Performance-based assessment of reconstructed images

Kenneth M. Hanson

Los Alamos National Laboratory, MS B283
Los Alamos, New Mexico 87545 USA

ABSTRACT

During the early 90s, I engaged in a productive and enjoyable collaboration with Robert Wagner and his colleague, Kyle Myers. We explored the ramifications of the principle that the quality of an image should be assessed on the basis of how well it facilitates the performance of appropriate visual tasks. We applied this principle to algorithms used to reconstruct scenes from incomplete and/or noisy projection data. For binary visual tasks, we used both the conventional disk detection and a new challenging task, inspired by the Rayleigh resolution criterion, of deciding whether an object was a blurred version of two dots or a bar. The results of human and machine observer tests were summarized with the detectability index based on the area under the ROC curve. We investigated a variety of reconstruction algorithms, including ART, with and without a nonnegativity constraint, and the MEMSYS3 algorithm. We concluded that the performance of the Raleigh task was optimized when the strength of the prior was near MEMSYS's default "classic" value for both human and machine observers. A notable result was that the most-often-used metric of rms error in the reconstruction was not necessarily indicative of the value of a reconstructed image for the purpose of performing visual tasks.

Keywords: tomographic reconstruction, assessment of image quality, Rayleigh discrimination task, ROC analysis, human observer, machine observer, entropic prior, ART reconstruction algorithm, MEMSYS, Robert Wagner

1. INTRODUCTION

I met Robert F. Wagner shortly after I began my career in medical imaging. We bumped into each other in the fall of 1976 when I attended my first conference on image science, which was held in Toronto and organized by Chris Dainty and Rodney Shaw. Bob and I immediately recognized that we viewed the world similarly, probably because of our common backgrounds in physics. Bob had recently begun working in medical imaging at the FDA's Center for Radiological Health and had started to come to grips with some of the fundamental issues of x-ray radiography. I was eager to learn about the ideas he was developing, which concerned how to characterize the limitations of quantum-limited imaging. The cornerstone of his approach was NEQ (Noise Equivalent Quanta), which had been highlighted by Dainty and Shaw in their recent book, Image Science,[1] which dealt with the technology of photographic processes.

Immediately after the Toronto meeting, Bob and I traveled together to Ottawa to attend the International Conference on Medical Imaging. We heard David Chesler[2]talk about the negative correlations in the noise in CT images, a talk that turned out to have a strong influence on both of our careers. Interestly, from signal-detection theory, these correlations should affect the performance of visual tasks, for example, the interpretation of CT scans by radiologists. We proceeded in slightly different directions but kept in close touch with each other.

Bob wrote several papers about the ultimate limitations of radiography. He focused on the implications of signal detection theory for the diagnostic capabilities of radiographic systems.[3] I studied the noise power spectra in CT scans and their consequences for the performance of visual tasks, culminating in Ref. 4. We worked together on a few papers in that era.[5,6] It was in that period that Bob and I got to know Art Burgess, another ex-physicist, who was doing some remarkable human-vision experiments[7] that explored how well humans could perform visual tasks relative to an optimum machine observer.

Further author information – E-mail: kmh@lanl.gov; WWW: http://www.lanl.gov/home/kmh/; Telephone: +1 505 662 4370

Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment,
edited by Berkman Sahiner, David J. Manning, Proc. of SPIE Vol. 7263, 72630G
© 2009 SPIE · CCC code: 1605-7422/09/$18 · doi: 10.1117/12.817793

For nearly two decades, I visited the FDA one week each year so I could discuss with Bob various topics in medical imaging. Our interactions were both fun and thought provoking. Bob was my mentor and colleague; we reasoned together to try to reach a consensus in our views of medical imaging. Kyle Myers joined Bob's group in 1987, after which she became an integral part of our ongoing discussions.

Soon after Kyle joined the FDA, I began a very productive and enjoyable collaboration with Bob and Kyle, which I describe below.

