 and 7.68 mm, respectively. Ideally, ''equivalent'' lesion sizes would be preferred to nominal sizes but it is not clear how equivalence should be defined for realistic lesions given their highly variable 2D profiles and irregular boundaries. Therefore, we chose to use nominal lesion sizes defined using rescaled array sizes.

The circle cues were bipolar, consisting of contiguous bright and dark rings with diameter dependent on lesion size. An nonlinear relationship relating cue diameter and lesion size was developed, based on subjective evaluation. If the circle was too close to the lesion it interfered with detection. If it was too far away then the lesion location became more difficult to determine. For the smallest lesion size (four pix-

els, 0.5 mm), the selected cue diameter was 32 pixels. The following list gives the other paired lesion and cue diameters in pixels: (8, 46), (16, 64), (32, 90), (64, 128), (96, 156), (128, 182). The contrast of the bipolar rings was also subjectively selected to maximize the benefit of the cue. The same contrast was used for all cue diameters.

Lesion amplitude was adjusted during the trials using a staircase procedure[59] modified to maintain a value close to that giving 90% correct responses. After each incorrect decision the amplitude was increased by one step. After six successive correct responses the amplitude was decreased by one step. The staircase amplitude steps were one-quarter octave (1.091 ratio). The staircase had a floor two steps below the estimated amplitude for 92 percent correct ($d'=2$) based on preliminary results. The goal was to ensure that decision trials were done with lesion amplitudes near those required to give the lowest coefficient of variation due to sampling statistics.[60] As an example of data value conversions, the amplitude threshold for 2AFC detection (defined using d' equal to two) for an 8 mm mass in the average mammographic background is $\sim 0.03 \log E$ units which corresponds to 0.14 OD units at the peak film gamma of 4.5 (which occurs at 2.2 OD).

Each observer did 256 trials for each experimental condition in blocks of 128 trials. Viewing time was not limited and feedback as to correctness of response was used. A block of 128 decision trials took between 6 and 12 min. Observers were encouraged to vary viewing distance to optimize performance for each lesion size. The average viewing distance (750 mm) was about 2300 times the pixel size on the monitor (0.3 mm) but the range varied as much as a factor of 2 either way, as a function of signal size. The 750 mm CRT viewing distance corresponds to viewing the original mammogram at 280 mm distance. In addition, both a magnifying and minifying lens were available for use. Percent correct results were transformed[61] to d' values and amplitude threshold estimates required for d' equal to 2 were obtained by linear regression, assuming proportionality (d' equal zero for zero signal amplitude). It has been repeatedly found that strict proportionality does not hold,[62,63] but since our data were all collected near 90% correct the use of an offset in the regression fit had virtually no effect (less than 1%) on the threshold estimates. For 2AFC experiments, the coefficient of variation of the estimate of d' due to sampling statistics[60] is about 9% for 256 decision trials and d' equal to 2. This sets a lower limit of the estimate of the standard error of the mean of d' for a given observer in a given experimental condition. When the data of two or three observers are combined to obtain an estimate of the average value of d', the minimum coefficients of variation are about 6% and 5%, respectively, since the observers are using randomly selected, independent pairs of images. The coefficient of variation of d' estimates in the search experiment are signal size dependent. For 256 trials and 90% correct, the value is about 5% for the largest signal and 3% for the smallest signal. Since humans have other sources of variability, all these estimates are only lower limits.

C. CD diagram with extracted masses and mammographic backgrounds

This experiment was designed to determine the CD diagram for detection of realistic mass lesions in mammographic backgrounds. Four lesions were used to provide variety and improve the statistical significance of the CD diagram slope estimates. Mammogram structure is highly variable, both within and between images, and the dynamic range of CRT display is limited. We wanted to be sure that human performance was not limited by image amplitude so we used an adaptive display contrast (window) method. Separate linear transformations were applied to the image data, $d_i(x,y)$, for the two background regions ($i=1,2$) used in a 2AFC decision trial. The mean, m_i and standard deviation, σ_i, in each image was measured within a 23 mm (192 pixel) diameter circular region centered on the potential lesion location. The transformation, carried out using floating point numbers, is given by

$$v_i(x,y)=g_i[d_i(x,y)-m_i] \quad \text{where} \quad g_i=f(\sigma_i). \tag{13}$$

The gain factor, $g_i=f(\sigma_i)$, that we used was a nonlinear function developed by subjective evaluation. The goal was to obtain maximum display contrast for a given background image while avoiding clipping at either end of the image data range. The widest and narrowest data ranges allowed by g_i were equal to 0.66 and $0.2 \log E$ units, respectively. Given a mean background image optical density of 1.3, these ranges in OD units correspond to ($0.6 \rightarrow 2.5$) and ($1.0 \rightarrow 1.6$) respectively (on average). For this adaptive method, the lesion was also scaled by the gain factor used for the background to which it was added. Finally the image data values were rescaled to the display memory data range ($0 \rightarrow 255$) with a mean of 128 and converted to byte form.

D. CD diagram with one extracted mass and white noise backgrounds

We did a control experiment using one realistic mass lesion (*B*, in Fig. 2) to determine the CD diagram with uncorrelated (white) image noise. The noise had a range of 256 gray levels, a mean of 128 and a pixel standard deviation of 25.6 gray levels. The images were generated using a random number generator.[52] All three observers took part in this experiment and all methods were identical with those used in the mammographic background experiments including the use of the Image Systems monitor.

E. Effect of display window width (image contrast)

CD diagrams for detection of mass *B* were determined using four different $\log E$ window width methods using the high brightness Clinton monitor. The goal was to determine whether results depended on the gain factor used to convert from $\log E$ data values to the 256 gray level data scale used to drive the CRT. The highest contrast strategy used the adaptive window width (gain factor) method described above. Three fixed window widths were also used, covering $\log E$ ranges of 0.66, 1.0, and 1.4, respectively. The last window width was wide enough to display the full dynamic

range of 95% of the mammograms in the data base. The central display level was always determined using the background mean value method described above and displayed background levels were shifted to have a common mean. Two observers took part in this experiment. The lesion size range for this and subsequent experiments was reduced by eliminating the 0.5 mm size.

F. CD diagrams with a simulated nodule and filtered ($1/f^3$) noise backgrounds

These experiments were done using a simulated (designer) nodule signal with the two collections of background images (mammographic and filtered noise) as well as mass *B* in filtered noise backgrounds. The mass *B* lesion was included to determine whether its detectability (by humans) in the filtered noise background was similar to that of the simulated nodule. The purpose of the combined nodule and filtered noise background condition was to allow precise comparison between human and model observer performance. This could be done because (*A*) the nodule signal was smooth and isotropic so radial basis functions could be used and (*B*) the filtered noise background statistics were known. The designer nodule profile with v equal to 1.5 was used in this work and is shown in Fig. 2(b). The filtered noise backgrounds were lowpass power-law noise images produced by frequency domain filtering of white images obtained using a Gaussian random number generator.[52] The collection had the same estimated second order statistics as the collection of mammogram regions, that is a power spectrum $P(f) = B/f^3$ in the frequency range above 1 cycle/mm. Care is needed in production because of the problem of ''data wrap-around'' due to the periodic nature of the DFT when doing the inverse transform from the frequency domain. The collection was created by starting with 2048×2048 pixel white noise images, transforming to the frequency domain, filtering with $H(f) = B^{0.5}/f^{1.5}$, doing the inverse Fourier transform and selecting the central 512×512 region. The spectrum of the collection was then measured using the DFT method with results (averaged over angle) shown in Fig. 3(a). As a final check, the statistics of the mammographic collection and the filtered noise collection were compared in the spatial domain using the sequential pixel variance measurement method. The filtered noise image amplitudes were then rescaled by a small amount so that pixel variances of the two collections were identical for a diameter of 32 pixels, the geometric mean of nodule signal size. These experiments were done with the medium display window width.

