Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 9758–9765, November 1993
Colloquium Paper

This paper was presented at a colloquium entitled "Images of Science: Science of Images," organized by Albert V. Crewe, held January 13 and 14, 1992, at the National Academy of Sciences, Washington, DC.

Model observers for assessment of image quality

HARRISON H. BARRETT*, JIE YAO*, JANNICK P. ROLLAND†, AND KYLE J. MYERS‡

*Department of Radiology and Optical Sciences Center, University of Arizona, Tucson, AZ 85724; †Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599; and ‡Center for Devices and Radiological Health, Food and Drug Administration, Rockville, MD 20857

ABSTRACT **Image quality can be defined objectively in terms of the performance of some "observer" (either a human or a mathematical model) for some task of practical interest. If the end user of the image will be a human, model observers are used to predict the task performance of the human, as measured by psychophysical studies, and hence to serve as the basis for optimization of image quality. In this paper, we consider the task of detection of a weak signal in a noisy image. The mathematical observers considered include the ideal Bayesian, the nonprewhitening matched filter, a model based on linear-discriminant analysis and referred to as the Hotelling observer, and the Hotelling and Bayesian observers modified to account for the spatial-frequency-selective channels in the human visual system. The theory behind these observer models is briefly reviewed, and several psychophysical studies relating to the choice among them are summarized. Only the Hotelling model with channels is mathematically tractable in all cases considered here and capable of accounting for all of these data. This model requires no adjustment of parameters to fit the data and is relatively insensitive to the details of the channel mechanism. We therefore suggest it as a useful model observer for the purpose of assessing and optimizing image quality with respect to simple detection tasks.**

Image quality, for scientific and medical purposes, can be defined in terms of how well desired information can be extracted from the image. In other words, image quality is measured by the performance of some "observer" on some specific task (1–3). The observer can be a human, such as a physician trying to make a diagnosis, or it can be a mathematical model or a computer algorithm. The tasks can be divided generically into classification and estimation tasks (4). In medical applications, an example of a classification task would be lesion detection, while an estimation task might be determination of the volume of blood expelled from the heart on each beat.

For classification tasks performed by a human observer, psychophysical studies and receiver operating characteristic (ROC) analysis provide a reproducible, quantitative measure of image quality (2, 3, 5), but such studies are time consuming and require large numbers of images. Moreover, they do not provide an easy way to see how image quality is related to various parameters of the imaging system or processing algorithm. For these reasons, there is considerable interest, especially in the radiological literature (6–8), in mathematical model observers. If the ultimate observer will be a human rather than a machine, the objective of the model is to predict accurately the performance of the human. Then the model observer can be used for system evaluation and optimization with some assurance that the system that is best for the model is also best for a human. Model observers may also be used either to replace the human entirely or to augment human performance (9, 10), but in this paper we focus on using the models to predict psychophysical results.

Early efforts on model observers for this purpose concentrated on the ideal Bayesian observer (6, 11–14), which sets an upper bound to the performance of any observer, including the human. Moreover, there are certain tasks for which the performance of the human observer correlates very well with that of the ideal observer. An important example is detection of a nonrandom signal (or discrimination between two nonrandom signals) in Gaussian noise. For this task, the strategy of the ideal observer is to perform a linear filtering operation on the image and to compare the result to a threshold in order to make a decision. Since the filter is linear, it is straightforward to calculate the performance of the ideal observer and to compare it to that of the human.

The performance of the ideal observer on binary (two-alternative) tasks is quantified by a detectability index d_{ideal}, defined below, which can be calculated from the characteristics of the signal and noise. A similar performance index for the human, denoted d_{human}, can be derived from psychophysical data, and an efficiency for the human observer (12, 13) relative to the ideal can be defined as $(d_{\text{human}}/d_{\text{ideal}})^2$. If the noise is uncorrelated (so-called "white" noise) or has only positive correlations induced by low-pass filtering, this efficiency has been found to be fairly consistently around 50% (14). If the noise has negative correlations induced by high-pass or bandpass filtering, however, the efficiency of the human can be much less (15, 16), and it can depend in a complicated way on the noise characteristics. Thus, the ideal observer has relatively little predictive value for human performance in this kind of noise.

Another drawback to the ideal observer is that it may not be possible to calculate its performance. If the noise is not Gaussian or if the signal or the background on which it is superimposed is random, then the ideal-observer strategy is to calculate a quantity called the likelihood ratio, which is a nonlinear function of the image data. In many realistic imaging situations, we do not have sufficient information about the data statistics to calculate the likelihood ratio or the detection performance associated with it, so the ideal observer is simply not an option.

When the ideal observer is either mathematically intractable or does not predict human performance, we must resort to other model observers. Ones that have been suggested in the radiology literature include a nonprewhitening matched filter (NPWMF) (17), a modified NPWMF incorporating the transfer characteristics of the human eye (18, 19, §), an ideal

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: ROC, receiver operating characteristic; NPW, nonprewhitening; NPWMF, NPW matched filter; SKE/BKE, signal known exactly/background known exactly; TP, true positive; FP, false positive; AUC, area under ROC curve; MTF, modulation transfer function.

§Burgess, A. E., Oral Presentation, Optical Society of America Annual Meeting, Sept. 20–25, 1992, Albuquerque, NM.

observer handicapped by spatial-frequency-selective channels on the front end (20), an optimum linear discriminant referred to as the Hotelling observer (21–25), and the Hotelling observer with spatial-frequency-selective channels (26). It is the goal of this paper to review the mathematics of these potential model observers and some of the psychophysical evidence that relates to the choice among them.

THEORY

Ideal Observer. A digital image consisting of M pixels can be represented as an $M \times 1$ column vector $\mathbf{g}$. The image is a random vector both because of measurement noise in the imaging system and because the objects being imaged are themselves random. We assume that the task of interest is to observe a particular image $\mathbf{g}$ and use it to classify the corresponding object $\mathbf{f}$ that produced the image into one of two classes (e.g., normal vs. abnormal or lesion present vs. lesion absent). The ideal observer performs this task (11) by first computing a scalar test statistic called the likelihood ratio $L(\mathbf{g})$, or equivalently its logarithm $\lambda_{\text{ideal}}(\mathbf{g})$, defined by

$$\lambda_{\text{ideal}}(\mathbf{g}) \equiv \log[L(\mathbf{g})] = \log\left[\frac{p(\mathbf{g}|1)}{p(\mathbf{g}|2)}\right], \qquad [1]$$

where $p(\mathbf{g}|k)$ is the probability density of $\mathbf{g}$ given that it was produced by an object in class k (where $k = 1$ or 2). The classification is performed by comparing this test statistic to a threshold λ_t; if $\lambda(\mathbf{g}) > \lambda_t$, $\mathbf{f}$ is said to belong to class 1, whereas otherwise it is classified into class 2.