In those days, and even now, the standard metric used to characterize the value of a reconstruction or image-restoration algorithm has been the rms error, the root-mean-squared difference between an original scene and the algorithm's output. Many of us in the SPIE medical-imaging community were unconvinced of the value of this metric. We believed the most appropriate means for assessing the value of medical images was in terms of how well a diagnosis could be performed using them. Wagner, Myers, and I (with some help from our friends) set out to underscore this principle by applying it to the problem of evaluating different algorithms for reconstructing a scene from incomplete and/or noisy projection data.

In our collaboration, we explored a variety of issues, including reconstruction algorithms, optimal machine-observer strategy, and more complex visual tasks. We considered the problem of choosing the strength of the entropic prior in the MEMSYS3 algorithm.[8] We devised the Rayleigh discrimination task, in which one must decide whether an object was a blurred version of a pair of dots or a bar. The results of numerous randomized tests were summarized with an ROC (Receiver Operating Characteristic) analysis. We employed both human and machine observers. The latter included various measures for comparing the reconstructed region surrounding each object to competing alternative signals, as well as neural networks. We eventually drew into our project a number of collaborators including David Brown, Mary Pastel (Anderson), Art Burgess, Harry Barrett, and Jannick Rolland.

2. DISK DETECTION

Around 1986, with ample encouragement from Bob Wagner, I began a project to demonstrate the use of task performance to determine whether the nonlinear constraint of nonnegativity was worthwhile when reconstructing from incomplete data scenes with a background close to zero.[9, 10] The outcome of the study[11] was that the nonnegativity constraint was beneficial when the projection measurements are incomplete, but not when they are complete and noisy. It also demonstrated that the most-often-used metric of rms error was not always indicative of the value of an image reconstruction for the purpose of performing visual tasks.

My approach to measuring task performance was inspired by the human-vision studies done by Art Burgess.[7] Perception studies in those days were most often based on detecting circular disks in a scene. Burgess employed the binary-decision version of that task, in which the decision is whether the disk is present or not in a displayed scene. I used the same task to test the performance of a machine observer of reconstructed scenes.

Figure 1 shows the kind of scene used to test the reconstruction algorithms. It contains ten high-contrast disks (with amplitude 1), placed randomly within the circle of reconstruction, but with a little space between them. The purpose of these disks is to generate streak artifacts. Their random placement leads to randomized artifacts. Ten low-contrast disks (with amplitude 0.1) are added to the scene, again, with random placement. The visual task to be performed is to detect these disks.

For each scene generated, its projections were calculated according to the measurement geometry being studied. For some scenarios, random noise was added. Then the Algebraic Reconstruction Technique (ART) algorithm[12] was used to reconstruct the original scene. ART is an iterative algorithm that converges quite rapidly to a solution and with relaxation, can yield a least-squares solution. One advantage of ART is that a nonnegativity constraint can be imposed on the solution.

Figure 2 shows the ART reconstructions from 12 noiseless projections of the scene in Fig. 1, with and without a nonnegativity constraint. The streak artifacts that are so prevalent in Fig. 2a definitely interfere with the detection of the low-contrast disks. The nonnegativity constraint used in Fig. 2b drastically reduces the streaks. Still, the low-contrast disks are not easy to see. It is clear from these images that the question of whether the nonnegativity constraint is efficacious for this situation is inherently a probabilistic one; it can only be answered through a testing procedure that involves many realizations of this type of scene.

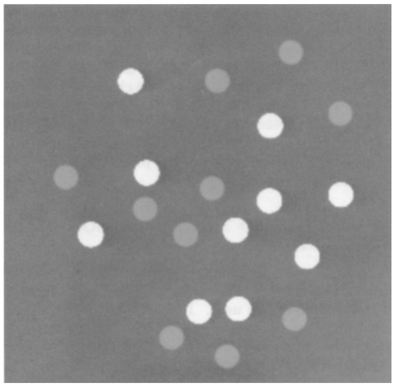

Figure 1. Scene containing ten high-contrast disks and ten low-contrast disks, placed in random locations. This type of scene was used to test reconstruction algorithms by evaluating how well the low-contrast disks can be detected.