G. CD diagrams for search in both types of backgrounds

In these experiments, the signal was always present at a randomly selected position somewhere within a 46×46 mm (384×384 pixel) region of a 61×61 mm background. The simulated nodule signal and one extracted mass (*B*) were used with both filtered noise and mammographic backgrounds. The same signal was used throughout each block of 128 decision trials and a reference copy was shown above the image to be searched. The observer's task was to try to identify the signal location using a cursor and trackball. There was no time limit and signal amplitude was adjusted using the staircase procedure. The observer response was scored as correct if the cursor was placed anywhere within the boundaries of the square centered on the signal. For the smallest (1 mm) signal, the square width was 1.5 times the signal size because of the problem of precise cursor positioning (± 4 pixels) if the actual square size had been used. This extra size tolerance would have little effect on results. During the search task, the observer might occasionally have two candidate locations in mind, but they would be well separated in the image. The observer was given feedback as to location correctness. If the selected location was incorrect, a bipolar circle (as used in the 2AFC experiments) would be placed in the image, centered on the true signal location. The observer was given an unlimited time to view this ''corrected'' image before continuing with the experiment. The results were recorded as percentages of correct responses, P_c, as a function of signal amplitude and thresholds were defined as the value required for 90% correct performance. It was necessary to estimate a best fit psychometric function through the P_c data to estimate this threshold. The standard signal detection theory MAFC method,[60,61] was used. MAFC analysis requires statistical independence of alternative decisions, which is violated in this experiment in two ways. There was no constraint on signal location, so alternative positions could overlap. Even if overlap was prevented, the alternative decisions would not be independent because of the long range correlations in the background structure. As an approximation, we defined an effective number of locations, M^*, using the ratio of search and signal areas. The value of M^* depends on signal size, ranging from 2900 for the smallest size to 11 for the largest. Given M^*, the value of $d'(M^*)$ was estimated for each amplitude, then linear regression was used, as in the 2AFC case, to estimate $d'(M^*)$ for 90% correct and the corresponding threshold amplitude was calculated. The value of $d'(M)$ is a weak function[61] of M (approximately logarithmic) so the precise details of defining M^* have little effect. In spite of the violation of the independence assumption, the M^* approach is a useful curve fitting method and also, as will be seen in the results section, gave a surprising good description of human search performance relative to 2AFC detection performance.

H. Model observer performance evaluation

Ideal and NPWE observer performances for SKE detection of the simulated nodule in filtered noise were determined by numerical solution of Eqs. (2) and (4). Ideal observer performance cannot be calculated in this way for the mammographic backgrounds because of the variation in second order statistics. However, we did use Eq. (2) to estimate performance of a PW model observer. Two different estimates of the mammographic power spectrum exponent were used: The radial average value of 3.0 for the collection and the mean value, 2.83, for the distribution of individual image radial averages. NPWE model performance was also esti-

mated using the template application method described below. NPWE results presented here are for a fixed viewing distance of 750 mm from the monitor. NPWE model nodule detection performance for the filtered noise case was also calculated numerically with variable (signal size dependent) viewing distances. There was no significant difference in results for nodule sizes greater than 2 mm.

NPWE model performance evaluation using the template approach proceeded as follows. First the Fourier transform, $S(u,v)$, of the discrete signal was calculated and multiplied by the square of the eye filter to give the template filter, $W(u,v)=E^2(u,v)S(u,v)$. The eye filter appears twice because the external signal is viewed through the eye and then the filtered signal, $E(u,v)S(u,v)$, is referred back to the image plane. Using the inverse Fourier transform, we obtained the observer template, $w(x,y)$, which can also be described by the vector, $\mathbf{w}$. The decision variable, λ, used by the observer to decide whether a signal is present in an image is determined by the cross-correlation (inner product) operation, $\mathbf{w}^t\mathbf{g}$. Since the signal is additive, the difference in the mean values of the decision variable with, $\mu_{\lambda,s}$, and without the signal, $\mu_{\lambda,n}$, will be given by

$$\Delta\mu_\lambda=\mu_{\lambda,s}-\mu_{\lambda,n}=\langle\mathbf{w}^t(\mathbf{s}+\mathbf{n}+\mathbf{b})\rangle-\langle\mathbf{w}^t(\mathbf{n}+\mathbf{b})\rangle=\mathbf{w}^t\mathbf{s}, \tag{14}$$

where $\langle\cdots\rangle$ indicates the expectation value. Similarly, the presence of the signal will not influence the standard deviation, σ_λ, of the decision variable. The detectability index is given by

$$d'=\frac{\Delta\mu_\lambda}{\sigma_\lambda}. \tag{15}$$

The value of $\Delta\mu_\lambda$ was obtained by cross correlating the template with the signal. The value of σ_λ was estimated by cross-correlating the template with a large number of potential signal locations in the image collections. Equation (15) assumes of a Gaussian probability distribution function (PDF) for λ. Serious errors in d' estimates[64] can arise if the PDF is markedly non-Gaussian. This was checked for both background types using 2000 locations and the NPWE model template response PDFs were well described by Gaussian distributions. Results for the numerical integration and template approaches agreed within 3% for $1/f^3$ filtered noise backgrounds. The NPWE template application results will be used for comparison with human results because they do not depend on spectral estimates.

Nodule detection performance for the channelized FH model (FH_{LG}) with Laguerre–Gauss basis functions[34] was determined in the spatial domain using the vector-matrix calculation approach, Eq. (6). The FH model vectors and the covariance matrices were determined using the following procedures.[23] First the model observer basis matrices, $\mathbf{T}$, were calculated using six basis functions; which include a distance scale, a, related to the signal radius, R, through a scaling ratio, $\rho=a/R$. Therefore it was necessary to determine the optimum value of ρ. This was discussed below, so let us assume ρ equal to unity for the moment. Once $\mathbf{T}$ was determined, the covariance matrix, $\mathbf{K_C}$, for the collection of background images as ''seen'' through the channel basis functions was estimated by first cross-correlating each basis vector with the ith image to obtain the result $\mathbf{r_{gi}}=\mathbf{T^t g_i}$, and then using the following equation.

$$\mathbf{K_C}=\langle(\mathbf{r_{g,i}}-\langle\mathbf{r_{g,i}}\rangle)(\mathbf{r_{g,i}}-\langle\mathbf{r_{g,i}}\rangle)^\mathbf{t}\rangle. \tag{16}$$

The 213 independent mammographic backgrounds that had been used for spectral analysis were also used for covariance matrix estimations. To maintain similar statistical uncertainty, 256 backgrounds were used from the filtered noise collection. This gave a coefficient of variation of about 10% for the individual matrix elements. The vector, $\mathbf{r_s}$, describing the signal as seen through the basis functions was determined by cross correlating each basis vector with the signal to give $\mathbf{r_s}=\mathbf{T^t s}$. Finally Eq. (6) was used, with φ and $\mathbf{K_{int}}$ equal to zero, to determine d' for each signal size. The optimum value of ρ was determined by iteratively adjusting it to maximize d' for each nodule size. The best value was between 0.70 and 0.75 for the size range from 2 to 15 mm. For sizes below 2 mm, the best value changed rapidly and equaled 0.90 at 1 mm (eight pixel) nodule diameter. This may be due to fact that the discrete nature of the basis functions became more important as nodule size decreased.

