Proc. of SPIE Vol. 1652, Medical Imaging VI: Image Processing, ed. M H Loew (Jun 1992) Copyright SPIE

Task Performance On Constrained Reconstructions: Human Observer Performance Compared With Sub-Optimal Bayesian Performance

Robert F. Wagner[+], Kyle J. Myers[+], 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 on the basis of how well binary-discrimination tasks can be performed by a machine algorithm that "views" the reconstructions.[1-3] Algorithms used in these investigations have been based on approximations to the ideal observer of Bayesian statistical decision theory. The present work examines the performance of an extended family of such algorithmic observers viewing tomographic images reconstructed from a small number of views using the Cambridge Maximum Entropy software, MEMSYS 3. We investigate the effects on the performance of these observers due to varying the parameter α; this parameter controls the stopping point of the iterative reconstruction technique and effectively determines the smoothness of the reconstruction. For the detection task considered here, performance is maximum at the lowest values of α studied; these values are encountered as one moves toward the limit of maximum likelihood estimation while maintaining the positivity constraint intrinsic to entropic priors. A breakdown in the validity of a Gaussian approximation used by one of the machine algorithms (the posterior probability) was observed in this region. Measurements on human observers performing the same task show that they perform comparably to the best machine observers in the region of highest machine scores, i.e., smallest values of α. For increasing values of α, both human and machine observer performance degrade. The falloff in human performance is more rapid than that of the machine observer at the largest values of α (lowest performance) studied. This behavior is common to all such studies of the so-called psychometric function.

1. INTRODUCTION

It has been recognized for several decades that the assessment of medical images or medical imaging systems requires the specification of a task to be performed using the images. Systems that rank in a certain order for the task of detecting large low-contrast lesions may rank differently for the task of detecting fine detailed structure. It has also been

recognized that the study of task performance may be expensive and time consuming because of the need for "ground truth" against which to judge the performance of the task, and the need for a sufficient number of images to obtain statistical significance in the results. These considerations have led to the study of task performance by machine or algorithmic observers who love the work. The most highly regarded algorithmic observers are those based on the optimal observers of Bayesian statistical decision theory, e.g., those based on the likelihood function.[4] 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. This paper addresses this question in the context of several of our investigations of reconstruction methods reported earlier.[1-3]

In this work the images are obtained from reconstructions derived from simulations of limited-angle two-dimensional tomography, a methodology required by applications ranging from single-photon emission CT (SPECT) in radionuclide imaging through tomography or tomosynthesis of coronary artery images in radiography. The reconstruction method used is based on the maximum a posteriori (MAP) method of image estimation[5] where the prior probability distribution on the reconstructed image is the so-called entropic prior.[6] The particular version of the reconstruction algorithm used here is due to the Cambridge school of Gull and Skilling and is named MEMSYS 3.[7] 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[8,9] or the method of the two-alternative-forced-choice (2AFC).[8] We shall now present some details of our present work in which we investigate the optimal use of the reconstruction technique, the detailed performance--given a specified detection task--by a number of algorithmic observers derived from statistical decision theory, and the performance of human observers given the same task.

2. THE SCENE AND THE TASK

As in previous work, the object class consists of a set of 10 scenes, each containing 20 randomly placed, non-overlapping disks on a zero background. Ten of the disks are low-contrast (amplitude = 0.1) and 10 are high-contrast (amplitude = 1.0). They are all 8 pixels in diameter in an overall field of 128 pixels in diameter. An example taken from the ensemble is shown in Figure 1a. The task is the detection of the low-contrast disks. The high-contrast disks are placed in the object to challenge the effects of limited-angle sampling which give rise to object-dependent artifacts. Here we have used 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 two. 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.

