Proc. of SPIE Vol. 1898, Medical Imaging 1993: Image Processing, ed. M H Loew (Sep 1993) Copyright SPIE

Rayleigh Task Performance In Tomographic Reconstructions: Comparison of Human and Machine Performance

Kyle J. Myers[+], Robert F. Wagner[+], and Kenneth M. Hanson[*]

[+]Center for Devices & Radiological Health/FDA, HFZ-142, Rockville MD 20857
[*]Los Alamos National Laboratory, MS P940, Los Alamos NM 87545

ABSTRACT

We have previously described how imaging systems and image reconstruction algorithms can be evaluated based on the ability of machine and human observers to perform a binary-discrimination task using the resulting images.[1-4] Machine observers used in these investigations have been based on approximations to the ideal observer of Bayesian statistical decision theory. The present work is an evaluation of tomographic images reconstructed from a small number of views using the Cambridge Maximum Entropy software, MEMSYS 3.[5] We compare the performance of machine and human viewers for the Rayleigh resolution task. Our results indicate that for both humans and machines a broad latitude exists in the choice of the parameter α that determines the smoothness of the reconstructions. We find human efficiency relative to the best machine observer to be approximately constant across the range of α values studied. The close correspondence between human and machine performance that we have now obtained over a variety of tasks indicates that our evaluation of imaging systems based on machine observers has relevance when the images are intended for human use.

1. INTRODUCTION

It has long been recognized that the assessment of medical imaging systems is task dependent. It has also been recognized that the study of task performance may be expensive and time consuming because of the cost associated with having clinicians or other trained observers participate in the study, the need for a sufficient number of images to obtain statistical significance in the results, and the need for "ground truth" against which to judge the performance of the task. These considerations have led to the study of task performance by machine observers using simulated images. The most highly regarded machine observers are those based on the optimal observers of Bayesian statistical decision theory, e.g., those based on the likelihood function.[6] The question of the comparative performance of such optimal observers--or attempts to approximate them in machine implementations--vis-a-vis the performance of the human observer then arises naturally.

 0-8194-1131-0/93/$6.(

The images in this particular study are obtained from reconstructions derived from simulations of limited-angle two-dimensional tomography. The reconstruction method used is based on the maximum a posteriori (MAP) method of image estimation[7] where the prior probability distribution on the reconstructed image is the so-called entropic prior.[8] The particular version of the reconstruction algorithm used here is from the Cambridge school of Gull and Skilling and is named MEMSYS 3.[5] The assessment of the images proceeds according to the paradigm presented by Hanson:[1] A large number of images are generated according to a Monte Carlo technique; a binary task is specified and performed by either a machine or a human observer; and the performance is scored according to either the method of the receiver operating characteristic (ROC) curve[9,10] or the method of the two-alternative forced-choice (2AFC).[9]

Last year we reported on the evaluation of MEMSYS 3 reconstructions through the comparison of humans and machines for the task of disk detection. We showed that the human and the various machine observers were all sensitive to the reconstruction parameter α, which controls the stopping rule for the algorithm and effectively determines the smoothness of the reconstruction. For both machine and human observers, we found disk detection performance to be maximum at the lowest values of α studied, so that the reconstruction approached the maximum-likelihood solution with a positivity constraint.

In varying the parameter α, the spatial frequency content in the MEMSYS 3 reconstructions is varied. Therefore, a task with different spatial frequency requirements might yield different human and/or machine performance functions with α. To investigate this possibility, the current work extends our investigations of MEMSYS 3 reconstructions to the comparison of human and machine observer performance for the Rayleigh discrimination task.