An important special case of binary classification is simple signal detection, where the object consists of a nonrandom background on which some weak, nonrandom signal can be superimposed. We refer to this case as SKE/BKE (signal known exactly/background known exactly). The measurement noise, the only source of randomness in this problem, can often be modeled as a correlated multivariate Gaussian process with the same covariance matrix $\mathbf{K}$ for both classes. With these assumptions, $\lambda_{\text{ideal}}(\mathbf{g})$ is given by (11)

$$\lambda_{\text{PW}}(\mathbf{g}) = (\bar{\mathbf{g}}_2 - \bar{\mathbf{g}}_1)^t \mathbf{K}^{-1}\mathbf{g}, \qquad [2]$$

where $\bar{\mathbf{g}}_k$ is the mean image for class k, and the superscript t denotes a matrix transpose. The difference in means $\bar{\mathbf{g}}_2 - \bar{\mathbf{g}}_1$ is the signal to be detected, so the ideal observer first filters the image with the matrix operator $\mathbf{K}^{-1}$ and then performs a matched filtering operation by taking the scalar product with the signal. This is equivalent to operating on both the data and the signal with $\mathbf{K}^{-1/2}$, then forming the scalar product of the two preprocessed vectors. Since $\mathbf{K}^{-1/2}$ serves to remove the correlations between the elements of $\mathbf{g}$, it is often called a prewhitening operation, and $\lambda_{\text{PW}}(\mathbf{g})$ as calculated from Eq. 2 is the output of a prewhitening matched filter. Note that $\lambda_{\text{PW}}(\mathbf{g})$ is linear in $\mathbf{g}$.

Except for this Gaussian SKE/BKE case, the log-likelihood ratio is often very difficult to determine and a highly nonlinear function of $\mathbf{g}$. In particular, if there is any randomness in the objects being imaged or the signal to be detected, the log-likelihood ratio is usually nonlinear. With rare exceptions, it is not possible to calculate the ideal-observer test statistic in these cases.

Performance Measures. We presume that any observer (including the human) makes a decision by computing some test statistic $\lambda(\mathbf{g})$ and comparing it to a decision threshold. For definiteness, we refer to class 1 as the signal-present or positive case, so the decision is "positive" if $\lambda(\mathbf{g})$ exceeds the decision threshold. If the object that produced the image was actually in class 1, we call the positive decision a true positive (TP), while otherwise it is a false positive (FP). The decision threshold controls the trade-off between TP and FP decisions. A plot of TP rate vs. FP rate as the threshold is varied is called a ROC curve, and the area under the ROC curve, denoted AUC, can be adopted as a figure of merit for the data set (1–3). This figure of merit ranges from 0.5 for a worthless test to 1.0 for a perfect one.

Another useful figure of merit is the signal-to-noise ratio associated with the test statistic (1–3), often called a detectability index and referred to as d' or d_a. We denote it simply as d, with subscripts to designate particular observers, and define it by

$$d^2 = \frac{[E(\lambda(\mathbf{g})|2) - E(\lambda(\mathbf{g})|1)]^2}{\frac{1}{2}\operatorname{var}(\lambda(\mathbf{g})|1) + \frac{1}{2}\operatorname{var}(\lambda(\mathbf{g})|2)}, \qquad [3]$$

where $E(\lambda(\mathbf{g})|k)$ is the conditional mean of the test statistic $\lambda(\mathbf{g})$ given that $\mathbf{g}$ comes from class k, while $\operatorname{var}(\lambda(\mathbf{g})|k)$ is the corresponding conditional variance. If $\lambda(\mathbf{g})$ obeys Gaussian statistics, d is related to the AUC by

$$\text{AUC} = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{d}{2}\right), \qquad [4]$$

where $\operatorname{erf}(d/2)$ is the error function (2). By use of this equation, d for the human observer can be derived from the ROC curve without knowledge of the specific test statistic.

Linear Discriminants. Since the log-likelihood ratio is usually mathematically intractable when it is a nonlinear function of $\mathbf{g}$, it is natural to consider linear discriminants in which the test statistic is constrained from the outset to be linear. A general linear test statistic is a scalar product of the form $\lambda_{\text{lin}} = \mathbf{u}^t\mathbf{g}$, and the objective of discriminant analysis is to choose the discriminant function $\mathbf{u}$ in such a way as to maximize performance of the classification task. If one has available a training set of data vectors of known classification, the optimum $\mathbf{u}$ was determined by Fisher in 1936 (27).

A closely related concept is that of class separability. In 1931, Hotelling (28) proposed a measure called T^2 to test the null hypothesis that the two samples of random vectors were drawn from populations with the same mean vector. If there is a difference in means of the two populations, T^2 is a quantitative estimate of how large it is. Later work showed the connection between T^2 and linear discriminants. Fisher's discriminant is a way of calculating a scalar test statistic that preserves the class separability inherent in the data, and the d figure of merit for the Fisher discriminant is proportional to T.

One conceptual way in which these ideas might be applied to image quality would be to draw sample images from two classes, compute T^2 from the pixel values, and use it to estimate the separability of the two classes with respect to the particular imaging system under study. Different imaging systems would yield different T^2 values, and we could say that the system that gave the largest separability, as quantified by T^2, was the best for that particular classification task. Equivalently, that system would be the one that gave best performance when the observer was the Fisher linear discriminant.

Unfortunately this approach does not work in practice. Calculation of either Hotelling's T^2 or the Fisher discriminant from a training set requires inverting a sample covariance matrix, but the inverse does not exist unless the number of random vectors in the sample exceeds the number of elements in each vector. With images, that means that the number of images must exceed the number of pixels in each image, a condition that is virtually impossible to achieve with images of any reasonable size. In pattern recognition, this dimensionality problem is dealt with by extracting a small

number of features from the images and calculating the linear discriminant from the features rather than from the original pixel values. This approach has not found use in image-quality assessment, probably because there is no general theory to guide feature selection.