3. THE RECONSTRUCTION ALGORITHM

The maximum-entropy algorithm investigated here is a member of a family of MAP techniques for image estimation or reconstruction.[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 scene under the assumption of Gaussian additive noise.[4] The term $-\alpha S$ derives from the exponent of the entropic prior probability distribution on the reconstruction.[6] Minimizing chi-squared is equivalent to finding the maximum likelihood (ML) reconstruction. Minimizing the term $-\alpha S$ is equivalent to maximizing the entropy S, which for practical purposes can be considered a measure of the degeneracy of the image, i.e., the number of ways the image could be formed with the same total energy, light intensity, silver halide grains, etc.[10]; a uniformly gray image achieves the unconstrained maximum entropy. Minimizing the overall sum is an attempt to find the "least committal image" consistent with the data (see Ref. 11 for this more axiomatic approach). The factor 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 ML solution. Several techniques for determining α have evolved over the last decade.

So-called "ad hoc" versions of maximum entropy were based on aiming for a solution that yielded a value of chi-squared equal to some value selected or "fixed" by the user. It was argued that "feasible" sets of solutions are those for which chi-squared is less than or equal to N, the number of independent measurements in the data set. This approach was motivated by the fact that the statistically expected value of a chi-squared distributed random variable with N degrees of freedom is N, and the reconstruction is then constrained to be, on average, within one standard deviation of the data (assuming Gaussian-distributed noise). Since many early authors set their aim at a value of chi-squared equal to N, this has been referred to as "historic" maximum entropy. (This discussion has followed Ref. 7.)

The more recent "Classic" MaxEnt determines α, and thereby also the final value of χ^2, from the data itself. Some motivation for this approach is provided in one version that is terminated when $\chi^2 + G = N$, where G is a measure of the goodness of the data.[7] This expression has been interpreted to indicate that only G "good" parameter measurements are expected to contribute to the reduction of the data misfit (χ^2) that occurs when a model is fitted to noisy data.[7,12] This is analogous to the result obtained when a calculation of a sample mean (or M sample parameters) reduces the degrees of freedom of the residual chi-squared random variable by one (or M). In the limit of $G = N$, chi-squared may be driven to zero, the center of the ML distribution, which is also the limiting case when α approaches zero. 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 is gradually reduced until the desired value of chi-squared is reached. In effect,

the algorithm is terminated by a "stopping rule," which renders the image smoother than that offered by the ML solution where the algorithm runs to "completion" ($\alpha = 0$).

