Method to evaluate image-recovery algorithms based on task performance

K. M. Hanson*
Los Alamos National Laboratory, MS P940
Los Alamos, NM 87545 USA

Abstract

A method for evaluating image-recovery algorithms is presented, which is based on the numerical assessment of how well a specified visual task may be performed using the reconstructed images. A Monte Carlo technique is used to simulate the complete imaging process including the generation of scenes appropriate to the desired application, subsequent data taking, image recovery, and performance of the stated task based on the final image. The use of a pseudo-random simulation process permits one to assess the response of an image-recovery algorithm to many different scenes. Nonlinear algorithms are readily evaluated. The usefulness of this method is demonstrated through a study of the algebraic reconstruction technique (ART), which reconstructs images from their projections. In the imaging situation studied, it is found that the use of the nonnegativity constraint in ART can dramatically increase the detectability of objects in some instances, especially when the data consist of a limited number of noiseless projections.

Introduction

For every indirect imaging application it is necessary to choose an image-recovery algorithm to obtain a final image. This choice becomes critically important when the available data are limited and/or are noisy. Several classes of measures have been employed in the past upon which to base image-recovery algorithms [1]. There are those based on the fidelity of the reconstructed images, such as the conventional measure of minimum rms difference between the reconstruction and the original image. Experience teaches us that this does not always seem to be correlated with the usefulness of images and so does not help one select an algorithm. There are measures based on how closely the estimated reconstruction reproduces the input data. The most popular of these, based on least-squares residual (or minimum chi-squared), is known to be ill-conditioned or even worse, ill-posed [1]. To better condition the problem, it is often proposed to constrain the least-squares objective in some way. Further, there are measures that combine the two previous ones, such as maximum *a posteriori* reconstruction, which attempts to balance the match to the data against the relationship of the reconstruction to the known ensemble probability distributions [2].

The fundamental tenet adopted in this paper is that the overall purpose of the imaging procedure is to provide certain specific information about the object or scene under investigation. Consequently, in the approach to algorithm evaluation presented here, an algorithm is to be judged on the basis of how well one can perform stated visual tasks using the reconstructed images.

Figure 1 displays the successive links in an imaging chain. It emphasizes that all the elements of the imaging system have an effect on the final interpretation of the scene and thus must all be considered in evaluating the effectiveness of the system. In actuality, the performance of an imaging task is a subtle process because there are many subsidiary paths along which information is passed. The influence of these implicit

*This work was supported by the United States Department of Energy under contract number W-7405-ENG-36 and by Thomson CGR.

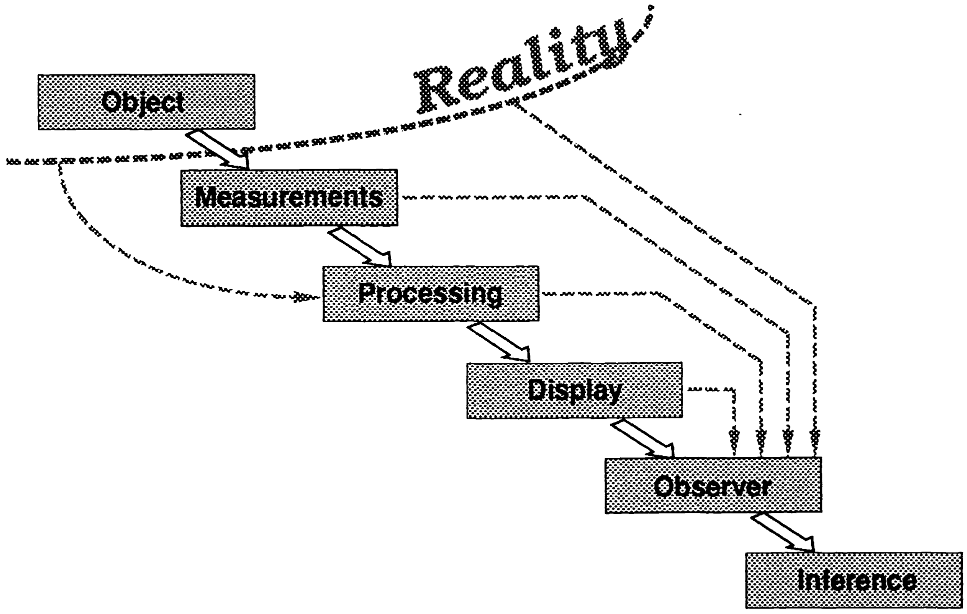

Figure 1: Diagram of the complete process involved in the performance of a visual task. The dashed lines indicate the paths taken by auxillary information, which **must** be available to the observer in order for him to make a decision about reality.

sources of prior information on the final inference is not well understood. Here, task performance will be evaluated under an explicit set of assumptions.