The Hotelling Observer. To overcome the dimensionality problem, we use simulated images, for which it is often possible to calculate the ensemble covariance matrices or at least to get good, full-rank estimates of them. If ensemble covariance matrices are available, the ensemble analog of Hotelling's T^2 can be used as a figure of merit for image quality (21–24). Denoted J, this metric can be regarded as the limit of T^2 when the number of sample images in each class approaches infinity. Specific examples of how J is calculated are given in refs. 23, 29, and 30.

To give a more precise definition of J, we define two "scatter matrices" $\mathbf{S}_1$ and $\mathbf{S}_2$. The interclass scatter matrix $\mathbf{S}_1$, which measures the separation of the two class means, is defined for binary problems by

$$\mathbf{S}_1 = P_1P_2(\bar{\mathbf{g}}_2 - \bar{\mathbf{g}}_1)(\bar{\mathbf{g}}_2 - \bar{\mathbf{g}}_1)^t, \qquad [5]$$

where P_k is the probability of occurrence of class k ($k = 1,2$), and $\bar{\mathbf{g}}_k$ is the class mean for the kth class.

The intraclass scatter matrix $\mathbf{S}_2$ is the arithmetic average of the ensemble covariance matrices for the two classes. It is given by

$$\mathbf{S}_2 \equiv \sum_{k=1}^{2} P_k\mathbf{K}_k = \sum_{k=1}^{2} P_k\langle(\mathbf{g} - \bar{\mathbf{g}}_k)(\mathbf{g} - \bar{\mathbf{g}}_k)^t\rangle_k, \qquad [6]$$

where $\mathbf{K}_k$ is the ensemble covariance matrix of the kth class, and the angular brackets denote a full ensemble average over all objects in class k and all realizations of the measurement noise.

In terms of these scatter matrices, a measure of (ensemble) class separability called the Hotelling trace J is defined by (22, 23)

$$J = \mathrm{tr}(\mathbf{S}_2^{-1}\mathbf{S}_1), \qquad [7]$$

where tr denotes the trace of the matrix. Hotelling's T^2 has exactly the same structure as J but with sample means and covariance matrices in place of the ensemble ones. A key difference is that the ensemble $\mathbf{S}_2$ matrix will usually be invertible, while the sample one will usually not be.

The linear test statistic associated with J is given by (30)

$$\lambda_{\mathrm{Hot}}(\mathbf{g}) = (\bar{\mathbf{g}}_2 - \bar{\mathbf{g}}_1)^t\mathbf{S}_2^{-1}\mathbf{g}. \qquad [8]$$

This form is quite similar to that for the Fisher discriminant except that ensemble means and covariances appear in place of sample ones.

Comparing Eq. 8 with Eq. 2, we see that the Hotelling test statistic has the same form as a prewhitening matched filter, which is the ideal observer for SKE/BKE problems. The main difference is that λ_{Hot} uses $\mathbf{S}_2^{-1}$ as the prewhitening filter while λ_{PW} uses $\mathbf{K}^{-1}$. In SKE/BKE problems, the only source of randomness is the measurement noise, so $\mathbf{S}_2 = \mathbf{K}$ in that case and the Hotelling observer is in fact ideal. More generally, however, the $\mathbf{S}_2$ matrix includes object variability, and the linear Hotelling test statistic is not equivalent to the nonlinear ideal-observer test statistic.

If we assume that λ_{Hot} obeys Gaussian statistics, which can often be argued from the central-limit theorem, J is related to d by (23)

$$J = P_1P_2(d_{\mathrm{Hot}})^2. \qquad [9]$$

NPW Observers. To devise a linear test statistic to predict the performance of human observers, it might be desirable to build in some characteristics of the human visual system. There is some evidence that humans cannot prewhiten correlated noise (14–17), so a number of workers have used the so-called NPWMF, with test statistic given by

$$\lambda_{\mathrm{NPW}}(\mathbf{g}) = (\bar{\mathbf{g}}_2 - \bar{\mathbf{g}}_1)^t\mathbf{g}. \qquad [10]$$

This test statistic is very easy to compute. No detailed knowledge of the noise statistics is needed, and the observer merely forms the scalar product of the mean difference image $\bar{\mathbf{g}}_2 - \bar{\mathbf{g}}_1$ and the image under test. In SKE/BKE problems with nonwhite noise, the performance (d) of the NPW observer is necessarily worse than that of the ideal prewhitening observer; how much worse depends on the nature of the noise correlations.

Eye Models. Another characteristic of the visual system that can be incorporated into the observer models is its modulation transfer function (MTF). If we ignore nonlinearities in the visual system, which is probably valid for low-contrast images, this transfer characteristic can be described by a matrix. The simplest model, treating the eye as a linear, shift-invariant filter, uses a square matrix $\mathbf{E}$ of the form $\mathbf{F}^{-1}\mathbf{M}\mathbf{F}$, where $\mathbf{F}$ represents the two-dimensional discrete Fourier transform and $\mathbf{M}$ specifies the MTF. Any of the observer models discussed above can incorporate the eye model by assuming that the test statistic is computed not from the original data vector $\mathbf{g}$ but rather from a modified vector $\mathbf{Eg}$. This modification does not affect the performance of the ideal or Hotelling observer if $\mathbf{E}$ is invertible, but it can affect the NPW model (18, 19).

Channels. A somewhat different characteristic of the visual system is spatial-frequency-selective channels. It is known that individual neurons in the visual cortex are responsive to grating stimuli in a certain band of two-dimensional spatial frequencies (31). Detection of a weak stimulus within one of these bands is masked by another, stronger, grating only if the two stimuli are within the same band. This observation suggests that the neuron not only functions as a bandpass filter but also computes the integral of the filter output over the band (32); if it were simply a bandpass filter, the human observer could still distinguish the signal to be detected from the masking stimulus. With the integration step, however, all stimuli within the same band are added together, and there is potentially a large loss of information in going through the channels.

As with the eye MTF, the simplest mathematical description of channels (15, 20) is a matrix $\mathbf{V}$, but there is an important difference. Since many different input spatial frequencies contribute to each channel output, the matrix $\mathbf{V}$, unlike $\mathbf{E}$, is highly rectangular and hence not invertible. Even the ideal observer suffers a loss of performance if it has to operate on $\mathbf{Vg}$ instead of $\mathbf{g}$.