2. THE SCENE AND THE DATA

The object class consists of a set of 10 scenes. Each scene contains 8 Gaussian doublets and 8 Gaussian bars, each randomly placed and randomly oriented in a circle of reconstruction inscribed in a 128x128 pixel array. The binary objects are pairs of points separated by 6 pixels and convolved with a 2D symmetric Gaussian function with a FWHM of 4 pixels. The bars are line segments 10.4 pixels long convolved with the same Gaussian used to create the binaries. The bar length and object amplitude were chosen to minimize the mean-square difference between the objects. An example scene taken from the ensemble is shown in Figure 1.

The data set consists of just 8 views, equally spaced over 180°, and parallel projections each containing 128 samples that include additive, zero-mean Gaussian noise with a standard deviation equal to one. The noise in the data is pre-smoothed prior to reconstruction by a triangular window with a FWHM of 3 pixels, reducing the rms noise level by a factor of 0.484. The object characteristics and noise variance were chosen to give signal-to-noise ratios that render the task neither too trivial nor too difficult, so that human and machine performance can be measured with a good degree of reliability.

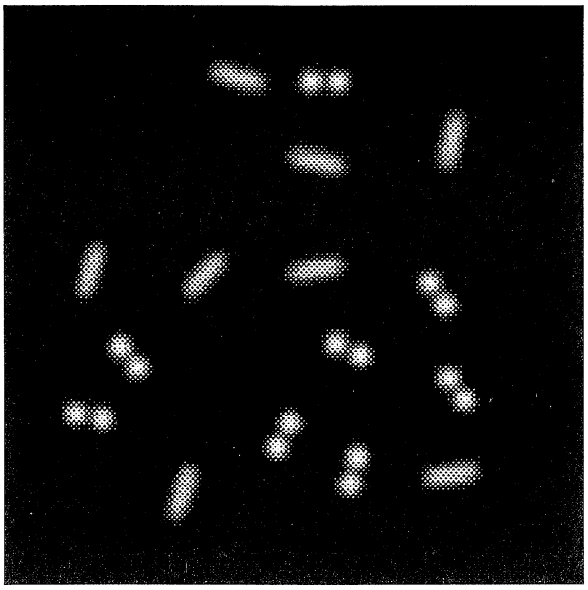


Figure 1. Sample scene containing 8 bar and 8 binary objects on a zero background. The circle of reconstruction is inscribed in a 128x128 pixel array.

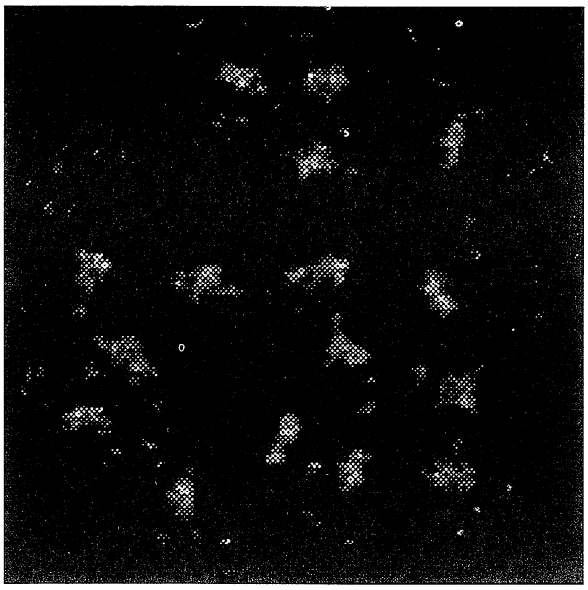
3. THE RECONSTRUCTION ALGORITHM

In previous proceedings we have described the reconstruction algorithm in detail.[3,4] Briefly, the maximum-entropy algorithm investigated here is a member of a family of MAP techniques for image estimation or reconstruction.[7] The particular version of the reconstruction algorithm used here was developed by Gull and Skilling and is named MEMSYS 3.[5] In effect the algorithm minimizes the expression