For linear imaging systems the effects of image noise on task performance can be predicted for a variety of simple tasks [3]. The same cannot be said of the effects of artifacts. The masking effects of measurement noise are truly random in nature. The random noise process results in each set of measurements being different, even when the scene being imaged does not change. Some kinds of artifacts appear as fixed patterns and do not often behave like stationary noise. However, those created by an insufficient number of measurements can manifest themselves as seemingly unpredictable irregularities that look like noise, but in a strict sense, they are not. These patterns are determined by the scene being imaged. Therefore, it is necessary to vary the scene in a realistic way to test how well an algorithm dispenses with artifacts. For example, the objects in the scene are normally randomly placed relative to the discretely sampled measurements as well as to the reconstruction grid. Both of these positionings might affect the reconstruction. Thus a single realization of a simple scene is completely inadequate to judge a reconstruction algorithm. It is necessary to obtain a statistically meaningful average of the response of an algorithm to many realizations of the ensemble of scenes with which it must cope. It is unclear whether or not such a global approach to task performance is amenable to theoretical treatment. The implied averaging over discrete samplings is difficult to handle analytically. Futhermore, it would be difficult to deal with nonlinear reconstruction or task performance algorithms. To overcome these deficiencies, the proposed method is based upon computer simulation of scenes appropriate to the desired application, subsequent data taking and analysis of the data. A Monte Carlo technique, one that employs pseudo-random numbers to generate its results, is used in this simulation because it can readily provide the above-noted variations within the ensemble. Furthermore, any new source of uncertainty can easily be incorporated into the simulation by simply adding randomness to the appropriate variable through the introduction of a pseudo-random number.

Method

The proposed method of evaluating image-recovery algorithms employs a Monte Carlo technique to simulate the entire imaging process from the beginning to the final task performance. To begin with, one randomly generates representative scenes and the corresponding sets of measurements. The specified tasks are then performed using the reconstructed scenes. Finally, the accuracy of the task performance is evaluated. The advantage of this numerical approach is that it readily handles complex imaging situations, nonstationary imaging characteristics, and nonlinear reconstruction algorithms. Its major disadvantage is that it provides an evaluation that is valid only for the specific imaging situation investigated.

The proposed method proceeds as follows. First, the whole problem must be completely specified:

a) Define the class of scenes to be imaged including as much complexity as exists in the intended application. Variations in scene from one realization to another should be fully specified.

b) Define the geometry of the measurements. The deficiencies in the measurements such as blur, uncertainties in the geometry, and uncertainties in the measurements (noise) should be specified. Variations of these uncertainties with position as well as intercorrelations between them could be included.

c) Define clearly the task to be performed. The task might be simple detection of a known object against a known background, for example. Alternatively, it could be discrimination between two types of objects, or something more complex, such as multiple discrimination, parameter estimation, etc. The fundamental assumptions made must be explicitly stated.

d) Define the method of task performance. This should be consistent with the intended application. If the task is to be performed by computer, then the intended analysis algorithm may be used. If the task is to be performed by a human observer, some approximation to the human should be used. Alternatively, a maximum-likelihood algorithm (ideal observer) may be employed to define the best possible performance (under the prevailing assumptions made about the extent of auxillary information).

The simulation procedure is then performed by doing the following:

e) Create a representative scene and the corresponding measurement data by means of a Monte Carlo simulation technique. All variations in scene content and uncertainties in the measurements are included by means of pseudo-random selection of the uncertain parameters.

f) Reconstruct the scene with the algorithm being tested.

g) Perform the specified task using the reconstructed image.

h) Repeat steps e) through g) a sufficient number of times to obtain the necessary statistics on the accuracy of the task performance.