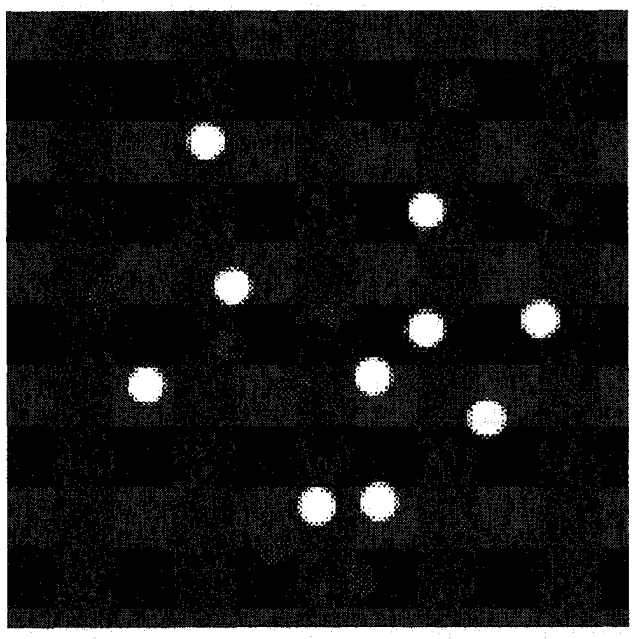


Fig. 1a

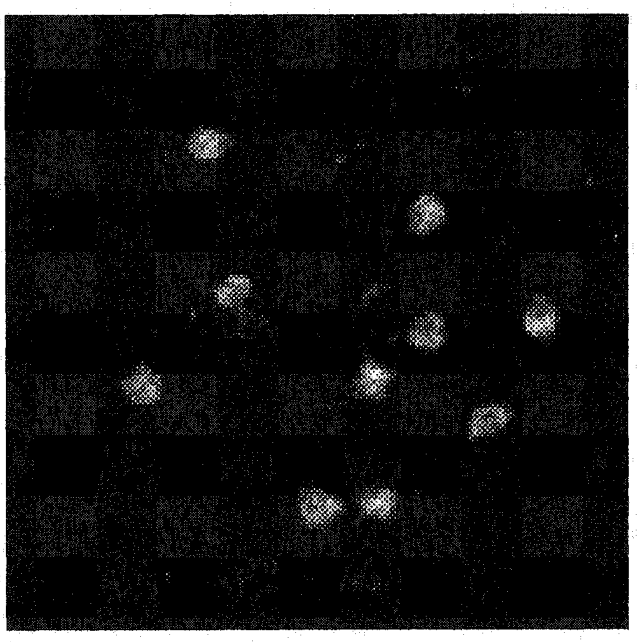


Fig. 1b

Figure 1. a) Sample scene containing 10 high-contrast disks and 10 low-contrast disks randomly placed on a zero background. b) Reconstruction of scene 1a with α=0.2.

A reconstruction of the object scene shown in Figure 1a, using MEMSYS 3 with alpha equal to 0.2, is given in Figure 1b. The high contrast disks are easily detected. Detectability of the low contrast disks has been adjusted to fall in a range such that the task is neither too trivial nor nearly impossible for the machine and the human observers, and can therefore be measured with a good degree of reliability. Figure 1b shows this to be plausible.

One of our major concerns with reconstruction algorithms in general, and the maximum entropy family in particular, is the fact that they are typically optimized in terms of minimizing some measure of error in the estimated image, without regard for the task to be performed by the reader of the image. In more technical terms, the question of optimal reconstruction when the estimated image is to be used for lesion detection or disease classification is not adressed. We shall return to this issue later in this paper.

4. ALGORITHMIC DECISION FUNCTIONS

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

(a) The exact expression for the log of the posterior probability of each hypothesis given the data, p(f|g). This function--consisting of the product of the pre-whitened likelihood p(g|f) and the exact expression for the entropy prior p(f)--is evaluated under the two hypotheses (disk present and disk absent). 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).[3,7] (Recall that quadratic in the log probability density is equivalent to Gaussian in the probability density.) Again, this calculation is done under two hypotheses (disk present and disk absent) and the difference forms the test statistic.

(c) The non-prewhitening matched filter (NPWMF) output, formed by summing all the pixels within the region of the expected disk signal. [13,14] This would be the likelihood of the signal in the case of Gaussian-distributed pixel values. However, the pixels in the reconstruction have some unknown (but non-Gaussian) distribution for a given object, and therefore this decision variable would be expected to be sub-optimal.

(d) The non-prewhitening matched filter, modified to include the background in an annular region centered on the location of the expected signal. The decision function is the difference between the activity in the central disk region and the estimated activity in the surrounding background. We shall refer to this as the disk contrast.

(e) The mean-squared-difference between the reconstruction and the expected object. This difference is calculated for each of the expected objects (disk present and absent) and the difference between the two calculations forms the test statistic.

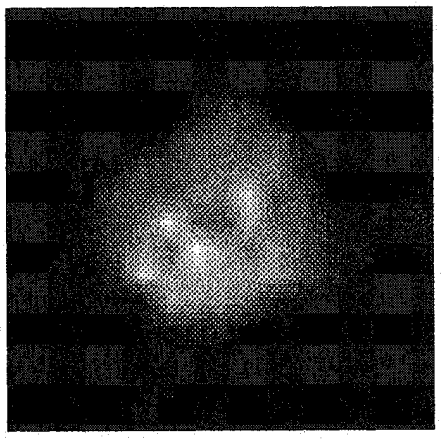
For each of the decision functions listed above, the following describes the decision-making procedure for the algorithmic observer. The decision function is applied to 100 subregions (16 pixels in diameter) in the reconstructions that contain background plus a disk (known and extracted by the investigator to form the H_1 test images) and the decision function output is recorded. The decision function is also applied to 100 regions in the reconstructions that contain only background (known and extracted by the investigator to form the H_0 test images) and the decision function output is recorded.