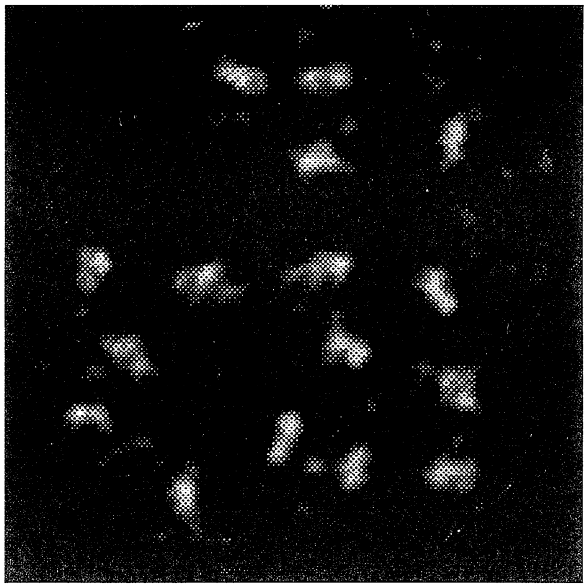
$$\chi^2/2 - \alpha S$$

where χ^2 is chi-squared, the exponent in the likelihood function that expresses the probability of the data given the object under the assumption of Gaussian additive noise.[6] The term $-\alpha S$ derives from the exponent of the entropic prior probability distribution on the reconstruction.[8] The parameter α selects one possible member of an infinite family of entropic priors; the smaller its value, the less one enforces the prior distribution, and the closer one approaches the minimum-chi-squared (or maximum likelihood) solution, while still retaining a positivity constraint through the entropy term. As α increases, the image becomes increasingly smooth, approaching the default of a uniform grey picture with mean level determined by the average intensity in the data set.

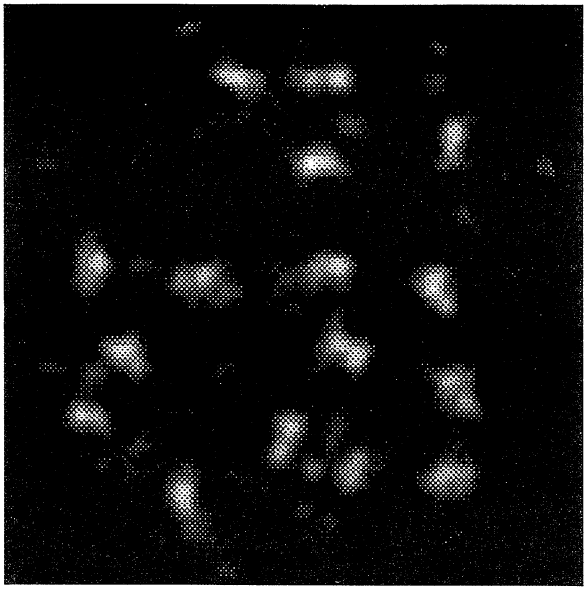
Reconstructions of the scene in Figure 1 are presented in Figure 2 for several values of α, including two values given the special labels "historic" and "classic." The historic solution is determined by setting α such that the value of chi-squared for the reconstruction is equal to N, the number of independent measurements in the data set. The classic algorithm determines α, and thereby the final value of χ^2, from the data itself. A further discussion of these particular choices for α is given in Wagner et al.[4] The MEMSYS 3 software also allows the user to specify an arbitrary ("ad hoc") value of the final or aimed-for value of chi-squared. In all cases, α is initialized at a very large value and gradually reduced until the desired value of chi-squared is reached. In effect, the algorithm is