Finally determine how well the task has been performed, on the average:

i) Evaluate the task performance. For binary discrimination tasks (of the yes-no variety), a receiver-operating characteristic (ROC) curve [4] may be generated. In a very precise treatment, one might use the Bayes' measure based on the relative costs of making false or true conclusions. For parameter estimation tasks, the standard measure of rms error might be employed.

Example - Evaluation of nonnegativity constraint

The usefulness of the nonnegativity constraint in the algebraic reconstruction technique (ART) [5] will now be explored to demonstrate how the proposed method can be used. It should be noted that such a constraint makes the response of the reconstruction algorithm nonlinear. As such, the task performance for either noise or artifacts is not amenable to linear analysis. For the present example, the scene is assumed to consist of a number of non-overlapping discs placed on a zero background. For this example, each scene contains 10 high-contrast discs of amplitude 1.0 and 10 low-contrast discs with amplitude 0.1. The discs are randomly placed within a circle of reconstruction, which has a diameter of 128 pixels in the reconstructed image. The diameter of each disc is 8 pixels. The first of the series of images generated for these tests is shown in Fig. 2. In this computed tomographic (CT) problem, the measurements are assumed to consist of a specified number of parallel projections, each containing 128 samples.

The above choices for this example are made because they provide a situation in which the nonnegativity constraint is likely to have a substantial effect. In the limited data-taking circumstances we will consider, the high-contrast discs produce serious artifacts in the reconstructions, which make it difficult to detect the low-contrast ones. The artifacts produced by these high-contrast discs depend on their positions. Thus, it is important to allow for random placement of the discs to randomize the artifacts. In some of the test cases

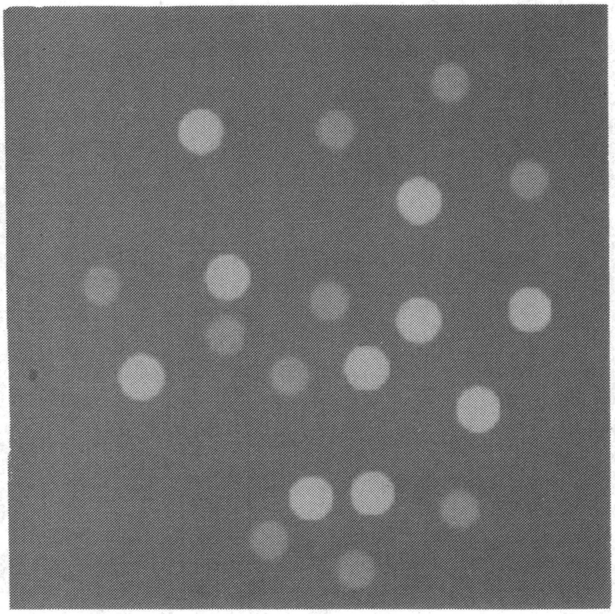

Figure 2: The first randomly generated scene consisting of 10 high-contrast and 10 low-contrast discs. The evaluation of d_A is based on an average over ten similar scenes.

described below, random noise is added to the projection measurements. For these, a Gaussian-distributed random number generator with zero mean is used. This means that negative projections are possible, even though the object itself is nonnegative. While this may seem ridiculous to theoreticians, it is not at variance with many experimental situations, as, for example, when the projections are derived from the measurement of the attenuation of x rays.

The ART algorithm is employed in these examples to reconstruct the original scene. Ten iterations of ART are used throughout. Variable relaxation (or damping) factors are used to attenuate successive updates during the reconstruction. The choice of the relaxation factor is a complex issue, which will not be discussed in detail here but will be addressed in a future publication dealing with the optimal choice of the relaxation factors. For the examples presented here, the algorithm begins with a relaxation factor of 0.2 for 100 views and 1.0 for the other cases, which involve limited numbers of projections. The relaxation factors are multiplied by 0.8 after each iteration, resulting in a final factor that is about seven times smaller than the initial one. This provides regularization in the estimation procedure, which converges to a least-squares solution in the limit that the relaxation factor approaches zero [6]. The result of reconstructing Fig. 2 from 12 noiseless views spanning 180° is shown in Fig. 3. The seemingly random fluctuations in the background are actually artifacts produced by the limited number of projections and arise mainly from the high-contrast discs. At first sight, it appears that the nonnegativity constraint improves the reconstruction considerably in that it has reduced the confusion caused by the fluctuations in the background. However, some of the low-contrast discs have not been reproduced. Also, there still remain many fluctuations in the background that may mislead one to suspect the presence of discs in places where none exist in reality. Thus, on the basis of this single example, one cannot say with certainty whether or not the nonnegativity constraint improves the detection of the low-contrast discs. A statistically significant comparison between reconstructions with and without the constraint must be made to assess its value.