The decision function outputs are histogrammed separately for the known signal and the known background locations. Then, by the well-known technique of varying the decision function threshold for calling "lesion present," the receiver operating characteristic (ROC) curve may be generated. The area under the ROC curve is measured and the summary measure d_a is derived from an inverse error function.[15] This measure will be the figure of merit used for evaluating the machine or algorithmic observers.

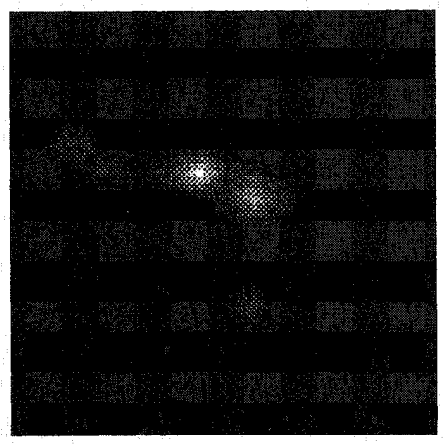
5. THE HUMAN OBSERVER

The human observers used the same 100 realizations of the signal-plus-background

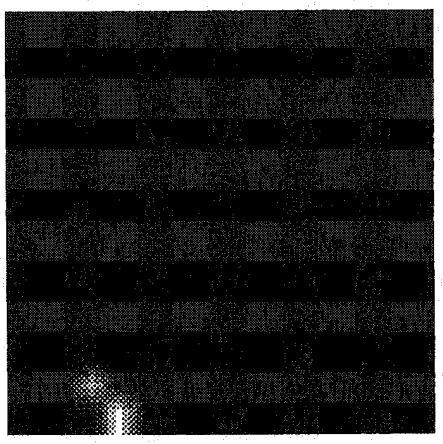
images and the same 100 realizations of the background-alone images. Each 16-pixel diameter test region was centered in a square of 16 pixels on a side, then bilinearly interpolated twice to form images that were 64 pixels on a side prior to display for the human observers. These images were presented to the observer in pairs: one member of the pair from the former class, and one member from the latter class. The side (left/right) containing the signal-present image is selected randomly. This is the usual two-alternative-forced-choice (2AFC) paradigm.[8] The choices of the observer are recorded, and his or her percentage correct score is calculated. This percentage correct corresponds to the area under the curve in the ROC paradigm[8]; 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.



(2a)



(2b)



(2c)

Figure 2. Local regions extracted from Fig. 1b: a) Sample region centered on a high-contrast disk; b) Sample region centered on a low-contrast disk. c) Sample noise-only test region.

The human observers viewed the images on a RAMTEK model 9460.[16] Originally the display was calibrated to obtain a dynamic range of about 100 using an SMPTE test pattern[17]; then the viewers were allowed to fine tune the one panel knob available to them until they felt most comfortable with the contrast and brightness. The human-observer experiments were conducted in a darkened room, after adaptation to the light level. Each observer had the benefit of about 700 practice trials before data collection began; although this is not a large number for such purposes, the observers' performance remained stable after the practice sessions. A typical set of disk-present/background-alone-present paired images is shown in Figure 2 ("disk present on the left" would be the correct choice in this example). Above the paired images is shown an example of a high-contrast disk. This indicates the region in which the low-contrast disk is expected to be found.