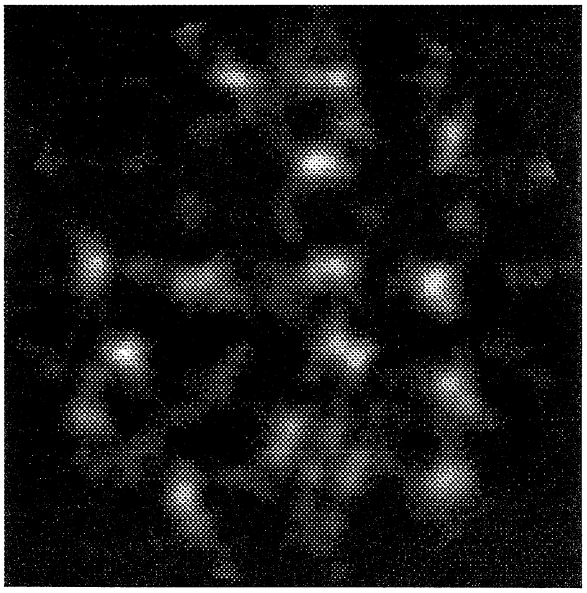
(a) α = .05, rms residual = .22



(b) α=.6 (classic), rms residual = .26



(c) α=4.03 (historic), rms residual = .5



(d) α=19.51, rms residual = 1.01

Figure 2. Sample reconstructions of the scene in Figure 1 showing increased image smoothness as α increases. The historic run results in a final χ^2=1024, the number of measurements. The classic algorithm attempts to find an optimal value for χ^2 that depends on the quality of the data.

terminated by a "stopping rule," which renders the image smoother than that offered by the ML solution where the algorithm runs to "completion" (α=0). It can be seen from the figure that the value of α dramatically affects the subjective appearance of the reconstruction, and only an objective evaluation of the ability of an observer to discriminate between the bars and binaries will reveal whether a particular choice for α is preferred.

4. MACHINE OBSERVERS

The machine decision functions are various approximations to decision functions that arise naturally in the study of Bayesian statistical decision theory. A list follows:

(a) The exact expression for the log of the posterior probability of each hypothesis given the data, p(f|g).[3,4] This function--consisting of the product of the likelihood p(g|f) and the exact expression for the entropy prior p(f)--is evaluated under the two hypotheses (binary present and bar present). The difference between the two evaluations at each location is the decision variable for that test region.

(b) The log of the posterior probability function, as in (a), but using a quadratic approximation obtained by expanding the expression for the log posterior probability in a Taylor series about the maximum (the reconstruction).[3,5] (Recall that quadratic in the log probability density is equivalent to Gaussian in the probability density.) Again, this calculation is done under two hypotheses (bar present and binary present) and the difference forms the test statistic.

(c) The mean-square difference between the reconstruction and the expected object. This difference is calculated for each of the expected objects (bar and binary) and the difference between the two calculations forms the test statistic. This decision strategy is approximately equivalent to the non-prewhitening matched filter.

Although the objects were randomly oriented in the original scene, the machine reader was given the orientation of each object under test. For each of the decision functions listed above, the following describes the decision-making procedure for the algorithmic observer. The decision function is applied to 80 subregions in the reconstructions that contain bar objects (known and extracted by the investigator to form the H_1 test images) and the decision-function output is recorded. The decision function is also applied to 80 regions in the reconstructions that contain binaries (known and extracted by the investigator to form the H_1 test images) and the decision-function output is recorded. The decision-function outputs are histogrammed separately for the known bar and binary locations. Then, by the well-known technique of varying the decision-function threshold, the receiver operating characteristic (ROC) curve is generated.[1] The area under the ROC curve is measured and the summary measure d_a is derived from an inverse error function.[11] This measure is the figure of merit used for evaluating the machine observers.