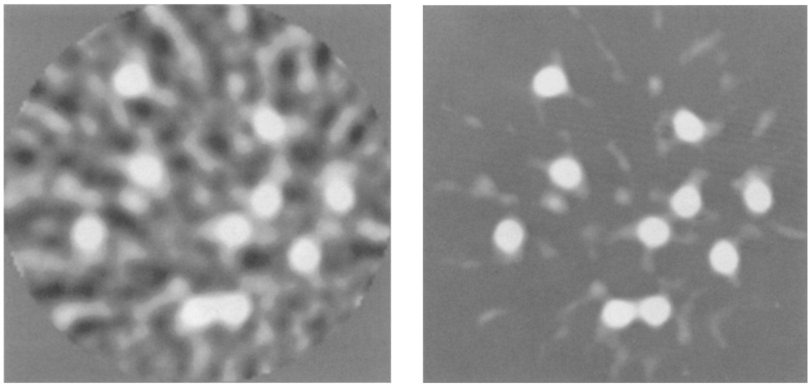

Figure 2. Reconstructions from 12 noiseless projections of the scene in Fig. 1 obtained using the ART algorithm a) without a nonnegativity constraint, on left, and b) with, on right.

For the disk detection task, the decision variable was the average image value over the known disk region. As well as the ten low-contrast disk regions in each reconstruction, 30 other circular regions of the same size were selected from random locations where no disk was located. Ten such reconstructions were used in all and an ROC curve was generated from the decision variable data. The area under the ROC curve was converted into the well-known detectability index d'_A.

For the above case of 12 noiseless projections, the result obtained from the ROC analysis was $d'_A = 0.90 \pm 0.15$ without the nonnegativity constraint, and $d'_A = 2.09 \pm 0.22$ with the constraint. It is clear that the constraint improved the performance of this task in this data-taking scenario. For comparison, the rms errors for the reconstructed scenes were 0.109 and 0.074, without and with the nonnegativity constraint, respectively. So, in this case, the detectability indices followed the rms error.

For a different measurement scenario, Fig. 3 shows the ART reconstructions from 180 projections with added random noise, with a rms value of 8, which is ten times larger than the projection value of a single low-contrast disk. For the image size used in this study, 256 × 256 pixels, 180 projections comprise an essentially complete data set. These two reconstructions, with and without nonnegativity, seem somewhat similar in appearance to the preceding examples. However, the ROC analysis yielded a surprising result.

For this situation, the ROC analysis yielded detectability indices of $d'_A = 1.96 \pm 0.21$ without the constraint, and $d'_A = 1.99 \pm 0.21$ with the constraint. The nonnegativity constraint did not help improve the performance of this task in this data-taking scenario. There is still no good explanation for this unexpected result, as far as I know.

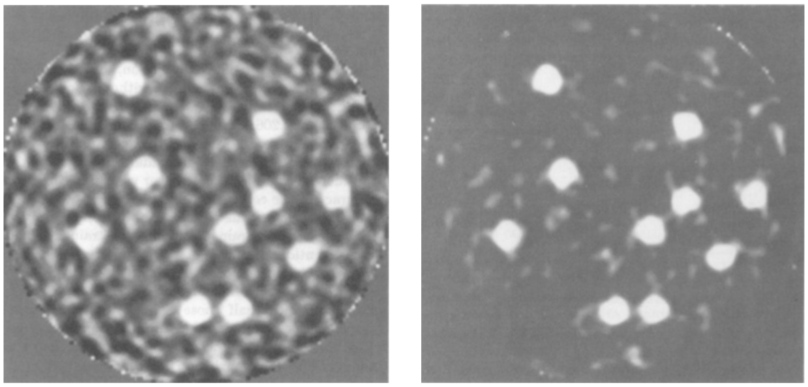

Figure 3. Reconstructions from 180 projections of the scene in Fig. 1, each with a modest amount of added noise. These reconstructions were obtained using the ART algorithm a) without a nonnegativity constraint, on left, and b) with, on right.