IV. RESULTS

A. Initial CD diagram experiments

The CD diagram results for the first experiment are shown in Fig. 4. These CD diagrams and all others will show the logarithm of amplitude thresholds as a function of the logarithm of nominal lesion sizes in millimeters. Thresholds are defined as the lesion amplitude (in $\log E$ units) required to obtain d' equal to two. The mammographic background results presented in Fig. 4(a) are for the four masses (A–D) averaged over data for the three observers. The CD diagram results for sizes of one mm and larger show increasing thresholds as predicted by theory. Regression fits to the data gave slopes of 0.25, 0.32, 0.31, and 0.28 for masses A to D in that order. The increased threshold for the smallest size (0.5 mm) is believed to be due to the fact that the power spectrum of the mammograms[43] is dominated by image noise for frequencies higher than 1 cycle/mm. Figure 4(b) shows mammographic backgrounds threshold results for the individual observers averaged over the four masses. As can be seen there was essentially no difference in performance between observers. The results for the control experiment, detecting mass B in white noise, are shown in Fig. 4(c). This CD diagram shows the familiar decrease in amplitude threshold as lesion size increases and a flattening of the curve at large sizes. The amplitudes are given relative to the image noise standard deviation per pixel. Given the white noise standard deviation of 20% of the mean (25.6 gray levels out of 128), the flattening occurs at a peak lesion amplitude of 5 gray levels on the monitor. The dashed line (slope $=-0.81$) is a regression fit to the average data for the four smallest sizes. The solid line is a smooth polynomial fit to average data for

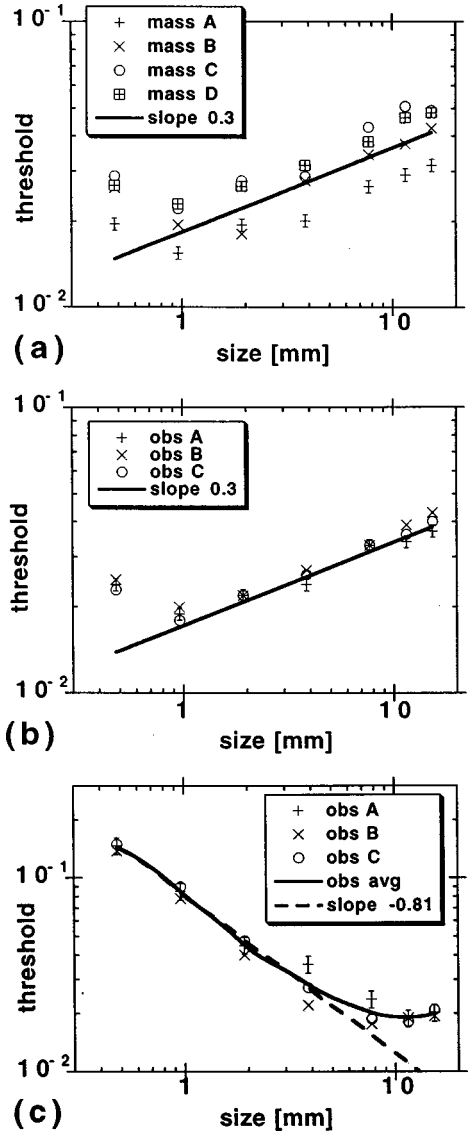

FIG. 4. (a) Threshold results (in log-exposure units) for detection of the four extracted masses (*A*-*D*) added to digital mammographic backgrounds as a function of scaled (nominal) mass size (in millimeters). The results are averages for three experienced observers (two physicists and a radiologist) with 256 trials for each observer per datum. The solid line has a slope of 0.30. The estimated coefficient of variation due to sampling statistics is 5%. Error bars are shown for mass *A*. The term ''nominal size'' is defined in the text. (b) Results for the same experiment shown by observer, with their results averaged for the four masses. (c) Threshold results (in relative amplitude units) for detection of mass *B* in white noise backgrounds (pixel standard deviation of 25.6 gray levels) as a function of lesion size. The dashed line has a slope of -0.81.

all sizes and is a guide to the eye with no theoretical significance. This white noise result is consistent with the many previous demonstrations in the literature.

B. Effect of display window

The next experiment was done using one lesion (mass *B*) and mammographic backgrounds to investigate the effect of display window width (gain factor). The four gain conditions gave a variation of approximately three in image contrast on the monitor. The results are shown in Fig. 5. The solid line has a slope of 0.30. The results demonstrate very little window width dependence. If lesion detectability was completely determined by patient structure then one would not expect any variation at all. The only systematic differences are at the smallest and largest lesion sizes. The thresholds for the one mm size at medium and wide windows are elevated by about 20% compared to the other two window conditions. The threshold for the 15 mm size and widest window is also elevated by about 20% compared to the other three window conditions. This clearly demonstrates that lesion detection is limited almost completely by breast structure complexity rather than displayed image contrast, *as long as* the lesion is near the middle of the gray scale range. Note that this does not suggest that display contrast is not important for clinical mammograms since our results do not say anything about what happens to detection thresholds when lesions are in a region near the darkest or brightest part of the gray scale range. The variable gain experiment was done with both monitors and the ratio of average thresholds was 1.06. The thresholds for the higher luminance monitor were lower but this is not regarded significant, since the ratio falls just within the 95% confidence limits based on sampling statistics.

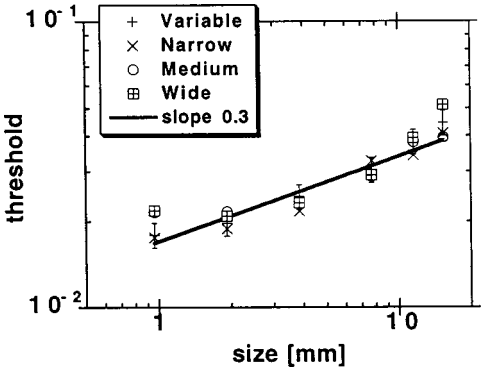

FIG. 5. CD diagram results for mass *B* (averaged for two observers) for four different display window widths. The windows were variable (*V*), narrow (*N*, range $=0.661 \log E$ units), medium (*M*, $1.0 \log E$ units), and wide (*W*, $1.4 \log E$ units). The variable window method is described in the text. The wide window could encompass the entire dynamic range of a mammogram (OD range 0.3–4). Error bars are shown for the variable window case. The solid line has slope 0.30.

C. CD diagrams for two background types

The amplitude threshold results for detection of the simulated nodule and mass *B* in mammographic backgrounds are shown in Fig. 6(a). There were no systematic differences in performance between observers or in detectability of the two signals. The ratio of average results for mass *B* and the nodule for the two observers who took part in both experiments was 1.00 ± 0.05. The solid line is a regression fit to the average data for sizes of two mm and larger. There is a significant departure from this line for the one mm size, just as was seen for the medium window width in the previous experiment. The threshold results for detection in $1/f^3$ filtered noise backgrounds are shown in Fig. 6(b). Once again, there were no systematic differences in performance between the three observers or in detectability of the two signals. The ratio of

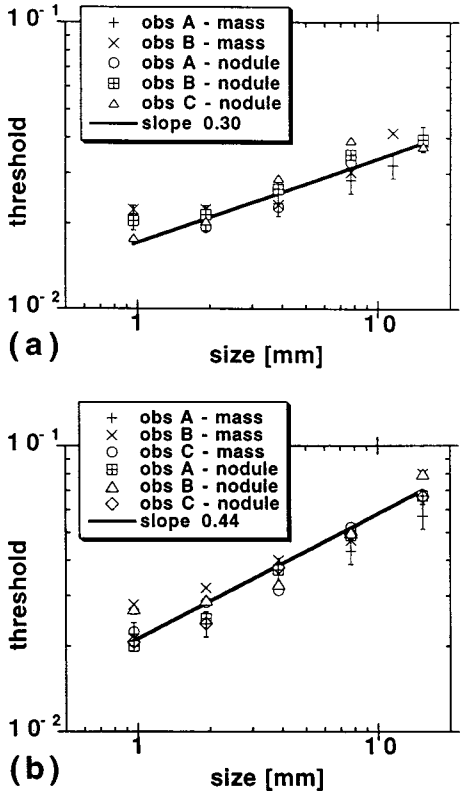

FIG. 6. (a) CD diagram results for mass *B* and the simulated nodule in mammographic backgrounds using the medium window width. The solid line has a slope of 0.30. (b) CD diagram results for mass *B* and the simulated nodule in filtered $1/f^3$, noise with the same power spectrum as the mammographic backgrounds. The solid line has a slope of 0.44.