5. HUMAN OBSERVERS

Human observer performance was measured using approximately the same sub-regions from the reconstructions containing binary and bar objects as were available to the machine observer. (The machine used a slightly smaller region for the binary objects than the bar objects, but to eliminate this possible cue for the humans, the same region size was displayed for both objects during the human studies.) Each binary and bar object was extracted and centered in a square 35 pixels on a side. These sub-regions were then bilinearly interpolated twice to form images that were 140 pixels on a side. Because the machine reader had prior knowledge of the object orientation, the objects were rotated prior to display for the human observers such that each object was displayed with its major axis horizontal. These images were presented to the observer in pairs: one member of the pair from the bar class, and one member from the binary class. The side (left/right) containing the binary image was selected randomly. Also displayed were samples of a bar and binary object from the original scene, so that the observer had full knowledge of each object's size, shape, etc. This is the usual two-alternative forced-choice (2AFC) paradigm.[9] Feedback on the correctness of individual choices was given to the observer after each decision. The choices of the observer were recorded, and his or her percentage correct score was calculated. This percentage correct corresponds to the area under the curve in the ROC paradigm;[9] the summary measure used in this case is also derived from an inverse error function and is often referred to as d', although it is also not uncommon to refer to it as d_a. This will be the figure of merit for evaluating the human observers.

6. RESULTS

We shall present our results as a function of the parameter α. This parameter was allowed to range from a low value of 0.05 to a high value of 20. In Figure 3, the figure of merit d_a is plotted for each of the machine observers described in Section 4. Arrows indicate the values of alpha corresponding to the so-called historic and classic solutions. As can be seen from the figure, the classic reconstructions have a smaller value of α (and hence a smaller χ^2) than the historic ones. For the historic run, χ^2=1024; the classic run gave χ^2=281. It can be seen from the figure that the decision variable based on the quadratic approximation to the log posterior probability fails catastrophically for small values of α. The same breakdown was observed in the detection task explored last year. The error bars, calculated based on the number of images studied and the resulting d', indicate that there is no significant difference in performance for decision variables based on the exact posterior probability and mean-squared difference over the range of α studied. While there are hints of diminished performance at the extremes, further study is required to determine whether any particular value of α is optimal for these observers. (The results are more significant than the error bars indicate, however, because the error bars were calculated based on uncorrelated data, when in reality the same data were used to form the reconstructions for each value of α, and both the machines and the humans viewed the same images.)

The results for two human observers are presented in Figure 4 with the performance curves for the two best machine observers from Figure 3. (Error bars would be similar to

those shown in Figure 3.) The human performance curves show trends similar to those seen for the machine observers. However, the humans suffer a constant performance penalty of approximately 30% relative to the machine observers. This constant performance lag is similar to that found in other psychophysical studies comparing machine and human observer performance for simple discrimination tasks. [12,13]

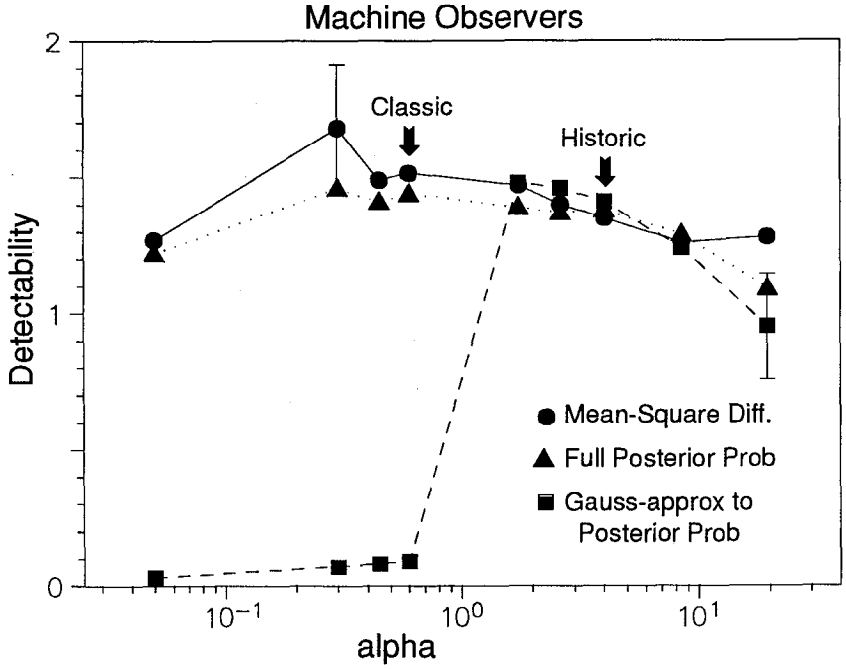


Figure 3. Rayleigh discrimination performance for the machine observers in Section 4 as a function of the parameter α that controls the degree of smoothness in the reconstructions.