For comparison, the rms error, averaged over all test reconstructions, was 0.101 and 0.063 for reconstruction without and with the nonnegativity constraint, respectively. Thus, while the nonnegativity constraint reduced the rms error, disk detection did not improve.

This study demonstrated that rms error is **not** a good metric for assessing image quality. It showed that the nonnegativity constraint improved detectability of the disks when the data suffer from being incomplete but not when complete data are noisy.

Reference 13 showed how to use task performance to optimize a reconstruction algorithm, in that case, the relaxation parameters in the ART algorithm with a fixed number of iterations, with and without a nonnegativity constraint.

The preceding work on disk detection was carried out with counsel and encouragement from Bob Wagner and Kyle Myers. It proved sufficiently interesting that Kyle, and soon thereafter, Bob, joined me in a fruitful collaboration aimed at exploring different reconstruction algorithms,[14] visual tasks, and types of observers.

We were intrigued by the MEMSYS code,[8] which was at the time highly touted as being the best approach to reconstruction; its promoters claimed it was essentially handed down from the heavens. This code came from the Cambridge group of Steve Gull, John Skilling, and collaborators. It represented a Bayesian approach to estimation based on an entropic prior, the strength of which is controlled by the parameter α. In an early study, Kyle and I showed MEMSYS3 reconstructions yielded modestly better detectability than those from the ART algorithm in a situation in which the data were incomplete.[15] We ended up using MEMSYS3 in much of our later work.

In one study with MEMSYS3, we showed[16] that it was possible to use the Bayesian posterior probability to obtain better disk detectability than with average disk amplitude, used in many of our other studies. We also did human observer performance studies with MEMSYS3 reconstructions,[17,18] with the goal of determining the optimum value of α. For disk detection, the best detectability was obtained at very small values of α.[17]

3. RAYLEIGH DISCRIMINATION TASK

We were very interested in exploring possibly more appropriate visual tasks. We knew that higher-order tasks[19,20] relied on mid spatial frequencies, whereas disk detection relied mostly on low spatial frequencies. We devised a challenging binary visual task, inspired by the Rayleigh resolution criterion, of deciding whether an object was a blurred version of two dots or a bar.

Each test scene was created by randomly placing eight pairs of dots and eight bars, each with random orientation and subjected to a Gaussian blur, inside the circle of reconstruction. Random noise was added to the calculated projections and the scene was reconstructed. Because of the strong statistical nature of the decisions, hundreds of binary tests had to be performed to obtain accurate estimates of d'.