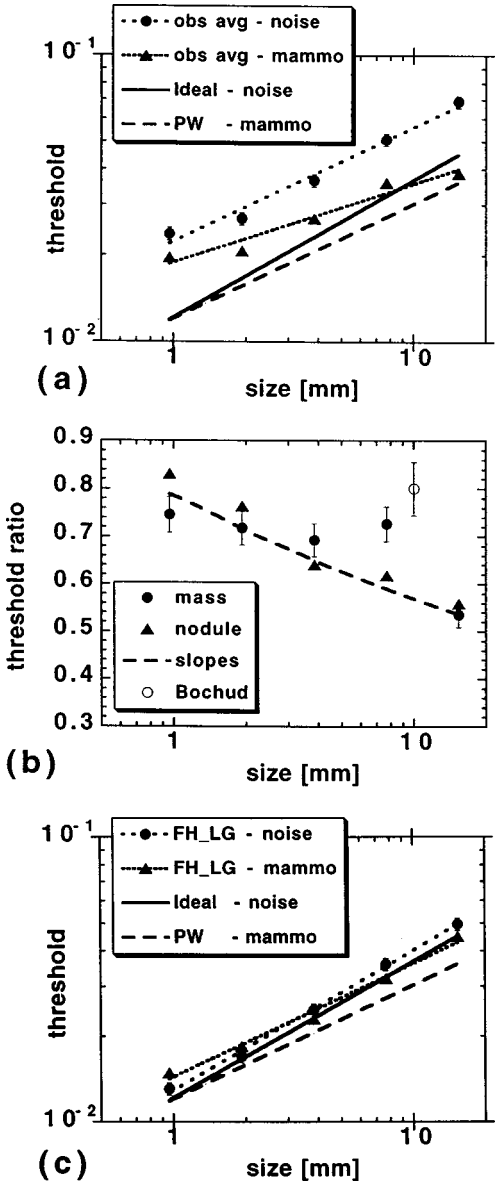

FIG. 7. (a) Average human results for nodule detection and model observer results calculated by numerical integration of Eq. (2). The model result for filtered noise corresponds to ideal observer performance since the noise is known to be stationary. The PW model calculations for the mammogram case were done using the spectral distribution mean exponent of 2.83. (b) The ratio of amplitude thresholds (averaged over observers) for the mass and nodule signals in the two background types are shown as a function of signal size. The solid line is the expected relationship given the different CD diagram slopes for the two background types. An additional point (*B*) is shown from Bochud *et al.* (Ref. 8). (c) Thresholds for nodule detection by the spatial domain FH model with Languerre–Gauss basis functions (LG) in the two background types compared with the frequency domain numerical integration results from (a).

amplitude thresholds for the two signals was 1.01 ± 0.03, averaged over the size range. The difference for the filtered noise background results was that the slope of the regression fit to the data changed to 0.44, in much closer agreement to the theoretical prediction of 0.50 for $1/f^3$ noise.

The average human results for detection of the simulated nodule in the two background types are combined in Fig. 7(a), with regression lines as guides to the eye. The main point of this figure is that there is a very significant difference in the absolute values of the human threshold levels. The nodule is easier to detect in the mammographic backgrounds. The figure also includes the numerical integration results for the PW model for mammographic backgrounds with the histogram-based power-law coefficients (slope of 2.83) and the ideal observer for filtered noise. Human efficiencies vary as a function of lesion size. The average absolute efficiency for detection in the filtered noise is ~40% except for the 1 mm size with 24% efficiency. The average relative efficiency for the mammographic background case is 57% with a range from 38% to 91%. Note that ideal observer performance and absolute efficiency cannot be calculated for the mammographic background case because of the uncertainty of its statistical properties. To further illustrate the human observer differences, the ratios of amplitude thresholds for the two backgrounds are shown in Fig. 7(b). There is a systematic and approximately linear decrease in ratio as a function of log (signal size). The line through the data was calculated using the ratio of CD diagram slopes followed by scaling to fit the average threshold ratio. In addition a point from a similar experiment by Bochud *et al.*[8] is shown in the figure.

The CD diagrams for nodule detection by the FH_{LG} model in the two types of backgrounds are shown in Fig. 7(c), together with the ideal observer results for the filtered noise case and PW model results for mammographic backgrounds

(using β equal to 2.83). The purpose was to compare frequency and spatial domain predictions as well as predictions for the two types of backgrounds. There is fairly good agreement between the frequency domain results (ideal and PW) and the spatial domain results (FH_{LG} model). This suggests that statistical nonstationarity of mammographic backgrounds is not a major concern. There is good agreement between the FH_{LG} model thresholds for the two backgrounds with a maximum disparity of 10% at the ends of the size range. This indicates that the statistics of the two background cases were well matched. However, there is a significant difference in slopes of the regression lines, 0.4 and 0.5 for the mammographic and filtered noise backgrounds, respectively. The regression lines cross at the middle signal size (4 mm). It should be noted that this corresponds to the region size (32 pixels) used for the final matching of the background fluctuation amplitudes using the sequential pixel variance method.

D. Comparison of human results and model predictions

Figures 8(a) and 8(b) show comparisons of human nodule and mass threshold results (averaged over observers) together with nodule threshold predictions for two model observers scaled using observer internal noise. Our variable display window experiments had shown little effect of image display contrast on human performance. This suggested that the static internal noise contribution was unimportant, so it was not included in further model analysis. For comparison of human and model performance the induced internal noise parameter, φ, was adjusted for each model and each experiment to give a match to average human results at a signal size of 4 mm (32 pixels). The results for the mammographic background case are shown in Fig. 8(a). The FH_{LG} model scaled result was obtained using φ equal to 0.1. The NPWE model was scaled using φ equal to -0.5. This negative value of φ arises because human performance was better than the model and can be interpreted as suggesting that the effective image noise level for humans doing the signal detection task is less than the noise level estimated by the NPWE observer. The negative value also indicates that NPWE model is not a realistic one and that it must be used with care. The results for the $1/f^3$ noise background case are shown in Fig. 8(b). The values of φ used to scale the model predictions to fit the human data were 1.08 and 0.30 for the FH_{LG} and NPWE models, respectively.

E. Search experiment results

The results for the search experiment using the mass signal are shown in Fig. 9. The results for the two observers agreed within experimental accuracy, so they were averaged. The average human results for the 2AFC experiments with the two types of backgrounds are also shown for comparison. The search results thresholds for the mammographic and filtered noise backgrounds are in very good agreement, in contrast to the 2AFC results. This will be discussed below. The solid curve through the search data was calculated as follows. Given model observer results for detection with two alternative signal locations (2AFC), performance for M statistically independent locations (MAFC) can be calculated.[24] It has been shown[62] that this is true for humans detecting simple signals in uncorrelated noise for values of M up to 450. The statistical independence requirement is important and was violated in this experiment. There was no constraint on lesion location in this experiment, so alternative positions could overlap. Even if overlap was prevented, the alternative decisions would not be independent because of the long range correlations in the background structure. However, as an approximation, an effective number of locations, M^*, was defined as the ratio of search and lesion areas. The value of M^* depended on lesion size, ranging from 15 for the largest size to 3550 for the smallest. Given M^*, the standard signal detection theory MAFC calculation method[24] was used to calculate human threshold estimates for the search tasks based on the human 2AFC results for the filtered noise background. The solid curve is a polynomial fit to these estimated thresholds with no additional scaling, which indicates that humans have the same efficiency for the M^* search task as for the 2AFC detection task with the filtered noise backgrounds. The implications of this result will be discussed below.