EXPERIMENTAL STUDIES

In this section we review a number of psychophysical studies, carried out at the University of Arizona, that relate to the choice of a model observer. The specific images considered were medical γ-ray images produced by a collimator or pinhole, but the results are certainly not particular to γ-ray imaging. All of the studies were carried out with the same methods. Eight to 10 observers viewed each of the images and assigned a certainty to the presence of a signal by means of a six-point rating scale. A ROC curve was calculated for each observer, AUC and d_{human} were derived, and the results were averaged over observers to get the final performance figures. Details of the procedure can be found in refs. 15, 22, and 31.

Correlated Noise. To address the question of whether the human observer could prewhiten the noise in SKE/BKE

problems, Myers *et al.* (15, 16) created a set of images with the same d for the ideal observer but with different degrees of noise correlation. In each case, the object being imaged was a simple flat background, and the task was to detect a weak, nonrandom signal that might be superimposed on the background. Each object was filtered by a low-pass filter representing predetection blur due to the collimator, and Poisson noise in the detection process was modeled by adding uncorrelated noise to the blurred image. The resulting noisy image was passed through a second filter, representing postdetection digital processing of the image.

To produce severely nonwhite noise, Myers chose the transfer function of the second filter to be given approximately by

$$P_2(\rho) = \rho^{n/2}\exp(-\beta\rho^2), \qquad [11]$$

where ρ is the two-dimensional spatial-frequency vector, ρ is its magnitude, and β and n are constants. Since the power spectral density $S(\rho)$ of filtered white noise is proportional to the squared modulus of the filter transfer function, $S(\rho)$ had the form $\rho^n\exp(-2\beta\rho^2)$. The noise autocorrelation function [the Fourier transform of $S(\rho)$] was therefore determined only by the second filter, while the overall transfer function of the system was the product of the transfer functions of the two filters. To focus on the noise correlations, the overall transfer function was constrained to be constant. The noise level was adjusted so that d for the ideal observer remained constant regardless of the exponent n. Since the task was SKE/BKE, $\mathbf{S}_2$ was the same as $\mathbf{K}$, and the ideal observer and the Hotelling observer for this study were identical and given by the prewhitening matched filter of Eq. **2**.

The results of the psychophysical studies (15, 16) are summarized in Fig. 1*A*, where it is seen that d_{human} fell off dramatically as the noise exponent n increased from 1 to 4. Since d_{ideal} was constant, this result indicated that the efficiency of the human relative to the ideal varied by almost a factor of 100, ruling out the prewhitening ideal/Hotelling model as a predictor of human performance for this task.

Fig. 1*B* shows the predictions of several other model observers for this problem. Included are the simple NPWMF and four versions of an ideal observer operating on image data prefiltered through channels as described above. The channels did not overlap in the frequency domain and had abrupt cutoffs, and a simple integral of the amplitude without any nonlinear rectification was performed after bandpass filtering in each channel. The bandwidth of each channel was proportional to its center frequency, and the different points in the figure are for different fractional bandwidths and starting frequency of the lowest channel. All of the channel models considered as well as the NPW model predict the behavior found in the psychophysical studies: strong decrease of d_{human} with increasing n. On this SKE/BKE task and a variety of others, the predictions of the channel models did not depend strongly on the details of the channels, and the channel models were virtually indistinguishable from the NPW model (20). Thus, the prewhitening ideal and Hotelling observers do not predict human performance for this task, while the NPW model and any of a variety of channelized ideal-observer models do so very well.

Random Signals and Backgrounds. To move away from SKE/BKE problems and investigate the effects of signal and background randomness, Fiete *et al.* (23) created a simple two-dimensional phantom of random, overlapping ellipses, roughly representing a liver. The task was to detect a small cold lesion of random size, shape, and contrast. The images were blurred with Gaussian blur functions of different widths, and Gaussian noise of various amplitudes was added; 32 normal and 32 abnormal images were generated for each of nine combinations of blur width and noise level.

Because of the object randomness, the ideal-observer performance was not calculable, but that of the Hotelling observer was. There was no postdetection processing, so the noise would have been uncorrelated if the object had been constant. The object variability, however, produced correlations when the whole ensemble of objects was considered, necessitating the use of the $\mathbf{S}_2^{-1}$ prewhitening filter of Eq. **8**. With some approximations, described by Fiete *et al.* (23, 24), it was possible to calculate both the test statistic and the performance of the Hotelling observer. Psychophysical studies were performed, and d_{human} correlated extremely well with the d_{Hot} ($r \simeq 0.98$).

Later, a more realistic extension of this study was performed, with the objective of determining the optimal colli-

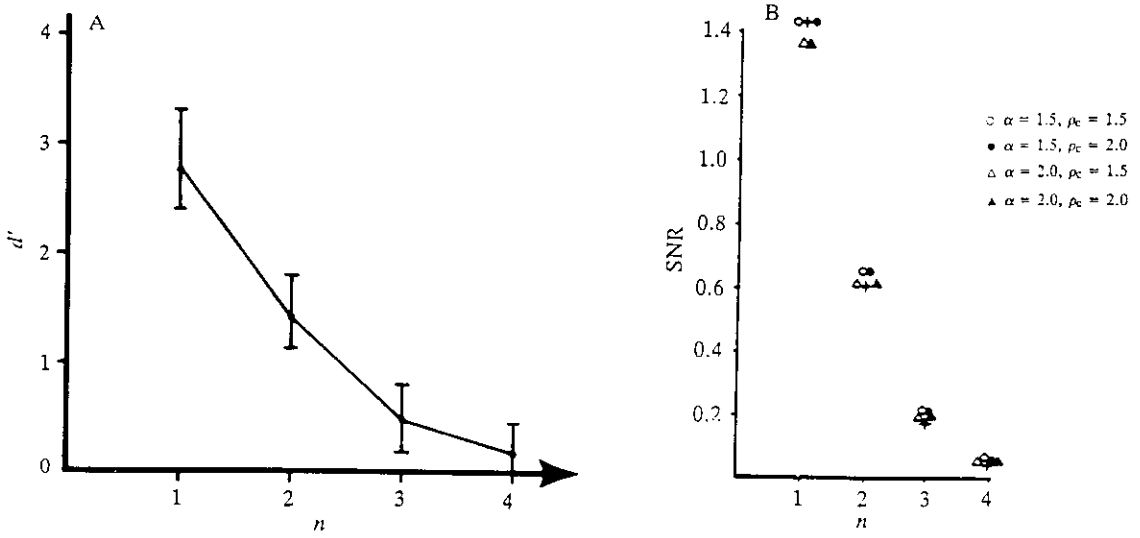

FIG. 1. Results from Myers *et al.* (15, 16) on SKE detection in correlated noise. (*A*) Psychophysical d values (here denoted d') derived from ROC curves vs. the noise exponent n (see Eq. **11**). (*B*) Theoretical predictions of the signal-to-noise ratio (SNR) of several model observers for this task. SNR is calculated directly from Eq. **4** in the text, while the psychophysical value is derived by measuring AUC and inverting Eq. **3**; the two quantities should be equal if the model is an accurate predictor of human performance. In this figure, the models considered are the NPWMF (+) and four different channelized ideal-observer models (○, ●, △, ▲). The channel models differ in the fractional bandwidth of the channels (α) and minimum frequency of the lowest channel (ρ_c). The five models considered here give very similar predictions, all of them showing the large fall-off in performance with n seen in the psychophysical data.

