BINARY TASK PERFORMANCE ON IMAGES RECONSTRUCTED USING MEMSYS 3: COMPARISON OF MACHINE AND HUMAN OBSERVERS

K. J. Myers and R. F. Wagner
Center for Devices & Radiological Health/FDA
HFZ-142, Rockville MD 20857 USA

K. M. Hanson
Los Alamos National Laboratory, MS P940
Los Alamos, New Mexico 87545 USA
email: kmh@lanl.gov

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, 2]. The present work examines the performance of a family of algorithmic observers viewing tomographic images reconstructed using the Cambridge Maximum Entropy software, MEMSYS 3. We investigate the effects on the performance of these observers due to varying the parameter α, which controls the strength of the prior in the iterative reconstruction technique. 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, a behavior 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. 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 and/or observers to obtain statistical significance in the results. These considerations have led to the study of task performance by machine or algorithmic observers. The question of the comparative performance of such machine observers relative to the performance of the human observer then naturally arises.

In this work images are obtained from reconstructions derived from simulations of limited-angle two-dimensional tomography. 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 or the method of the two-alternative-forced-choice (2AFC) [3, 4].

A. Mohammad-Djafari and G. Demoments (eds.), Maximum Entropy and Bayesian Methods, 415–421.

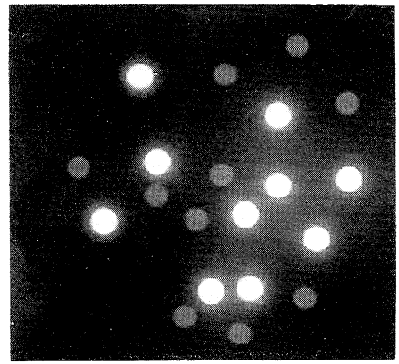

Figure 1: a) Sample scene containing 10 high-contrast disks and 10 low-contrast disks randomly placed on a zero background.

2. The Scene and the Task

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 1. The task is the detection of the low-contrast disks. Here we have used just 8 equally spaced views, and parallel projections each containing 128 samples that include additive, zero-mean Gaussian noise with a standard deviation of 2, which is about twice the peak projection value of the low-contrast disks. 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 reconstruction algorithm used here, named MEMSYS 3 [6], minimizes the expression

$$\frac{1}{2}\chi^2 - \alpha S \tag{1}$$

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, and $-\alpha S$ is the exponent of the entropic prior probability distribution on the reconstruction [5]. Minimizing chi-squared is equivalent to finding the maximum likelihood (ML) reconstruction. Minimizing $-\alpha S$ is equivalent to maximizing the entropy S, which can be considered a measure of the degeneracy of the image; a uniformly gray image achieves the unconstrained maximum entropy. Minimizing the expression in Eq. 1 amounts to finding the "least committal image" consistent with the data.

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. Since many early authors picked α so that chi-squared equaled the number of measurements, this has been referred to as "historic" maximum entropy. The more recent "classic" MaxEnt determines α from the data itself. 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. Reconstructions of the object scene shown in Figure 1 are given in Figure 2 for 4 values of α.

4. Algorithmic Decision Functions

The machine decision functions are various approximations to decision functions that arise in the study of Bayesian statistical decision theory:

(a) The difference in the log of the posterior probability for each hypothesis given the data, $p(\mathbf{f} \mid \mathbf{g})$.

(b) Same 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) [6, 2].

(c) The non-prewhitening matched filter (NPWMF) output, formed by summing all the pixels within the region of the expected signal [7, 8].

(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, referred to as the disk contrast, is the difference between the activity in the central disk region and the estimated activity in the surrounding disk.

(e) The difference in the mean-squared-difference between the reconstruction and the expected object calculated under each of the hypotheses (disk present and absent).

To determine each machine observer's figure of merit, the decision function is applied to 100 subregions (16 pixels in diameter) in the reconstructions that contain background plus a disk. The decision function is also applied to 100 regions in the reconstructions that contain only background. The decision function outputs are histogrammed separately for the known signal and the known background locations and the receiver operating characteristic (ROC) curve is then generated [3]. The figure of merit, d_a, is derived from the area under the ROC curve via an inverse error function.