V. DISCUSSION

Mammography is the primary method of detecting breast cancer. However, lesions are often extremely difficult to detect because of the complex and highly variable normal parenchymal patterns. Our work was designed to develop a quantitative understanding the effect of this structure on detectability of realistic lesions. Most previous experimental CD diagram measurements have given similar results, observers required less contrast for larger signals to achieve a particular detection accuracy. We obtained the same result with our control experiment using one realistic lesion and white noise. Our results for lesion detection in mammographic backgrounds are completely different. Amplitude thresholds increased as lesion size increased for lesions larger than 1 mm. In this region, the CD diagram for 2AFC experiments had a positive slope of 0.3. This result was found to be robust under a range of experimental conditions. The threshold increase below 1 mm (the size range for microcalcifications) is assumed to be due to image noise dominance in the mammographic spectrum at spatial frequencies greater than 1 cycle/mm.

This effect of a positive CD diagram slope is not completely novel. It is a common radiological experience that some large lesions with unsharp boundaries (edge gradients) can go undetected if the observer is too close to the image. This implies that the CD diagram slope, for these examples, has a positive value above some particular lesion size, expressed in *visual angle* rather than image distance units. This experience is consistent with the literature on detection of simulated signals with unsharp edges. The effect is most marked for detection of signals without image noise, which suggests that amplitude thresholds for these signals are in-

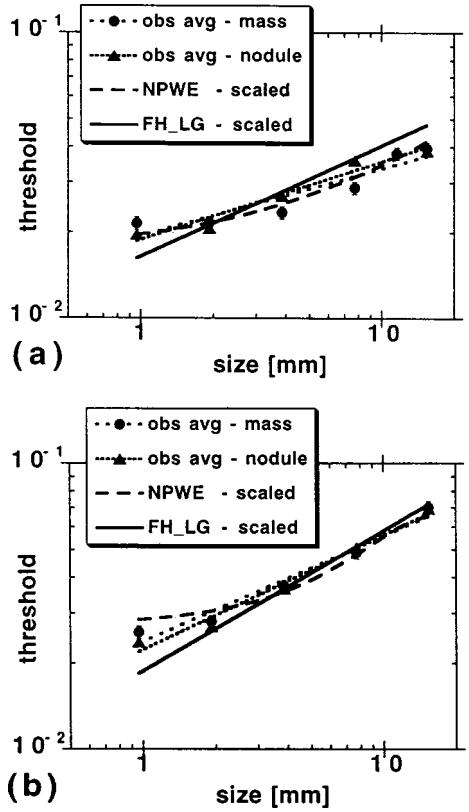

FIG. 8. (a) Comparison of average human results for detection of the mass and nodule signals in mammographic backgrounds with NPWE and FH_{LG} model observer thresholds, scaled using induced internal noise to fit human results for the 4 mm size. The solid line (slope=0.40) is a regression fit to the FH_{LG} model data. (b) Comparison of average human results for detection of the mass and nodule signals in filtered noise backgrounds. Induced internal noise was used for scaling models to fit human results for the 4 mm size. The solid line (slope=0.50) is a regression fit to the scaled FH_{LG} model data.

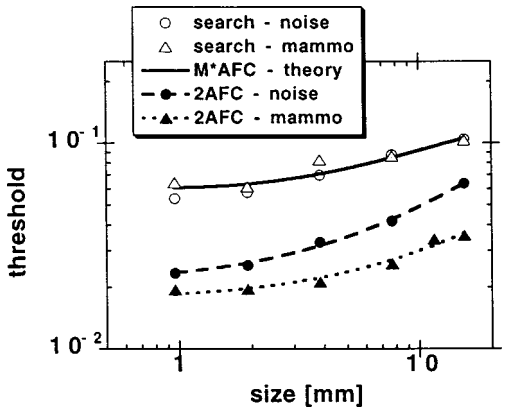

FIG. 9. Threshold results (averaged over human observers) for the search experiments for the mass signal in the two types of backgrounds. The 2AFC results are shown for reference with smooth dashed curves as guides to the eye. The solid curve is a fit to the search data using an M^*AFC model described in the text.

versely related to retinal luminance gradients. Examples include[65–67] detection of both bar and disc-like signals with unsharp edges (Gaussian and half-sine profiles) in noiseless images. These measurements demonstrated a minimum in the CD diagram for bar signal widths of ~0.05 degrees (1 mm viewed at 1 meter) and disc-like signal diameters of about 0.5 degrees (10 mm viewed at 1 meter). Burgess *et al.*[67] did threshold measurements for bar and disc signals with unsharp edges in radiographic image noise at a fixed viewing distance. Threshold elevation at large signal sizes was modest, but still present. Observers in our experiments were not constrained to a particular viewing distance. In addition, both magnifying and magnifying lenses were provided. Additional evidence that our unusual mammographic CD diagram results is not due to edge gradient effects is provided by the white noise control experiment, which showed no threshold increases at large lesion sizes. We conclude that the novel CD diagram threshold slope in our experiments is due to the unusual nature of the background structure spatial statistics. Chakraborty and Kundel found quite different CD results[68] in investigations done using sharp-edged disc signals smoothed by a Gaussian filter with $1/f^3$ noise. They kept the standard deviation of the Gaussian filter fixed as disc diameter increased. The predicted CD diagram had a negative slope of −0.5. This can be understood by the following considerations. First, the (radius normalized) form of the filtered signal violated the assumption of fixed profile used in the development of Eqs. (8)–(12). Secondly, the bandwidth of the smoothed signal was fixed because to the fixed smoothed filter. We evaluated smoothed disc detectability in $1/f^3$ noise using a Gaussian smoothing filter with standard deviation proportional to disc radius. This ensured that smoothed signal profile did not change as size increased. The CD diagram for this case had a positive slope of 0.5, consistent with the prediction of Eq. (12).

A crude intuitive explanation for the positive CD diagram slope with fixed signal profile and power-law processes is as follows. In white noise, pixel variance is independent of measurement region size. Signal strength increases in proportion to its area while the total noise standard deviation within the signal area increases in proportion to the square root of area. So SNR at constant signal amplitude increases in proportion to signal diameter and amplitude thresholds decrease as signal size increases for a constant SNR (or d' value). A similar effect is seen for lowpass image noise, which has a well defined correlation distance. Pixel variance becomes independent of measurement region size for diameters large compared this distance. Power-law noise does not have a defined correlation distance[69] and has increasing variability as the measurement scale increases [Fig. 3(c)]. For power-law exponents greater than two, SNR at constant amplitude actually decreases with increasing size. So amplitude thresholds increase as signal size increases at constant SNR. The above explanation is based on a nonprewhitening assumption and should not be used for calculations.

The CD diagram results for the extracted masses and the simulated nodule were similar under all experimental conditions, suggesting that details of our 2D signal profiles were not an important factor. The results for the display window width experiment demonstrated that neither the CD diagram slope or amplitude threshold levels were markedly dependent

on displayed image contrast. However, one must be careful not to assume that this applies to the clinical lesion detection situation. Our experiments were done with the mean levels at the potential signal locations in the mammographic backgrounds shifted to fall in the middle of the CRT luminance range. This condition is obviously does not apply in the display of real mammograms, where the local backgrounds of lesions can fall anywhere within the dynamic range of the image.

The model and human observer CD diagrams for the 2AFC experiments had positive slopes, as predicted by the ideal observer model. This qualitative agreement was encouraging, but there was quantitative disagreement. The estimated power spectra of the filtered noise and mammograms gave $1/f^{\beta}$, power-law exponents of 3.0. Equation (12) indicates that the CD slopes should have been equal to 0.50. The FH_{LG} model threshold slope for filtered noise agreed with this prediction, whereas the slope was 0.4 for the mammographic backgrounds. The human CD slopes were 0.44 for filtered noise backgrounds and about 0.3 for mammographic backgrounds. These slope differences demonstrate that the two background types were not equivalent, in spite of our effort to match second-order statistics using power spectrum and sequential pixel variance measurements. There are several possible reasons for the slope differences. One is that spectral estimation for the breast structure may be unreliable. This would not be surprising since it was based on the assumption of stationarity. Using the relationship, $m=0.5(\beta-2)$ and β equal to 3: A variation of 0.1 (3%) in power-law exponent (β) estimate will lead to a 0.045 (9%) change in CD diagram slope(m) estimate. A second possible explanation is the difference in statistical homogeneity in the two collections of backgrounds. The distribution of radial average exponents for individual periodograms in the images in the noise collection was very narrow, with a mean of 2.99 and a standard deviation of 0.03. In order to make our results as general as possible, we made no effort to select one particular type of breast structure (e.g., fatty or dense) so there was a wide variety of background complexity. The distribution of radial average exponents for the mammographic images was broad, with a mean of 2.83 and a standard deviation of 0.35. Our detection results, for both humans and models, are based on averages over this distribution.