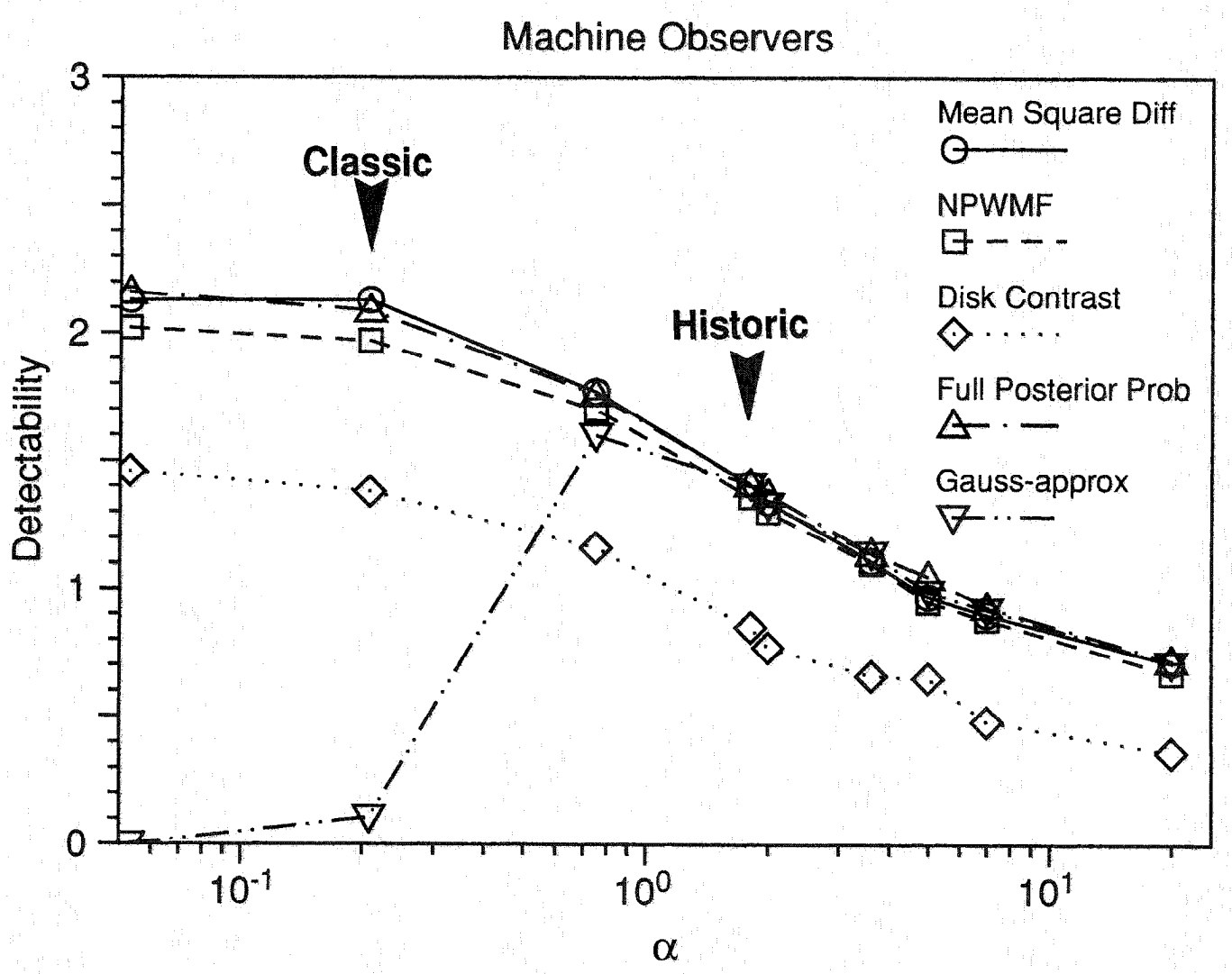


Figure 3. The detectability figure of merit d_a as a function of the parameter α for each of the machine or algorithmic observers described in Section 4. Arrows indicate the values of α corresponding to the "historic" and the "classic" maximum entropy solutions.

6. RESULTS FOR ALGORITHMIC OBSERVERS

We shall present our results as a function of the MaxEnt 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 algorithmic observers listed above. Generally the figure of merit is stable at a high value in the neighborhood of 2.0 at small values of α (approximating the ML solution) and falls off at high values of α (corresponding to extreme smoothing in the reconstruction). Arrows indicate the values of alpha corresponding to the so-called historic and classic MaxEnt solutions. As can be seen from the figure, the classic reconstructions have a smaller value of α than the historic ones. The classic reconstructions also resulted in values of χ^2 that are reduced compared to those obtained with the historic run. For the historic run, χ^2=1024; the classic run gave χ^2=473. All of the decision variables except the Gaussian approximation to the posterior probability function perform better using the classic MaxEnt reconstructions over the historic solution.