mator to use in planar radiocolloid imaging of the liver (24, 33). In this study, three-dimensional mathematical liver phantoms (34) were used to model a healthy class, while another group of mathematical phantoms with elliptical cold regions in the liver simulated a diseased class. There was considerable randomness in both the liver background and the signal to be detected. Images of these objects through parallel-hole collimators with various bore diameters and bore lengths were calculated. The Hotelling trace J was calculated from these images for each collimator, and psychophysical studies were performed. The results again showed a good correlation between human and Hotelling performance.

Lumpy Backgrounds. The next study in this series, performed by Rolland and Barrett (35, 36), considered detection of an exactly known signal on a random, spatially inhomogeneous background, which we refer to as a lumpy background. In contrast to the Fiete studies, the background in this case was not intended to represent any realistic medical object. Instead it was a stationary random process with a Gaussian autocorrelation function, which made it possible to calculate the performance of the Hotelling observer without any approximations (29, 36, 37). The objects being imaged, consisting of random background samples plus a nonrandom signal in half of the cases, were imaged through a pinhole aperture. The main variables were the diameter of the pinhole, which controlled both the blur and the noise level in the images, and the exposure time, which affected only the noise level and not the blur.

Since the performance of the ideal observer was not calculable for this problem, we compared the Hotelling and NPW models. As in the Fiete studies, there was no postdetection processing, so there were no noise correlations for a single realization of the background. Nevertheless, there were strong correlations when the whole ensemble of random backgrounds was considered, and the prewhitening operation (S_2^{-1}) made a large difference in calculated observer performance (29).

Fig. 2 *A* and *C* shows the performance of the two model observers as a function of pinhole diameter and exposure time for three different levels of background lumpiness, including zero, which corresponds to a nonrandom, spatially homogeneous background. Fig. 2 *B* and *D* shows the results of the psychophysical study by Rolland and Barrett (35, 36). In interpreting this figure, it must be kept in mind that in the case of zero lumpiness there is no noise correlation at all, so