The other aspect of the 2AFC results that must be considered is that there was a very significant difference between threshold values for human observers with the two backgrounds, but only a small difference for model observers. This is related to the question of whether mammographic structure can be considered to be purely random noise or whether should it be treated as a combination of noise and deterministic (masking) components.[4,7] Our human 2AFC results are consistent with previous findings by Bochud *et al.*[7,8] The lower thresholds for mammographic backgrounds suggest that, for humans in 2AFC experiments, the effect of mammographic structure is a mixture of noise and masking. An equivalent statement is that we are able to use anatomical knowledge to some degree in the decision process. This is consistent with subjective impressions. At a small scale, we can identify a variety of structure in individual images (ducts, skin pores, etc.). At large scales we can identify separate regions of fatty and dense tissue with markedly different lesion detection amplitude thresholds. The human mammogram/noise threshold ratio plot, Fig. 7(b), suggests that the effect of breast structure on lesion detection becomes less ''noise-like'' as lesion size increases. The threshold results for the FH_{LG} observer model for the two types of backgrounds are quite similar in spite of the slope differences, with equality at the 4 mm size. This suggests that for the model, which has no knowledge of anatomy, breast structure is equivalent to noise to a good approximation.

Absolute human efficiency for 2AFC simulated nodule detection in filtered noise backgrounds was about 40%. This value is similar to a variety of previous results for detection, discrimination and localization of simple aperiodic signals in white noise [29,62] and lowpass filtered noise.[20,39] Absolute efficiency for nodule detection in filtered noise was 90% for FH_{LG} model. Average human efficiency for 2AFC nodule detection in the mammographic backgrounds was 57% relative to the PW observer model, but was as high as 90% (for the largest signal size). The scaled NPWE results agreed well with human results, however humans were more efficient than the model for the mammographic backgrounds so this agreement should not be taken as anything but a modeling convenience. The NPWE model, however, has the useful feature that viewing distance can be adjusted to give a good fit to human data at small signal sizes.

The human results for the search experiment were surprising, in that the thresholds were the same for the two background classes. This suggests that, for this task, breast structure is equivalent to random noise and that anatomical knowledge does not help. It was also surprising that the simple M^*AFC model could be used to estimate search threshold amplitudes directly from 2AFC results with filtered noise backgrounds, in spite of the fact that the assumption of statistical independence of alternative locations was violated. A second surprise is that no additional scaling of human results was necessary. This suggests that our efficiency for the search task is the same as for the 2AFC task—consistent with a similar finding for signal location identification in white noise.[62]

Variation in breast structure statistical properties, both between patients and within a mammogram, has interesting implications for lesion detectability. Equation (12) gives a linear relationship, $m=0.5(\beta-2)$, between CD diagram slope, m, and spectral exponent β for the ideal observer—subject to the constraints discussed after Eq. (12). Let us assume that the same linear relationship holds for humans with a coefficient of 0.3 for realistic mammographic lesions and backgrounds and consider the CD diagram threshold lines in Fig. 10 for exponents 2, 3, and 4. The threshold line represents a boundary between detecting and not detecting a lesion at a selected decision accuracy (area under the ROC or percent correct in 2AFC detection). The amplitude (contrast) of a growing tumor can also be represented on the CD diagram plot. As the tumor grows, its projected contrast will be determined by the difference in x-ray attenuation coefficient

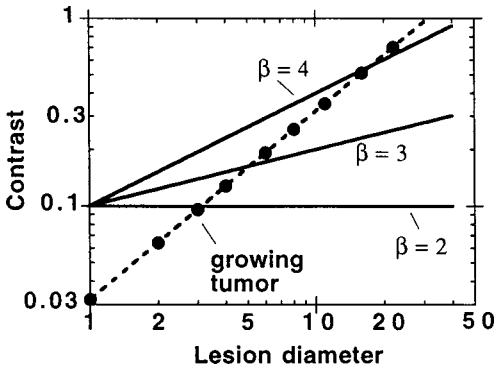

FIG. 10. The contrast trajectory of a growing mass (●, with slope +1) is superimposed on the PW observer model CD diagram threshold boundary lines for three power-law noise exponents (2, 3, and 4). This covers the range found for small (10×10 mm) regions on mammograms. For a given exponent, the tumor would be detectable above the corresponding boundary.

between the tumor and the surrounding tissue and by the tumor thickness perpendicular to the x-ray beam. We assume thickness proportional to diameter, so contrast will increase linearly with diameter as long as the composition of the tumor and surrounding tissue do not change. The contrast trajectory for such a growing tumor is shown in Fig. 10. The tumor becomes detectable, at the selected accuracy criterion, when its trajectory crosses the appropriate CD threshold line. If the tumor is in a region of low spectral exponent, detection probability will change rapidly with size. If the exponent is large, detection probability will change slowly with size. If the spectra, $P(f)=B/f^{\beta}$, have similar values of B, then the lesion will be detectable at smaller sizes for mammogram regions with smaller power-law exponents.

The search for breast cancer in projection mammograms is one of the most difficult tasks undertaken by radiologists. This is most frequently attributed to the superimposition of breast structures and the lack of anatomical assistance in sorting out the complex range of parenchymal patterns. This point of view suggests that the patterns are essentially a form of random noise. Our search experiment results are consistent with this assumption. It is clear that the statistical properties of the parenchymal patterns and the resulting positive slope of the CD diagram are unusual. Our finding that more lesion contrast (amplitude) is needed as the size of the lesion increases helps explain why large lesions can be missed despite careful search of a mammogram. Van Gils *et al.*[70] reported that 19% (40/211) of cancers missed in a large screening study were larger than 22 mm. The theoretical prediction that lesion detectability will be dependent on the local statistical properties of patient structure is also very germane to this point. It is well known that tumors are more difficult to detect in dense breast tissue.[70,71] This is usually attributed to a reduction in linear attenuation coefficient difference between the lesion and surrounding tissue. We have done preliminary investigations indicating that power-law spectrum exponents increase as breast tissue density increases. This suggests that a second factor may need to be considered in explaining the dense breast effect. There was encouragingly good agreement between our human observer results and the model predictions when internal noise is included. We do recognize that our investigations must be extended to even more clinically realistic decision tasks. Examples include discrimination between benign and malignant lesions, identification of lesions from a menu of possibilities and search with uncertainty as to which of a number of lesions is present.

Our investigation was limited to digitally rescaled lesions to satisfy the constraints of our initial theoretical analysis. The Chakraborty and Kundel results[68] and the constraints on the use of Eq. (12) indicate that we need to develop an understanding of the statistics of breast lesion profiles. It would also be more realistic to use each lesion at its true physical size. Of course one would need to use a fairly large and representative sample of lesion images to produce more realistic CD diagrams. The problem with this approach is mathematical compensation for the 2D profile variations or alternatively, defining ''effective signal amplitude'' and ''effective signal size'' in this context. The answers are straightforward for model observer detection in white noise. The answers are not known for detection in power-law noise or mammograms.