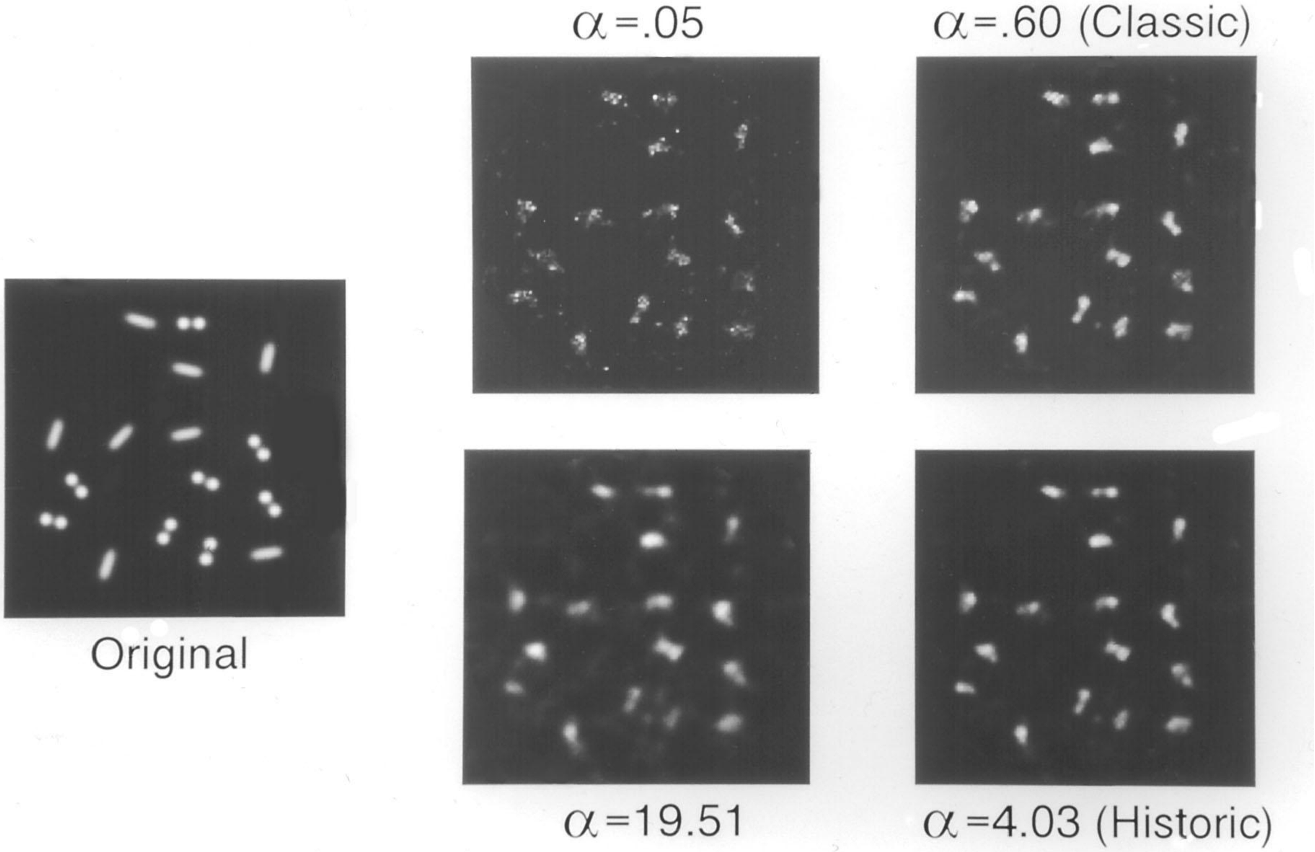

Figure 4. Reconstructions, on the right, from eight projections of the original Rayleigh discrimination scene on the left, each with a modest amount of noise added; from Ref. 21. These reconstructions were obtained using the MEMSYS3 algorithm, which inherently enforces nonnegativity, with varying strengths of the prior, controlled by the parameter α.

The complete specification of the two targets in the Rayleigh discrimination task[14] is as follows: For a 128×128-pixel image, the separation of the dots was 6 pixels and the lines were 10.4 pixels long. These structures were placed given a random orientation and random position in the scene, while making sure they do not come come too close to any other object. Each of these types of objects was convolved with a symmetric 2D Gaussian function with a FWHM of 4 pixels. The line length and amplitude was chosen to minimize the mean-square difference between the two types of obhects. The goal was to force the binary decision to be based on the fine details of the degraded image, not on gross features such as integrated intensity and overall size of the structure.