Figure 4 also contains a plot of the mean-squared error (MSE) between the reconstruction and the original scene. Mean-squared error is used by many researchers in the field of image processing as a metric for image quality. In our previous investigation of disk detection, we found MSE to be a very poor predictor of human and machine performance. We have seen other instances where MSE can decrease greatly when a non-negativity constraint is invoked, compared to reconstructions where the constraint is not invoked, although observer performance may change very little. We see from Figure 4 that MSE shows a broad concave shape as a function of α for this task. For the set of reconstructions considered here, MSE changes little because we are considering a limited part of the reconstruction space -- the same algorithm is employed, with only the single parameter α being varied. Because it makes no reference to the task that the images are intended for, and for a host of other reasons,[14] MSE cannot generally be recommended as a measure of image quality.

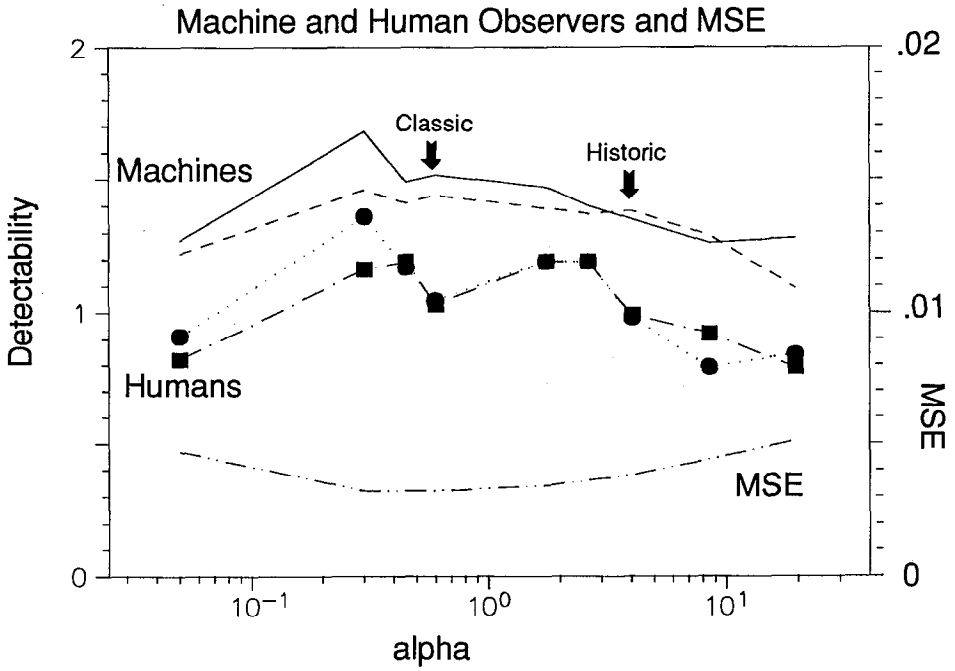


Figure 4. Performance of two human observers and the two best machine observers from the previous figure. Also shown is the mean-squared error (MSE) between the reconstructions and the original scenes.