It is clear that much more work is needed to obtain a better understanding of the spatial statistics of mammograms and lesions. There is also a need for more realistic experiments. We intend to pursue these lines of investigation. The importance of this work may increase in the future, when mammograms images are viewed on CRTs for primary interpretation. The unexpected relationship between lesion contrast and detectability plus our finding that human results are in reasonable agreement with theoretical observer predictions suggest that it may be possible to use these models to develop image processing algorithms and display methods that will help increase the accuracy of digital mammogram interpretation.

ACKNOWLEDGMENTS

We would like to thank a number of people who gave valuable assistance. Larry Clarke and Maria Kallergi of University of South Florida provided the digitized mammograms and a variety of related information. Jack Beutel, Sterling Diagnostic Imaging, digitized the tumor specimen radiographs and provided Min-R/Microvision H&D curve data. We would also like to thank Craig Abbey, Douglas Adams, John Boone, Dev Chakraborty, Miguel Eckstein, John Heine, Harold Kundel, Kyle Myers, Darrell Smith, and Robert Wagner for helpful discussions and suggestions. This research was supported by Grant No. R01-CA58302 from the US National Cancer Institute. Some preliminary results related to this work were presented at the 1999 SPIE Medical Imaging conference in San Diego and were published in SPIE Proceedings.[72]

[a]Electronic mail: burgess@bwh.harvard.edu

[1] G. Revesz, H. L. Kundel, and M. A. Graber, ''The influence of structured noise on detection of radiologic abnormalities,'' Invest. Radiol. **9**, 479–486 (1974).

[2] H. L. Kundel *et al.*, ''Nodule detection with and without a chest image,'' Invest. Radiol. **20**, 94–99 (1985).
[3] E. Samei, W. Eyler, and L. Baron, ''Effects of anatomical structure on signal detection,'' in *Handbook Of Medical Imaging: Physics and Psychophysics*, edited by J. Beutel, H. Kundel, and R. Van Metter (SPIE Press, Bellingham WA, 2000), Vol. 1, Physics and Psychophysics, pp. 655–682.
[4] N. V. S. Graham, *Visual Pattern Analyzers* (Oxford University Press, New York, 1989).
[5] D. J. Swift and R. A. Smith, ''Spatial frequency masking and Weber's law,'' Vision Res. **23**, 495–505 (1983).
[6] R. A. Smith and D. J. Swift, ''Spatial-frequency masking and Birdsall's theorem,'' J. Opt. Soc. Am. A **2**, 1593–1599 (1985).
[7] F. O. Bochud *et al.*, ''Estimate of the noisy component of anatomical backgrounds,'' Med. Phys. **26**, 1365–1370 (1999).
[8] F. O. Bochud, C. K. Abbey, and M. P. Eckstein, ''Further investigation of the effect of phase spectrum on visual detection in structured backgrounds,'' *Medical Imaging 1999, Image Perception*, San Diego, CA, edited by E. Krupinski (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1999), Vol. 3663, pp. 273–281.
[9] F. Bochud, C. K. Abbey, and M. P. Eckstein, ''Visual signal detection in structured backgrounds. III. Calculation of figures of merit for model observers in statistically nonstationary backgrounds,'' J. Opt. Soc. Am. A **17**, 193–205 (2000).
[10] R. E. Sturm and R. H. Morgan, ''Screen intensification systems and their limitations,'' Am. J. Roentgenol. **62**, 617–634 (1949).
[11] G. Cohen, D. L. McDaniel, and L. K. Wagner, ''Analysis of the variations in contrast-detail experiments,'' Med. Phys. **11**, 469–473 (1984).
[12] M. P. Eckstein, C. K. Abbey, and F. O. Bouchud, ''A practical guide to model observers for visual detection in synthetic and natural noisy images,'' in *Handbook of Medical Imaging*, edited by J. Beutel, H. L. Kundel, and R. L. van Metter (SPIE Press, Bellingham, WA, 2000), Vol. 1, Physics and Psychophysics, pp. 593–628.
[13] M. S. Landy and J. A. Movshon, *Computational Models of Visual Processing* (MIT, Cambridge, MA, 1991).
[14] A. E. Burgess and B. Colborne, ''Visual signal detection IV: Observer inconsistency,'' J. Opt. Soc. Am. A **5**, 617–627 (1988).
[15] A. E. Burgess, ''Prewhitening revisited,'' *Medical Imaging 1998, Image Perception*, San Diego, CA, edited by H. L. Kundel (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1998), Vol. 3340, pp. 55–64.
[16] K. J. Myers *et al.*, ''Effect of noise correlation on detectability of disk signals in medical imaging,'' J. Opt. Soc. Am. A **2**, 1752–1759 (1985).
[17] J. P. Rolland and H. H. Barrett, ''Effect of random background inhomogeneity on observer detection performance,'' J. Opt. Soc. Am. A **9**, 649–658 (1992).
[18] A. E. Burgess, ''Statistically defined backgrounds: Performance of a modified nonprewhitening matched filter model,'' J. Opt. Soc. Am. A **11**, 1237–42 (1994).
[19] A. E. Burgess, X. Li, and C. K. Abbey, ''Visual signal detectability with two noise components: anomalous masking effects,'' J. Opt. Soc. Am. A **14**, 2420–2442 (1997).
[20] A. E. Burgess, ''Visual signal detection with two-component noise: low-pass spectrum effects,'' J. Opt. Soc. Am. A **16**, 694–704 (1999).
[21] K. J. Myers and H. H. Barrett, ''Addition of a channel mechanism to the ideal-observer model,'' J. Opt. Soc. Am. A **4**, 2447–2457 (1987).
[22] H. H. Barrett *et al.*, ''Model observers for assessment of image quality,'' Proc. Natl. Acad. Sci. U.S.A. **90**, 9758–9765 (1993).
[23] C. K. Abbey and F. O. Bouchud, ''Modelling visual signal detection tasks in correlated image noise with linear observer models,'' in *Handbook Of Medical Imaging: Physics and Psychophysics*, edited by J. Beutel, H. Kundel, and R. Van Metter (SPIE Press, Bellingham, WA, 2000), Vol. 1. Physics and Psychophysics, pp. 629–654.
[24] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics. Reprinted by Peninsula Publishing, Los Altos, CA.* (1988) (Wiley, New York, 1966).
[25] R. F. Wagner and D. G. Brown, ''Unified SNR analysis of medical imaging systems,'' Phys. Med. Biol. **30**, 489–518 (1985).
[26] K. J. Myers, ''Ideal observer models of visual signal detection,'' in *Handbook of Medical Imaging: Physics and Psychophysics*, edited by J. Beutel, H. Kundel, and R. Van Metter (SPIE, Bellingham, WA, 2000), Vol. 1, Physics and Psychophysics, pp. 558–592.
[27] R. F. Wagner, D. G. Brown, and M. S. Pastel, ''Application of information theory to the assessment of computed tomography,'' Med. Phys. **6**, 83–94 (1979).
[28] P. F. Judy, R. G. Swensson, and M. Szulc, ''Lesion detection and signal-to-noise ratio in CT,'' Med. Phys. **8**, 13–23 (1981).
[29] A. E. Burgess *et al.*, ''Efficiency of human visual discrimination,'' Science **214**, 93–94 (1981).
[30] M. J. Tapiovaara and R. F. Wagner, ''SNR and noise measurements for medical imaging: I. A practical approach based on statistical decision theory,'' Phys. Med. Biol. **38**, 71–92 (1993).
[31] M. Ishida *et al.*, ''Digital image processing: effect on detectability of simulated low-contrast radiographic patterns,'' Radiology **150**, 569–575 (1984).
[32] H. H. Barrett, K. J. Myers, and R. F. Wagner, ''Beyond signal detection theory,'' *Application of Optical Instrumentation in Medicine XIV and Picture Archiving and Communications (PACS IV) for Medical Applications, Newport Beach, CA* (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1986), Vol. 626, pp. 231–239.
[33] R. D. Fiete *et al.*, ''Hotelling trace criterion and its correlation with human-observer performance,'' J. Opt. Soc. Am. A **4**, 945–953 (1987).
[34] H. H. Barrett *et al.*, ''Stabilized estimates of Hotelling observer detection performance in patient structured noise,'' *Medical Imaging 1998, Image Perception*, San Diego, CA, edited by E. Krupinski (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1998), Vol. 3340, pp. 27–43.
[35] M. P. Eckstein, ''Models for human signal detection in spatiotemporal noise,'' Ph.D. dissertation, Univ. Calif. Los Angeles, Los Angeles, CA, 1994.
[36] S. Daly, ''The visual differences predictor: an algorithm for the assessment of image fidelity,'' in *Digital Images and Human Vision*, edited by A. B. Watson (MIT, Cambridge, Massachusetts, 1993).
[37] M. P. Eckstein *et al.*, ''The effect of image compression in model and human performance,'' *Medical Imaging 1999, Image Perception*, San Diego, CA, edited by E. Krupinski (Proceedings of the Society of Photo-optical Instrumentation Engineers, San Diego, CA, 1999), Vol. 3663, pp. 243–252.
[38] A. B. Watson, ''Detection and recognition of simple spatial forms,'' in *Physical and Biological Processing of Images*, edited by O. J. Braddick and A. C. Sleigh (Springer-Verlag, New York, 1983).
[39] C. K. Abbey, ''Assessment of reconstructed images,'' Ph.D. dissertation, Univ. of Arizona, 1998.
[40] H. Wilson and J. Bergen, ''A four-mechanism model for threshold spatial vision,'' Vision Res. **19**, 19–32 (1979).
[41] A. Papoulis, *Systems and Transforms with Applications in Optics* (MacGraw-Hill, New York, 1968).
[42] B. Zheng, Y.-H. Chang, and D. Gur, ''Adaptive computer-aided diagnosis scheme of digitized mammograms,'' Acad. Radiol. **3**, 806–814 (1996).
[43] A. E. Burgess, ''Mammographic structure: Data preparation and spatial statistics analysis,'' *Medical Imaging 1998, Image Processing*, San Diego, CA, edited by K. Hanson (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1999), Vol. 3661, pp. 642–653.
[44] J. J. Heine, S. R. Deans, and L. P. Clarke, ''Multiresolution probability analysis of random fields,'' J. Opt. Soc. Am. A **16**, 6–16 (1999).
[45] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products* (Academic, New York, 1994).
[46] E. Samei *et al.*, ''Comparison of observer performance for real and simulated nodules in chest radiography,'' in *Medical Imaging 1996: Image Perception*, Newport Beach, CA, edited by H. L. Kundel (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1996), Vol. 2712, pp. 60–70.
[47] A. E. Burgess and S. Chakraborty, ''Producing lesions for hybrid images: extracted tumours and simulated microcalcifications,'' *Medical Imaging 1999, Image Perception*, San Diego, CA, edited by E. Krupinski (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1999), Vol. 3663, pp. 316–322.
[48] M. Kallergi *et al.*, ''Evaluation of a CCD-based film digitizer for digital mammography,'' in *Medical Imaging 1997: Physics of Medical Imaging*, Newport Beach, CA, edited by R. L. Van Metter and J. Beutel (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1997), Vol. 3032, pp. 282–291.