5. The Human Observer

The human observers used the same 100 realizations of the signal-plus-background images and background-alone images. Each 16-pixel diameter test region was centered in a 16×16 square, then bilinearly interpolated twice to form 64×64 pixel images for display. These images were presented to the observer in pairs in the usual two-alternative-forced-choice (2AFC) paradigm [3]. The observer's percentage correct in a 2AFC experiment corresponds to the area under the curve in the ROC paradigm when the same images are used. Thus, the detectability figure of merit for the human observers is derived via an inverse error function from the percent correct, and can be compared directly to the machine d_a.

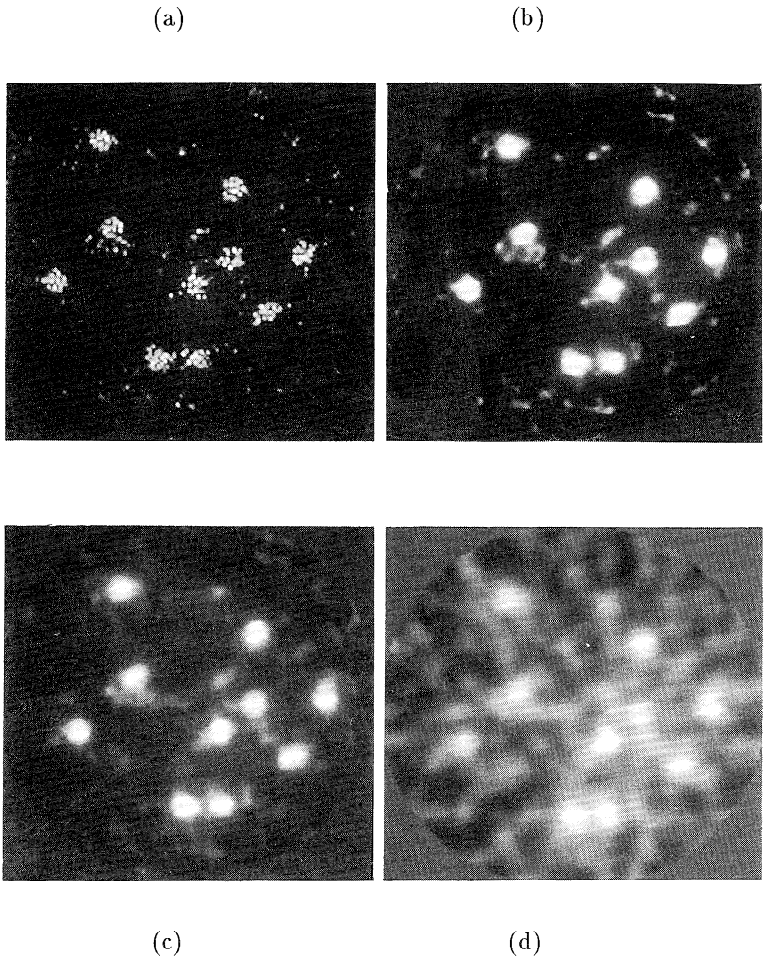

Figure 2: Reconstructions of the object scene in Figure 1 for values of α equal to a)0.002, b)0.2, c)1.8, d)20.

6. Results

In Figure 3, d_a is plotted for each of the algorithmic observers. We see that the performance of the algorithmic observers is a function of the parameter α. Generally the figure of merit is stable at small values of α and falls off at high values of α. Arrows indicate the values of α corresponding to the historic and classic MaxEnt solutions. As can be seen from the figure, the classic reconstructions have a smaller value of α (and hence χ^2) than the historic ones. For the historic run, α=01.8 and χ^2=1024; the classic run gave α=0.2 and χ^2=473. All of the decision variables except the Gaussian approximation to the posterior probability

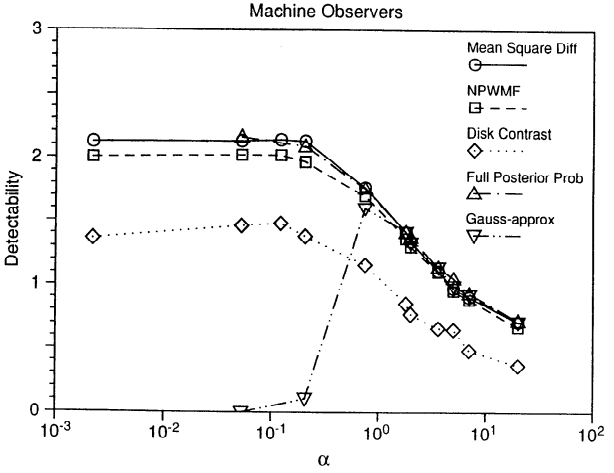

Figure 3: The detectability d_a as a function of the parameter α for each of the machine or algorithmic observers.

function perform better using the classic MaxEnt reconstructions over using 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 catastrophically for small values of α.