Figure 4 shows four reconstructions of the original scene, on the left, obtained with the MEMSYS3 algorithm for various values of the α parameter, which controls the strength of the entropic prior. As α approaches zero, the solution approaches the maximum-likelihood (minimum-χ^2) solution and the reconstruction tends to become spiky. As α increases, χ^2 increases and the reconstruction becomes smoother. In the limit that α goes to infinity, the reconstruction will become flat. There are two special values of α called out by MEMSYS. One is called the "historic" solution, when $\chi^2 = N$, where N is the number of measurements. It is well known that this condition results in "underfitting" the data in some situations. The other, called the "classic" solution is attained when $\chi^2 = N - G$, where G is a measure of the "goodness" of the measurements, which is estimated by the code.

In the situation represented in Fig. 4, the value of $\alpha = 0.6$ corresponds to the "classic" solution and $\alpha = 4.03$ to the "historic" one, whose reconstruction is somewhat smoother.

We employed both human and machine observers. The testing procedures for the humans (Bob and Kyle)

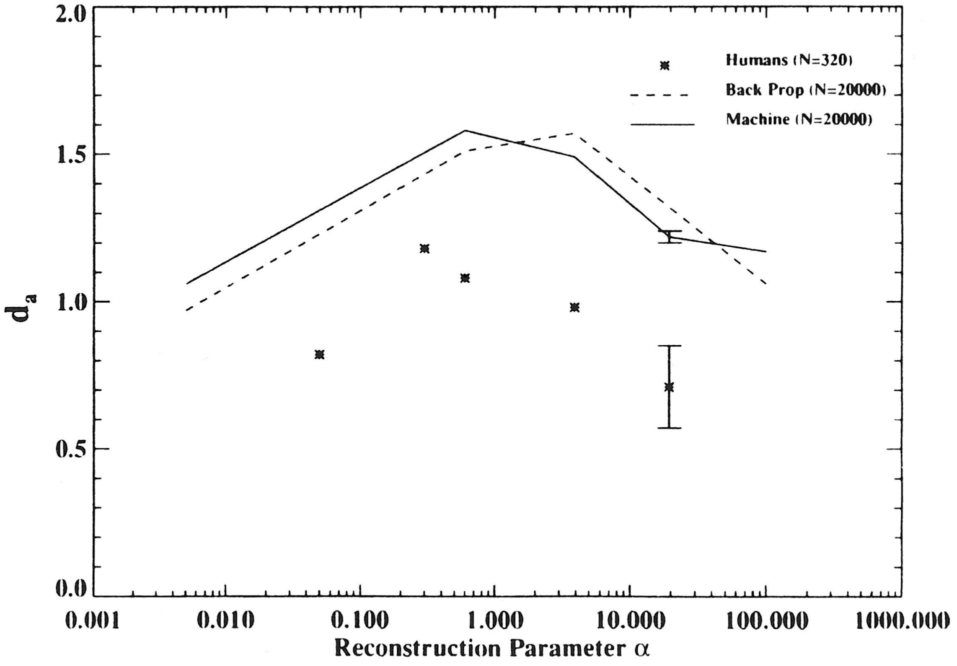

Figure 5. Plot of d'_A vs. α, the regularization parameter for MEMSYS3 reconstructions from eight noisy projections. The error bars shown represent the one-standard-deviation uncertainty in all data points of the same type.

involved a set up similar to what Burgess[7] used, namely a two-alternative forced-choice test for two sub-images, one containing two dots and the other a bar, displayed side by side. Each sub-image was randomly extracted from an appropriate region in a reconstruction. The sub-images were rotated to make sure they all had the same, known orientation. The observer had to try to select the sub-image with the two dots. In all, each human observer was shown 320 pairs of sub-images for each value of α and their d'_A's were determined from the fraction of correct choices they recorded.[7]

We tried numerous types of machine observers in our studies.[22–25] The best performances were obtained with two techniques that used the full two-dimensional sub-image, template matching and a neural network. The decision variable for template matching was based on calculating the mean-square difference between the reconstructed and known sub-images for the doublet and for the bar and taking their difference. The neural network consisted of two layers and was trained by backpropagation.