7. SUMMARY

This study was undertaken to determine whether machine and human discrimination performance on the Rayleigh task would behave differently as α varies compared to the performance with α we found for the task of disk detection. In our earlier investigations of the detection problem, we found that the performance of both human and machine readers was poorest at the largest values of α studied, with detectability increasing steadily as α decreased until a performance plateau was reached near the classic solution. We have found that the Rayleigh task gives generally quite different performance curves as a function of α than those found for the detection task. For the Rayleigh task there appears to be much broader latitude in the choice of α. While there are hints of performance degradation for the humans and machines at the highest and lowest values of α studied, further investigation is needed with larger numbers of images and a greater range of α values. It is certain that observer performance must drop when α is infinite, because in that limit the reconstruction is a uniform grey, and no discrimination between the bar and binary objects is possible. However, we do not know how or where a performance decline will occur as α approaches that limit. We also do not yet know how closely the maximum likelihood limit may be approached (while maintaining the positivity constraint inherent to the entropy prior) before numerical difficulties will be encountered during the reconstruction process or observer performance declines.

For both a detection and a discrimination task, we have found close correspondence between the performance of human and certain machine observers as the character of the reconstructions is varied through the adjustment of the parameter α. The close correspondence between the robust machine observers and the human observer performance indicates that the machine observers that we have been using in this and previous work are indeed relevant when the images are intended for human use.

8. ACKNOWLEDGEMENTS

The authors have enjoyed many helpful conversations with Harrison H. Barrett, David G. Brown, and Arthur E. Burgess in the course of this work. We are indebted to Robert J. Jennings for several generations of his 2AFC display software and to Bruce C. Danielson for computer systems support on countless occasions. This work was partially supported by the U.S. Department of Energy under contract number W-7405-ENG-36.

9. REFERENCES

1. K.M. Hanson, "Method of evaluating image-recovery algorithms based on task performance," J. Opt. Soc. Am. A **7**, 1294-1304 (1990).

2. K.J. Myers and K.M. Hanson, "Comparison of the algebraic reconstruction technique with the maximum entropy reconstruction technique for a variety of detection tasks," Proc. SPIE **1231**, 176-187 (1990).

3. K.J. Myers and K.M. Hanson, "Task performance based on the posterior probability of maximum-entropy reconstructions obtained with MEMSYS 3," Proc. SPIE **1443**, 172-182 (1991).

4. R.F. Wagner, K.J. Myers, and K.M. Hanson, "Task performance on constrained reconstructions: human observer performance compared with sub-optimal Bayesian performance," Proc. of the SPIE **1652**, 352-362 (1992).

5. S.F. Gull and J. Skilling, *Quantified Maximum Entropy - MEMSYS 3 Users' Manual*, Maximum Entropy Data Consultants Ltd., Royston, England (1989).

6. A.D. Whalen, Detection of Signals in Noise (Academic, New York, 1971).

7. K.M. Hanson, "Bayesian and related methods in image reconstruction from incomplete data," in *Image Recovery: Theory and Application*, Henry Stark, editor (Academic, Orlando, 1987).

8. S.F. Gull and J. Skilling, "Maximum entropy method in image processing," IEE Proc. **131(F)**, 646-659 (1984).

9. D.M. Green and J.A. Swets, *Signal Detection Theory and Psychophysics* (Robert E. Krieger, Huntington NY, 1974).

10. C.E. Metz, "ROC methodology in radiologic imaging," Invest. Radiol. **21**, 720-733 (1986).

11. A.J. Simpson and M.J. Fitter, "What is the best index of detectability?" Psych. Bull. **80**, 481-488 (1973).

12. A.E. Burgess, R.F. Wagner, R.J. Jennings, H.B. Barlow, "Efficiency of human visual signal discrimination," Science **214**, 93-94 (1981).

13. K.J. Myers, H.H. Barrett, M.C. Borgstrom, D.D. Patton, G.W. Seeley, "Effect of noise correlation on detectability of disk signals in medical imaging," J. Opt. Soc. Am. A **2**, 1752-1759 (1985).

14. H.H. Barrett, presentation given at the Image Processing Technical Group Meeting, Annual Meeting of the Optical Society of America, 20-25 Sept. 1992, Albuquerque, NM. (manuscript in preparation).