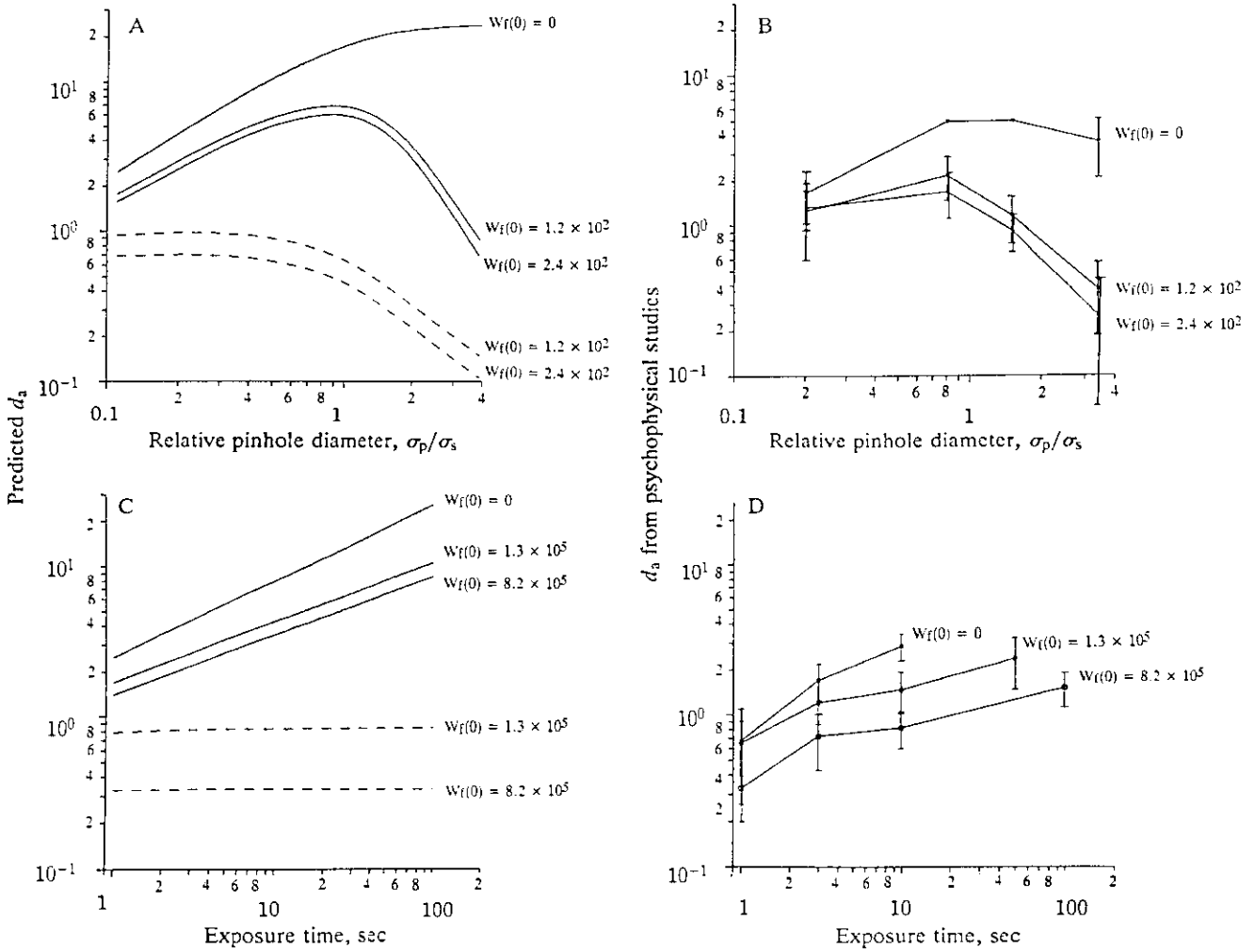

FIG. 2. Results from Rolland and Barrett (35, 36) on SKE detection of a signal superimposed on a random, spatially inhomogeneous background, modeled as a stationary random process. The inhomogeneity ("lumpiness") of the background is specified by the power spectral density of the background $W_f(\rho)$ evaluated at zero spatial frequency ($\rho = 0$). The correlation length of the background was fixed and equal to 3 times the width of the signal to be detected. Theoretical results are shown in *A* and *C* and comparable psychophysical data are shown in *B* and *D*. The independent variables are pinhole diameter (*A* and *B*) or exposure time (*C* and *D*). In all cases, the vertical axes are detectability indices d, here denoted d_a. For the theoretical curves, two model observers are considered: Hotelling (solid lines) and NPW (dashed lines). If the lumpiness $W_f(0) = 0$, these two models are identical, and the predictions are given by the upper curves in *A* and *C*. For nonzero lumpiness, however, there is a large difference in the predictions of the two models, with the NPW giving consistently lower performance. The psychophysical data in *B* and *D* show the same qualitative behavior as the Hotelling model (compare *A* to *B* and *C* to *D*).

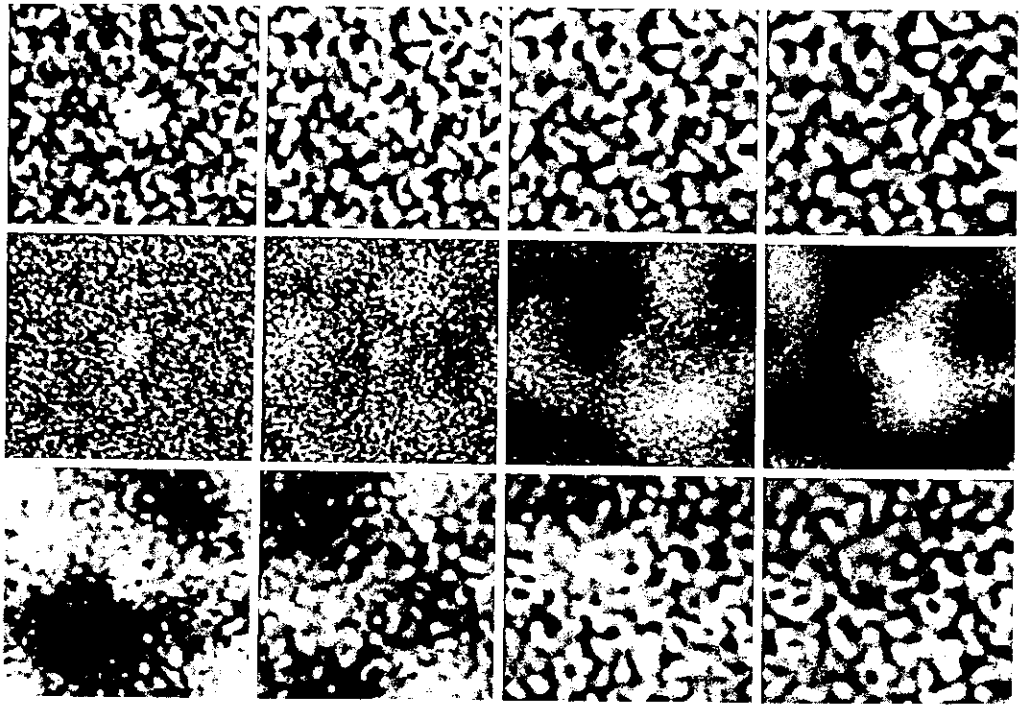

FIG. 3. Images used in various psychophysical studies discussed in this paper. (*Top*) Images used by Myers *et al.* (15, 16); the noise exponent n varies from 1 at left to 4 at right. The d for the ideal observer was held constant, but the signal, readily visible at the top left ($n = 1$), is not perceptible for higher values of n. (*Middle*) Images used by Rolland and Barrett (35, 36). Here the lumpiness $W_f(0)$ increases from left to right. (*Bottom*) Images used by Yao and Barrett (26). Here the lumpiness $W_f(0)$ is fixed at a nonzero value while the noise exponent n goes from 1 at left to 4 at right.

the NPW and Hotelling observers coincide (and are also ideal). Thus, the upper theoretical curves in Fig. 2 *A* and *C* are applicable to both of these observers.

The general trends in the psychophysical data are in excellent accord with the Hotelling model. For both human and Hotelling observers, there is a clear optimum in d vs. pinhole diameter, an effect not predicted by the NPW model. The performance of both human and Hotelling observers increases with exposure time, while that of the NPW observer does not. Furthermore, the introduction of background lumpiness degrades the performance of the NPW observer, relative to the zero-lumpiness case, far more than it degrades either the human or the Hotelling observer. Overall, the efficiency of the human relative to the Hotelling model, $(d_{human}/d_{Hot})^2$, is substantially constant, while $(d_{human}/d_{NPW})^2$ varies by almost 2 orders of magnitude over the range of parameters explored in Fig. 2.

A second study using the lumpy-background paradigm was reported by Yao *et al.*[¶] They noted that the performance of the Hotelling observer is determined solely by the mean vectors and covariance matrices and does not depend on the detailed shape of the probability density function for the grey levels in the image. To test whether this prediction was also true for humans, Yao constructed three sets of backgrounds with the same means and covariances (hence the same S_1 and S_2) but with different grey-level histograms. The Hotelling performance was the same for all three sets by construction, and Yao's psychophysical study showed that human performance was also the same within experimental error.

Can Humans Prewhiten? The studies by Rolland and Barrett (35, 36) and Yao and Barrett (26) lead to a clear and unequivocal conclusion: The psychophysical results for SKE signal detection in a lumpy background correlate very well with the prewhitening model (Hotelling in this case) and not at all with the NPW model. This conclusion, though demanded by the data, is in striking contrast to that derived from the work of Myers *et al.* (15, 16), which gives an equally unequivocal conclusion: For SKE/BKE detection tasks with noise correlations induced by postdetection filtering, the psychophysical data correlate very well with the NPW model and not at all with the ideal/Hotelling prewhitening matched filter. These two sets of studies, performed in the same laboratory with the same methods, thus give opposite answers to the question of whether humans can prewhiten correlated noise. To be sure, the nature of the correlations is somewhat different in the two cases. In the work of Myers the correlation produced by the second filter would be visible in a single image (Fig. 3), while in the Rolland and Yao work it would be evident only when an ensemble of images was analyzed. The prewhitening filter also has a different form, $\mathbf{K}^{-1}$ in Myers' work and the more general $\mathbf{S}_2^{-1}$ for Rolland and Yao. Nevertheless, the striking difference in the results must be resolved before we can adopt a single model observer for assessment of image quality.