The machine observer testing involved 20000 sub-image pairs, 5000 of which were used for training. The advantage of machine observers is that the testing can be automated, which permits attaining estimates of detectability with arbitrary accuracy.

Figure 5 summarizes the testing results for human and machine observers for the Raleigh task.[21] The d'_A values for humans have roughly the same dependence on α as for the machines, but reduced by a factor of about 0.7. The square of this factor is typical for human observer efficiency compared to ideal observers.[7] These results seem to show that for the best performance of the Rayleigh discrimination task in this data-taking scenario, the value of α should be around unity, but with a broad optimal range that includes both the classic and historic solutions.

4. DISCUSSION

The main outcome of my collaboration with Bob Wagner and Kyle Myers was the demonstration of how one can use task performance to compare image processing and reconstruction algorithms. Our Rayleigh discrimination task appeared to provide a challenging and appropriate basis for assessing image usefulness. We established that the rms error in a reconstruction is **not** a realiable indicator of the usefulness of the image.

With each study, we felt we had advanced our understanding. However, much of our work begged for follow-up studies, which we could not find enough time to do. We hope others may find our results tantalizing and will pursue the quest.

On a personal level, my interaction with Bob Wagner involved innumerable delightful hours of discussion and camaraderie. My association with him led to a long and close friendship.

To pay tribute to Bob Wagner, I have created a memorial site (http://sites.google.com/site/robertfwagnermemorial), as well as an online photo album (http://picasa.google.com/kmhpix/BobWagner). I encourage anyone who would like to share memories or photos of Bob on these sites to contact me at kennethmhanson@msn.com.

ACKNOWLEDGEMENTS

The work reported here was done in collaboration with numerous colleagues, including Kyle Myers, Art Burgess, David Brown, Mary Pastel (Anderson), Harry Barrett, and Jannick Rolland. This work was supported in part by the United States Department of Energy under contract W-7405-ENG-36.

REFERENCES

[1] Dainty, J. C. and Shaw, R., [*Image Science*], Academic, London (1974).

[2] Reiderer, S. J., Pelc, N. J., and Chesler, D. A., "Statistical aspects of computed x-ray tomography," in [*45th Inter. Conf. Medical Physics*], (1976).

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

[4] Hanson, K. M., "Detectability in computed tomographic images," *Med. Phys.* **6**, 441–451 (1979).

[5] Sandrik, J. M., Wagner, R. F., and Hanson, K. M., "Radiographic screen-film noise power spectrum: calibration and intercomparison," *Appl. Opt.* **21**, 3597–3601 (1982).

[6] Wagner, R. F., Brown, D. G., Burgess, A. E., and Hanson, K. M., "The observer SNR penalty for reconstructions from projections," *Magn. Reson. Med.* **1**, 76–77 (1984).

[7] Burgess, A. E. and Ghandeharian, H., "Visual signal detection. I. Ability to use phase information," *J. Opt. Soc. Amer. A* **1**, 900–905 (1984).

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

[9] Hanson, K. M., "Evaluation of image-recovery algorithms on the basis of task performance," in [*Proc. 11eme Colloque sur le Traitement du Signal et des Images*], 547–550 (1987).

[10] Hanson, K. M., "Method to evaluate image-recovery algorithms based on task performance," in [*Medical Imaging II: Image Formation*], Schneider, R. H. and III, S. J. D., eds., *Proc. SPIE* **914**, 336–343, SPIE (1988).

[11] Hanson, K. M., "Method to evaluate image-recovery algorithms based on task performance," *J. Opt. Soc. Amer. A* **7**, 45–57 (1990).

[12] Gordon, R., Bender, R., and Herman, G., "Algebraic reconstruction techniques for three-dimensional electron microscopy and x-ray photography," *J. Theor. Biol.* **29**, 471–481 (1970).