It can be seen from Figure 3 that the decision variable based on the quadratic approximation to the log posterior probability fails catastropically for small values of α. At small values of α, the majority of the image values are extremely small, even within the region of a low-contrast disk (see Figure 2). Near the limit of the reconstruction values going to zero, the positivity constraint inherent in the entropy prior causes the distribution of image values to be extremely non-Gaussian, and the Gaussian approximation falls apart.

7. RESULTS FOR HUMAN OBSERVERS

The results for two human observers are presented in Figure 4. They are seen to follow the better machine observer results to within the error bars at the best performance levels (lower values of alpha) and to fall off somewhat faster than the best machine results at the lower performance levels (higher values of alpha). The performance of human observers typically lags behind that of Bayesian-based observers at the lowest performance levels due, presumably, to the increase in uncertainty in his or her prior information induced by the highly degenerated versions of the signal, and this also degrades his or her internal consistency.

8. DISCUSSION

In this work we see an extension of trends seen in our earlier work, namely, that the task performance of algorithmic observers is indeed a function of the prior probability parameter alpha. The most obvious feature of the machine observer results, other than that based on the Gaussian approximation, is the stability of performance over a significant range as the value of α tends to lower values. We do not yet know how closely the maximum likelihood limit may be approached--maintaining the positivity constraint inherent to the entropy prior--before numerical difficulties will be encountered. Another obvious feature is the progression toward inferior performance as α is allowed to increase to higher values, effectively moving further away from the maximum likelihood solution. The performance curves for three out

of four of the robust algorithmic observers cluster at a level significantly higher than the performance curve for the machine observer referred to above as the disk contrast. The reason for the lower performance of the latter observer is not yet understood.

The close correspondence between the robust algorithmic observers and the human observer performance indicates that the degree of sharpness/smoothing represented by the variation over alpha is significant when the images are to be used for visual tasks. In fact, this correspondence 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. That is to say, the image assessment paradigm that we have been developing does in fact efficiently serve the purpose for which it was designed.