[49] H. C. Lee, S. Daly, and R. L. Van Metter, ''Visual optimization of radiographic tone scale,'' *Medical Imaging 1997: Image Perception*, Newport Beach, CA, edited by H. L. Kundel (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1997), Vol. 3036, pp. 118–129.
[50] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications* (Cambridge University Press, Cambridge, 1993).
[51] J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement procedures* (Wiley, New York, 1986).
[52] W. H. Press *et al.*, *Numerical Recipes in Fortran*, 2nd ed. (Cambridge University Press, Cambridge 1992).
[53] B. B. Mandelbrot, ''New methods in statistical economics,'' J. Political Economy **71**, 421–440 (1963).
[54] J. C. Russ, *Fractal Surfaces* (Plenum, New York, 1994).
[55] P. G. J. Barten, ''Subjective image quality of high-definition television pictures,'' SID Symp. Digest **31**, 239–243 (1990).
[56] H. L. Kundel *et al.*, ''A circle cue enhances detection of simulated masses on mammographic backgrounds,'' in *Medical Imaging 1997, Image Perception*, Newport Beach, CA, edited by H. L. Kundel (Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1997), Vol. 3036, pp. 81–84.
[57] P. F. Judy *et al.*, ''Observer detection efficiency with target size uncertainty,'' *Medical Imaging 1995, Image Perception*, San Diego, CA, edited by H. L. Kundel, San Diego, CA, (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1995), Vol. 2436, pp. 10–17.
[58] D. Reintgen *et al.*, ''The anatomy of missed breast cancers,'' Surgical Oncology **2**, 65–75 (1993).
[59] A. B. Watson and D. Pelli, ''QUEST: a Bayesian adaptive psychometric method,'' Percept. Psychophys. **33**, 113–120 (1983).
[60] A. E. Burgess, ''Comparison of receiver operating characteristic and forced choice observer performance measurement methods,'' Med. Phys. **22**, 643–655 (1995).
[61] J. A. Swets, *Signal Detection and Recognition by Human Observers* (Wiley, New York, 1964).
[62] A. E. Burgess and H. Ghandeharian, ''Visual signal detection II. Signal location identification,'' J. Opt. Soc. Am. A **1**, 906–910 (1984).
[63] M. P. Eckstein, J. S. Whiting, and J. P. Thomas, ''Role of knowledge in human visual temporal integration in spatiotemporal noise,'' J. Opt. Soc. Am. A **13**, 1960–1968 (1996).
[64] D. G. Brown, M. F. Insana, and M. Tapiovaara, ''Detection performance of the ideal decision function and its MacLaurin expansion,'' J. Acoust. Soc. Am. **97**, 379–398 (1995).
[65] J. J. Kulikowski and P. E. King-Smith, ''Spatial arrangement of line, edge, and grating detectors revealed by subthreshold summation,'' Vision Res. **13**, 1455–1478 (1973).
[66] R. M. Shapley, ''Gaussian bars and rectangular bars: the influence of width and gradient on visibility,'' Vision Res. **14**, 1457–1462 (1974).
[67] A. E. Burgess, K. Humphrey, and R. F. Wagner, ''Detection of bars and discs in quantum noise,'' in *Application of Optical Instrumentation in Medicine VII*, edited by J. E. Gray (Proceedings of the Society of Photo-optical Instrumentation Engineers, Toronto, Ont., 1979), Vol. 173, pp. 34–40.
[68] D. Chakraborty and H. L. Kundel, ''Anomalous nodule visibility effects in mammographic images,'' *Medical Imaging 2001, Image Perception and Performance*, San Diego, CA, edited by E. Krupinski (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 2001), Vol. 4324 (in press).
[69] R. F. Voss, ''Fractals in nature: from characterization to simulation,'' in *The Science of Fractal Images*, edited by M. F. Barnsley *et al.* (Springer-Verlag, New York, 1988).
[70] C. van Gils *et al.*, ''Effect of mammographic breast density on breast cancer screening performance: a study in Nijmegen, the Netherlands,'' J. Epidemiol. Community Health **52**, 267–271 (1998).
[71] R. D. Rosenberg *et al.*, ''Effects of age, breast density, ethnicity, and estrogen replacement therapy on screening mammographic sensitivity and cancer stage at diagnosis: review of 183,134 screening mammograms in Albuquerque, New Mexico,'' Radiology **209**, 511–518 (1998).
[72] A. E. Burgess, F. L. Jacobson, and P. F. Judy, ''On the detection of lesions in mammographic structure,'' *Medical Imaging 1998, Image Perception*, San Diego, CA, edited by E. Krupinski (Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 1999), Vol. 3663, pp. 304–315.