The task to be performed is assumed to be the simple detection of the low-contrast discs. It is assumed that the position of a possible disc is known beforehand as is the background. To perform the stated task of detection, it is assumed that the sum over the area of the disc provides an appropriate decision variable. This is an approximation to the matched filter, which is known to be the optimum decision variable when the image is corrupted by additive uncorrelated Gaussian noise [4]. This ignores the blurring effects of the finite resolution of the discretely-sampled reconstruction. It also does not take into account the known correlation

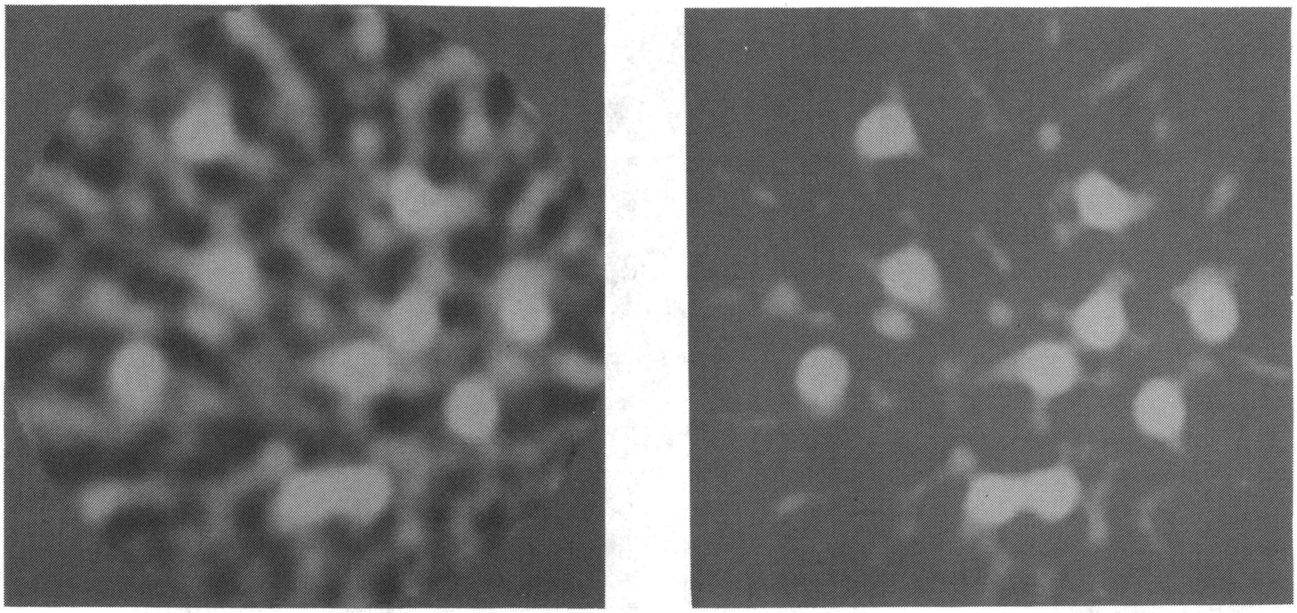

Figure 3: Reconstructions of Fig. 2 from 12 noiseless parallel projections subtending 180° obtained with the ART algorithm (right) with and (left) without the nonnegativity constraint. These images are displayed at high contrast to show the low-contrast discs of interest.