The results for two human observers are presented in Figure 4.

They are seen to follow the best machine observer results to within the error bars for

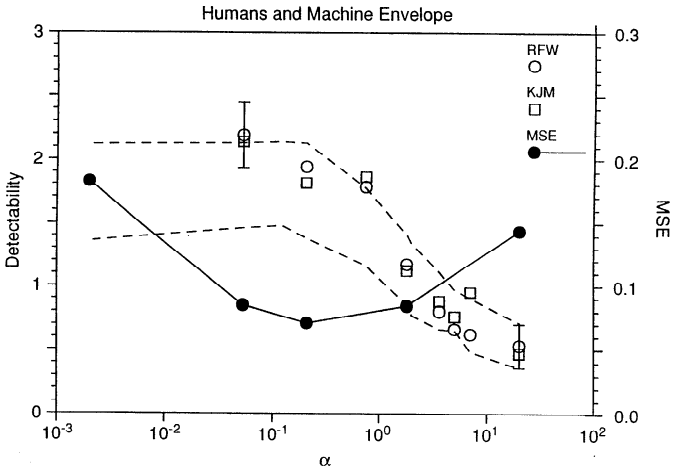

Figure 4: Detectability as a function of α for two human observers (circles and squares), bracketed by the envelope of the machine observer performance functions from Figure 3 (neglecting the Gaussian approximation to the posterior probability). The mean-squared error (MSE) between the original scene and the reconstructions is also given as a function of α.

lower values of α, and to fall off somewhat faster than the machine results as α increases. The close correspondence between the performance of the algorithmic observers and the human observers indicates that the degree of sharpness/smoothing represented by the variation over α is significant when the images are to be used for visual tasks, and that the machine observers we have studied are indeed relevant when the images are intended for human use. Also shown in Figure 4 is the mean-squared-error between the original scene and the reconstructions obtained for each value of α. It is clear from the figure that MSE is a poor predictor of performance for any of the machine or human observers.

7. Future Issues

This work indicates that, for algorithmic (excluding the Gaussian approximation to the log posterior probability) and human observers and a simple disk detection task, high detectability is found for values of α from about 0.2 all the way down to the ML limit–with the positivity constraint inherent to the entropy prior. Different conclusions might be drawn from the study of more detailed detection and discrimination tasks.

A general question for investigation is: How does one optimize an image reconstruction algorithm when that estimation step is to be followed by an image classification step? At present, most optimizers of image reconstruction routines use a figure of merit related to the MSE or rms pixel noise. Although such figures of merit can be related to certain detectability measures used here (at least for linear reconstruction schemes) [9], the relationship is neither direct nor necessarily monotonic. And, for this iterative reconstruction method, we have found that MSE does not predict human or machine detection performance. 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.

8. Acknowledgements

This work was partially supported by the U.S. Department of Energy under contract number W-7405-ENG-36.

References

[1] K. M. Hanson. Method to evaluate image-recovery algorithms based on task performance. *J. Opt. Soc. Amer.*, A7:45–57, 1990.

[2] 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.

[3] D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics.* Krieger, Huntington, NY, 1974.

[4] C. E. Metz. ROC methodology in radiologic imaging. *Invest. Radiol.*, 21:720–733, 1986.

[5] S. F. Gull and J. Skilling. Maximum entropy method in image processing. *IEE Proc*, 131(F):646–659, 1984.

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

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

[8] 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. Amer.*, A7:1279–1293, 1990.

[9] H. H. Barrett. Objective assessment of image quality: effects of quantum noise and object variability. *J. Opt. Soc. Amer.*, A7:1266–1278, 1990.