A Synthesis: The Channelized Hotelling Model. Since the Myers *et al.* (15, 16), Rolland and Barrett (35, 36), and Yao *et al.* experiments had correlations of different origin, Yao and Barrett performed another experiment (26) in which both kinds of correlation were present. The task was again detection of a known signal on a lumpy background, but the same postprocessing filters used by Myers *et al.* (15, 16) were also employed. The experimental conditions included zero lumpiness, reproducing the setup used by Myers *et al.*, and noise exponent $n = 0$, reproducing the setup used by Rolland and Barrett (35, 36).

Since the Myers results had been equally well predicted by the somewhat ad hoc NPW model and the physiologically based channel models, Yao considered addition of a channel mechanism to the Hotelling model. The nonoverlapping channels had constant fractional bandwidths, abrupt cutoffs and a simple amplitude integration just as in the Myers model. The scatter matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ as well as λ_{Hot} and d_{Hot} were calculated on the basis of $\mathbf{Vg}$ rather than $\mathbf{g}$.

[¶]Yao, J., Barrett, H. H. & Rolland, J. P., Oral Presentation, Optical Society of America Annual Meeting, Nov. 3–8, 1991, San Jose, CA.

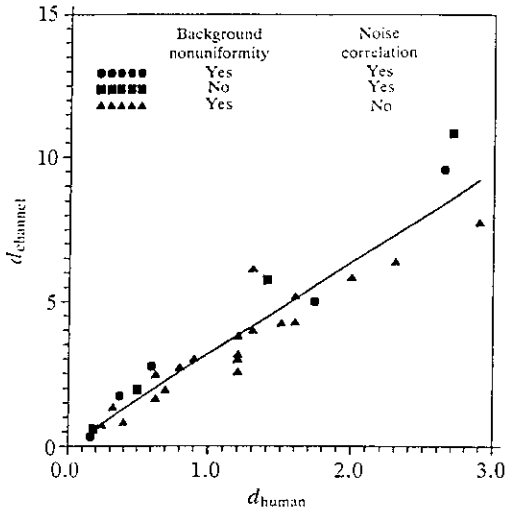

FIG. 4. Compilation of psychophysical data obtained by Myers *et al.* (15, 16) (■), Rolland and Barrett (35, 36) (▲), and Yao and Barrett (26) (●) plotted vs. the predictions of the channelized Hotelling model. This common model fits all three data sets within experimental error.

The psychophysical results obtained in this study (26) are plotted in Fig. 4 along with the results of Myers (15, 16) and Rolland and Barrett (35, 36). In each case a plot of d_{human} against the d for the channelized Hotelling model ($d_{channel}$) is given. All three sets of results are found to fall closely on a straight line, indicating that the channelized Hotelling model is capable of predicting human performance with both kinds of correlation separately or together.

DISCUSSION AND CONCLUSIONS

In this paper, we have considered a number of possible model observers for the task of detection of a "blob"—i.e., a spatially compact signal of relatively low contrast. In some cases, the images were more-or-less realistic simulations of γ-ray images in nuclear medicine, and the blob represented a tumor or other lesion. In other cases, however, the choice of image was dictated by the desire to produce large and easily detectable differences between the predictions of competing models.

In contrast to most psychophysical investigations, our goal was not so much to understand the human visual system as to devise a useful figure of merit for quantitative assessment and optimization of image quality. The ideal Bayesian observer cannot be widely used for this purpose for two reasons: It is not mathematically tractable if there is significant object randomness, and it does not predict human performance in the presence of correlated noise.

To overcome the first objection, we introduced the Hotelling observer, a linear-discriminant model that is often tractable even when the ideal observer is not. In all situations we investigated, the Hotelling model was a good predictor of human performance if there was no postdetection filter to introduce noise correlations. Based on some theoretical studies not reported here, we also suspect that the Hotelling model will be successful with postdetection filtering if it has a low-pass character. With high-pass or bandpass filtering, however, large deviations between human and Hotelling performance were found. In the work of Myers *et al.* (15, 16), the efficiency of the human relative to the ideal fell by 2 orders of magnitude as the degree of high-pass correlation (as measured by the noise exponent n) was increased. The ideal observer, identical with the Hotelling observer on these tasks, did not fit these data, but the NPW model fit very well.

On the other hand, the NPW model did not predict any of the salient features of Rolland's data on SKE detection on a random background. In that work, the efficiency of the human relative to the NPW model varied by some 2 orders of magnitude. The Hotelling model, however, accounted for all of the qualitative behavior of the results and gave an efficiency that was much more nearly constant. The Hotelling model also worked quite well in predicting human performance on realistic tumor-detection tasks in simulated liver images.

The apparent contradiction between these two sets of results is resolved by incorporating channels into the Hotelling model. The resulting channelized Hotelling model fits a large body of data without adjustment of parameters. The channels have very little effect on performance with white or low-pass noise, so the Hotelling and channelized Hotelling models appear to work equally well in those cases. With noise correlations produced by high-pass filtering, the channels cause the same kind of degradation in performance as does the ad hoc assumption that the human cannot prewhiten. Thus, the channelized Hotelling model seems to be in good accord with all available psychophysical evidence for SKE/BKE blob-detection tasks, with or without noise correlations. It also accounts for all of our data on lumpy backgrounds, including images processed through high-pass filters.

Of course, we can never rule out the possibility that the same data could be explained by other models. In particular, we have not investigated the use of various models for the MTF of the eye in conjunction with the NPW assumption. Even if other models could be found to fit the data, however, the channelized Hotelling model would remain attractive on several accounts. It is mathematically tractable in a wide variety of problems, including those with realistic object and signal models. It accounts for arbitrary random variability in the object and signal, and it allows any form of postprocessing, including high-pass filtering and even nonlinear processing. The channels have a physiological basis, and the predictions of the model are robust with respect to the details of the channels.