[13] Hanson, K. M., "POPART - Performance OPtimized Algebraic Reconstruction Technique," in [*Visual Comm. and Image Processing*], Hsing, T. R., ed., *Proc. SPIE* **1001**, 318–325, SPIE (1988).

[14] Hanson, K. M. and Myers, K. J., "Rayleigh task performance as a method to evaluate image reconstruction algorithms," in [*Maximum Entropy and Bayesian Methods*], Grandy, W. T. and Schick, L. H., eds., 303–312, Kluwer Academic, Dordrecht (1991).

[15] Myers, K. J. and Hanson, K. M., "Comparison of the algebraic reconstruction technique with the maximum entropy reconstruction technique for a variety of detection tasks," in [*Image Formation*], Schneider, R. H., ed., *Proc. SPIE* **1231**, 176–187, SPIE (1990).

[16] Myers, K. J. and Hanson, K. M., "Task performance based on the posterior probability of maximum-entropy reconstructions with MEMSYS 3," in [*Image Physics*], Schneider, R. H., ed., *Proc. SPIE* **1443**, 172–182, SPIE (1991).

[17] Wagner, R. F., Myers, K. J., and Hanson, K. M., "Task performance on constrained reconstructions: Human observers compared with suboptimal Bayesian performance," in [*Image Processing*], Loew, M. H., ed., *Proc. SPIE* **1652**, 352–362, SPIE (1992).

[18] Myers, K. J., Wagner, R. F., and Hanson, K. M., "Binary task performance in images reconstructed with MEMSYS3: comparison of machine and human observers," in [*Maximum Entropy and Bayesian Methods*], Mohammad-Djafari, A. and Demoment, G., eds., 415–421, Kluwer Academic, Dordrecht (1993).

[19] Hanson, K. M., "Variations in task and the ideal observer," in [*Appl. of Optical Instru. in Medicine XI: Image Formation*], Fullerton, G. D., ed., *Proc. SPIE* **419**, 60–67, SPIE (1983).

[20] Wagner, R. F., K. J. Myers, G. D. Brown, M. J. T., and Burgess, A. E., "Higher order tasks: human vs. machine performance," in [*Medical Imaging III: Image Formation*], *Proc. SPIE* **1090**, 183–194, SPIE (1989).

[21] Wagner, R. F., Myers, K. J., Brown, D. G., Anderson, M. P., and Hanson, K. M., "Toward optimal human and algorithmic observer performance of detection and discrimination tasks on reconstruction from sparse data," in [*Maximum Entropy and Bayesian Methods*], Hanson, K. M. and Silver, R. N., eds., 211–220, Kluwer Academic, Dordrecht (1995).

[22] Wagner, R. F., Myers, K. J., and Hanson, K. M., "Rayleigh task performance in tomographic reconstructions: Comparison of human and machine performance," in [*Image Processing*], Loew, M. H., ed., *Proc. SPIE* **1898**, 628–637, SPIE (1993).

[23] Myers, K. J., Anderson, M. P., Brown, D. G., Wagner, R. F., and Hanson, K. M., "Neural network performance for binary discrimination tasks, part II: effect of task training, and feature pre-selection," in [*Medical Imaging: Image Processing*], Loew, M. H., ed., *Proc. SPIE* **2434**, 828–837, SPIE (1995).

[24] Myers, K. J., Wagner, R. F., Hanson, K. M., Barrett, H. H., and Rolland, J. P., "Human and quasi-Bayesian observers of quantum-, artifact-, and object-variability-limited images," in [*Image Perception*], Kundel, H. L., ed., *Proc. SPIE* **2166**, 180–190, SPIE (1994).

[25] Myers, K. J., Wagner, R. F., and Hanson, K. M., "Rayleigh task performance in tomographic reconstructions: comparison of human and machine performance," in [*Image Processing*], Loew, M. H., ed., *Proc. SPIE* **1898**, 628–637, SPIE (1993).