in the noise in CT reconstructions [7] that have been derived from projections containing uncorrelated noise. Nor does it take into account the effects that the nonnegativity contraint has on the character of the noise. After reconstruction, the sums in each region where the low-contrast objects are known to exist are calculated, as well as those over each region where none exist. These two data sets may be displayed as histograms in this decision variable ψ as shown in Fig. 4. A disc will be said to be present at each location where the value of the decision variable is above a chosen threshold. The probability that the presence of a disc is correctly detected, called the true-positive probability, is estimated as the area under the dashed histogram above the threshold. The probability of falsely stating a disc to be present, the false-positive probability, is the area under the solid curve above the threshold. As the threshold is lowered to increase the true-positive rate, the false-positive rate also increases. According to Bayes, a theoretically optimum choice of the threshold value can be made on the basis of the relative costs associated with correctly and incorrectly detecting discs. However, when dealing with human observers, these histograms are not explicitly observable and the choice of the threshold is implicitly made by the observer. The relationship between the two histogram distributions is often characterized by the detectability index d', given by

$$d' = \frac{\psi_1 - \psi_0}{\sqrt{\frac{\sigma_1^2 + \sigma_0^2}{2}}}, \tag{1}$$

where ψ_1 and σ_1 are the mean and rms deviation of the frequency distribution when the object is present and those with the subscript 0 are when the object is not present. This is sometimes called the signal-to-noise ratio (SNR) for detection. For the histograms shown in Fig. 4, d' is 0.87.

The same results may be displayed as an ROC curve [4], which shows the variation of true-positive probability with false-positive probability as the decision threshold changes. The ROC curve completely summarizes binary detection task performance. Figure 5 shows the ROC curve generated directly from the distributions in Fig. 4. Comparison between the ROC curves produced by unconstrained ART and constrained ART shows that the nonnegativity constraint has dramatically enhanced the performance of this detection task. The area under the ROC curve A is known to be the same as the fraction of correct scores that would be obtained in a two-alternative forced choice experiment [8]. The area under the ROC

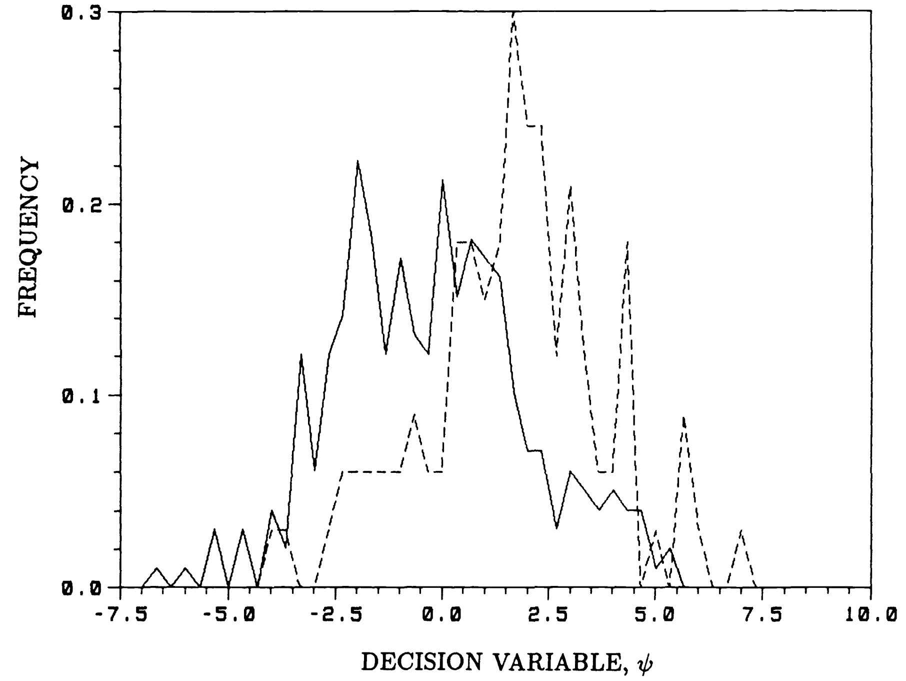

Figure 4: The frequency distributions of the decision variable (the sum over a circular region) evaluated where a low-contrast disc is known to exist (dashed line) and where none exists (solid line) for ART reconstructions without the nonnegativity constraint. These results summarize the performance obtained from reconstructions from 12 views for 10 randomly-generated scenes.