In summary, we offer the working hypothesis that the channelized Hotelling observer model can be used as a general tool for assessment of image quality with respect to detection and discrimination tasks, and we encourage further psychophysical investigations to discover the limits of its applicability. In particular, further studies to compare the performance of channelized Hotelling and human observers with different kinds of signal and background variability would be worthwhile. For example, the mathematical framework of the channelized Hotelling model allows us to analyze the effects of varying signal size, signal location, or background correlation structure. It remains an open question whether the mathematical model and the human observer will behave the same way in response to these variations.

Note Added in Proof. Burgess (38) has recently considered a variety of eye models used in conjunction with a NPW observer. He compared the predictions of these models to the data of Rolland and Barrett and found that he could account for many, but not all, features of the data.

We have benefited greatly from discussions with Robert Wagner, David Brown, and Arthur Burgess. This work was supported in part by the National Cancer Institute under Grants PO1 CA23417 and RO1 CA52643.

1. Swets, J. (1988) *Science* **240**, 1285–1293.
2. Swets, J. & Pickett, R. M. (1982) *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory* (Academic, New York).
3. Metz, C. E. (1986) *Invest. Radiol.* **21**, 720–733.

4. Barrett, H. H. (1990) *J. Opt. Soc. Am. A Opt. Image Sci.* **7**, 1266–1278.
5. Massof, R. W. & Emmel, T. C. (1987) *Appl. Opt.* **26**, 1395–1408.
6. Wagner, R. F. & Brown, D. G. (1985) *Phys. Med. Biol.* **30**, 489–518.
7. Judy, P. F. & Swensson, R. G. (1987) *J. Opt. Soc. Am. A Opt. Image Sci.* **4**, 954–965.
8. Chesters, M. S. (1992) *Phys. Med. Biol.* **7**, 1433–1476.
9. Chan, H.-P., Doi, K., Galhotra, S., Vyborny, C. J., Schmidt, R. A., Metz, C. E., Lam, K. L., Ogura, T., Wu, Y. & MacMahon, H. (1990) *Invest. Radiol.* **25**, 1102–1110.
10. Doi, K., Giger, M. L., MacMahon, H., Hoffman, K. R., Nishikawa, R. M., Schmidt, R. A., Chua, K.-G., Katsuragawa, S., Nakamori, N., Sanada, S., Yoshimura, H., Metz, C. E., Montner, S. M., Matsumoto, T., Chen, S. & Vyborny, C. J. (1992) *Semin. Ultrasound Comput. Tomography Magn. Reson. Imaging* **13**, 140–152.
11. Whalen, A. D. (1971) *Detection of Signals in Noise* (Academic, New York).
12. Tanner, W. P., Jr., & Birdsall, T. G. (1958) *J. Acoust. Soc. Am.* **30**, 922–928.
13. Barlow, H. B. (1962) *J. Physiol. (London)* **160**, 155–168.
14. Burgess, A. E., Wagner, R. F., Jennings, R. J. & Barlow, H. B. (1981) *Science* **214**, 93–94.
15. Myers, K. J. (1985) Ph.D. dissertation (Univ. of Arizona, Tucson).
16. Myers, K. J., Barrett, H. H., Borgstrom, M. C., Patton, D. D. & Seeley, G. W. (1985) *J. Opt. Soc. Am. A Opt. Image Sci.* **2**, 1752–1759.
17. Wagner, R. F. (1978) *Proc. SPSE* **22**, 41–26.
18. Loo, L.-N. D. (1982) Ph.D. thesis (Univ. of Chicago, Chicago).
19. Giger, M. L. & Doi, K. (1985) *Med. Phys.* **12**, 201–208.
20. Myers, K. J. & Barrett, H. H. (1987) *J. Opt. Soc. Am. A Opt. Image Sci.* **4**, 2447–2457.
21. Barrett, H. H., Myers, K. J. & Wagner, R. F. (1986) *Proc. SPIE Int. Soc. Opt. Eng.* **626**, 231–239.
22. Smith, W. E. & Barrett, H. H. (1986) *J. Opt. Soc. Am. A Opt. Image Sci.* **3**, 717–725.
23. Fiete, R. D., Barrett, H. H., Smith, W. E. & Myers, K. J. (1987) *J. Opt. Soc. Am. A Opt. Image Sci.* **4**, 945–953.
24. Fiete, R. D., Barrett, H. H., Cargill, E. B., Myers, K. J. & Smith, W. E. (1987) *Proc. SPIE Int. Soc. Opt. Eng.* **727**, 298–305.
25. Fiete, R. D. (1987) Ph.D. thesis (Univ. of Arizona, Tucson).
26. Yao, J. & Barrett, H. H. (1992) *Proc. SPIE Int. Soc. Opt. Eng.* **1768**, 161–168.
27. Fisher, R. A. (1936) *Ann. Eugenics* **7**, 179–188.
28. Hotelling, H. (1931) *Ann. Math. Stat.* **2**, 360–378.
29. Myers, K. J., Rolland, J. P., Barrett, H. H. & Wagner, R. F. (1990) *J. Opt. Soc. Am. A Opt. Image Sci.* **7**, 1279–1293.
30. Barrett, H. H., Gooley, T. A., Girodias, K. A., Rolland, J. P., White, T. A. & Yao, J. (1991) in *XIIth International Conference on Information Processing in Medical Imaging*, eds. Colchester, A. C. F. & Hawkes, D. J. (Springer, New York), pp. 458–473.
31. Sachs, M., Nachmias, J. & Robson, J. (1971) *J. Opt. Soc. Am.* **61**, 1176–1186.
32. Wagner, R. F. & Weaver, K. E. (1972) *Proc. SPIE Int. Soc. Opt. Eng.* **35**, 83–94.
33. White, T. A., Barrett, H. H., Cargill, E. B., Fiete, R. D. & Ker, M. (1989) *J. Nucl. Med.* **30**, 892 (abstr.).
34. Cargill, E. B. (1989) Ph.D. dissertation (Univ. of Arizona, Tucson).
35. Rolland, J. P. & Barrett, H. H. (1992) *J. Opt. Soc. Am. A Opt. Image Sci.* **9**, 649–658.
36. Rolland, J. P. Y. (1990) Ph.D. dissertation (Univ. of Arizona, Tucson).
37. Barrett, H. H., Rolland, J. P., Wagner, R. F. & Myers, K. J. (1989) *Proc. SPIE Int. Soc. Opt. Eng.* **1090**, 176–182.
38. Burgess, A. E., *J. Opt. Soc. Am. A. Opt. Image Sci.*, in press.