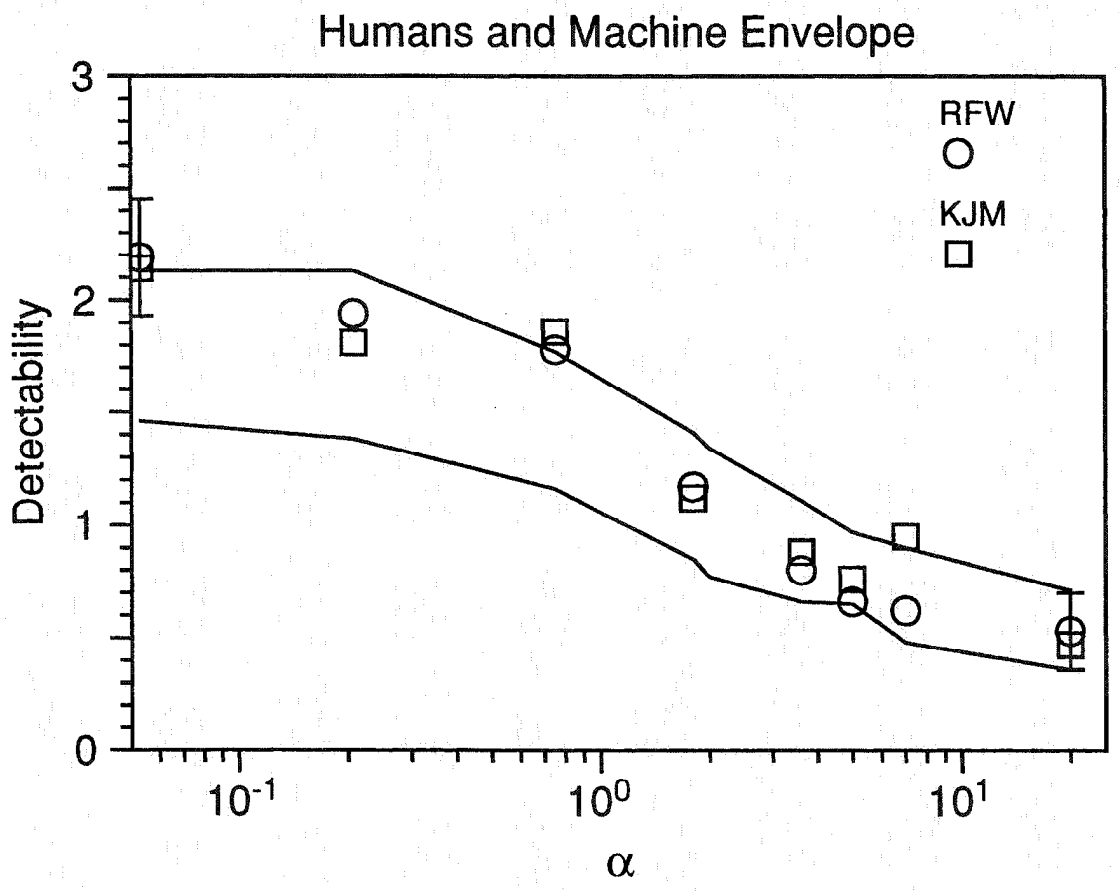


Figure 4. The detectability figure of merit as a function of the parameter α for two human observers (circles and squares), shown bracketed by the envelope of the detectability for the machine observers from Figure 3 (neglecting the Gaussian approximation to the posterior probability).

9. FUTURE ISSUES

It remains to be seen whether, for the detection task considered here, there is an optimal value of α closer to the ML limit. If not, this work would indicate that, for both algorithmic (excluding the Gaussian approximation to the log posterior probability) and human observers, the ML solution--with the positivity constraint inherent to the entropy prior--was optimal. Different conclusions might be drawn from the study of more detailed detection and discrimination tasks.

An outstanding question still remains to be seriously investigated by our community. How does one optimize an image estimation algorithm when the estimation step is to be followed by an image classification step. We are systematically studying this problem. It would be very gratifying if a general approach could be developed. At present, most optimizers of image reconstruction routines use a figure of merit related to the residual variance or rms pixel noise. Although such figures of merit can be related to certain detectability measures used here (at least for linear reconstruction schemes),[18] the relationship is neither direct nor necessarily monotonic. A more complete understanding of the steps that lead from estimation or reconstruction through a machine or human observer to a final detection or classification decision is required in order to optimize the procedure for the performance of the task for which the image was acquired.

10. ACKNOWLEDGEMENTS

The authors have enjoyed many helpful conversations with Harrison 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.

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

5. 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).

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

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

8. D.M. Green and J.A. Swets, *Signal Detection Theory and Psychophysics* (Robert E. Kreiger, Huntington NY, 1966).

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

10. H.C. Andrews and B.R. Hunt, *Digital Image Restoration* (Prentice-Hall, Englewood Cliffs NJ, 1977).

11. J. Skilling, "Classic maximum entropy," in *Maximum Entropy and Bayesian Methods in Science and Engineering, Vol 1, Foundations,* ed. G.J. Erickson and C.R. Smith (Kluwer, 1988).

12. D.J.C. MacKay, "Bayesian interpolation," [Submitted to] Neural Computation (1991).

13. R.F. Wagner, and D.G. Brown, "Unified SNR analysis of medical imaging systems," Phys. Med. Biol. **30**, 489-518 (1985).

14. K.J. Myers, J.P. Rolland, H.H. Barrett, and R.F. Wagner, "Aperture optimization for emission imaging: effect of a spatially varying background," J. Opt. Soc. Am. A **7**, 1279-1293 (1990).

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

16. Ramtek Corporation, 1525 Atteberry Lane, San Jose CA 95131.

17. SMPTE Medical Imaging Test Pattern, Standard RP-133, Society of Motion Picture and Television Engineers, 862 Scarsdale Ave, Scarsdale NY 10583.

18. H.H. Barrett, "Objective assessment of image quality: effects of quantum noise and object variability," J. Opt. Soc. Am. A **7**, 1266- 1278 (1990).