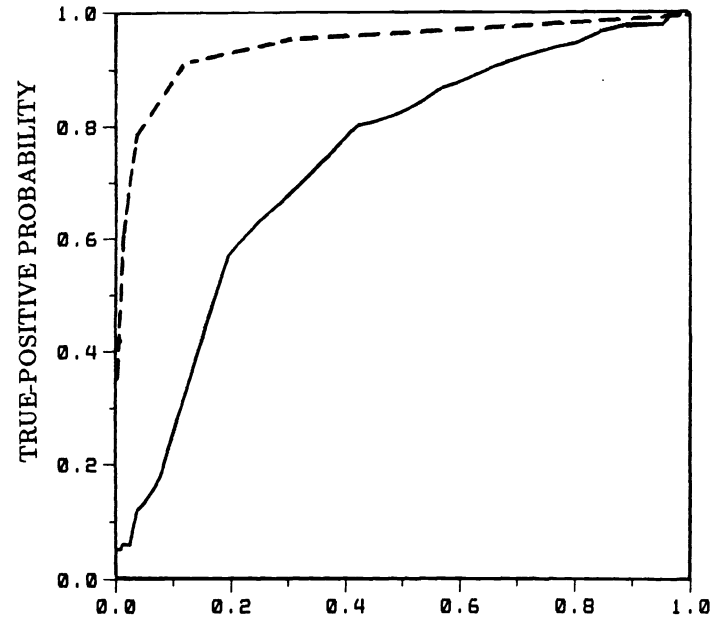

Figure 5: The receiver operating characteristic (ROC) curve derived from the frequency distributions shown in Fig. 4; that is, for unconstrained ART (solid line) and the same for reconstructions obtained with the nonnegativity constraint (dashed line). The nonnegativity constraint is seen to markedly improve task performance since its ROC curve is always significantly above that for no constraints.

curve in Fig. 5 with no constraint is 0.736 compared to 0.948 with the constraint. The area under the ROC curve may be expressed in terms of an effective index for detectability d_A, calculated to be the same as d', under the assumption of Gaussian-shaped frequency histograms

$$d_A = 2\, erf^{-1}\{2\,(1 - A\,)\}, \tag{2}$$

where erf^{-1} is the inverse of the error function. For Fig. 5, a value for d_A of 0.89 is obtained for the case without the constraint and 2.30 with the constraint. Thus, the use of the nonnegativity constraint has increased the detectability by 158% in this case of a limited number of views.

Table 1 tabulates the results obtained under varying data-taking conditions. To refer the noise level to the magnitude of the projections, the peak projection value for the low-contrast discs is 0.80. The nonnegativity constraint is seen to be generally useful. The constraint is particularly helpful when the data are limited by the measurement geometry. It has little effect when the data are complete but noisy. The CPU time required to calculate the entries in the table took as long as one hour on a VAX 8700, which is about four times faster than a VAX 785.

It was noted in essentially all situations tested that, even though the histograms did not always appear to possess Gaussian shapes, d' was very close to being the same as d_A. This is useful to know because d' has better statistical accuracy than d_A and is more likely to be a continuous function of the parameters that can be varied in the reconstruction procedure. This makes d' the performance index of choice for the purpose of optimizing the reconstruction technique.

Table 1: Summary of the effect of the nonnegativity constraint on the detectability index d_A determined from the area under the ROC curve, for various kinds of projection data. These results were obtained from tests made on the CT reconstructions of 10 randomly generated scenes for various kinds of deficiencies in the data. Note that * indicates d' (calculated from the means and variances of the frequency distributions using (1)) was used instead of d_A because of the statistical inaccuracy of the latter.

no. proj.	$\Delta\theta$ (deg.)	rms noise	d_A without	constraint	with		Improvement
100	180	8	1.98		2.00		+1%
100	180	4	4.10	*	3.94	*	–4%
8	180	0	0.45		0.79		+76%
12	180	0	0.89		2.30		+158%
16	180	0	1.94		5.51	*	+184%
16	90	0	1.19		2.49		+109%
32	90	0	1.25		3.57		+186%
16	180	2	1.63		2.74		+69%

Discussion

We have presented a new method to test the effectiveness of reconstruction algorithms. This method is based on a Monte Carlo simulation of the complete imaging process from the composition of the original scene to the final interpretation of the reconstructed image. The accuracy with which a specified task is performed is the goal of the simulation. This method is in accordance with the notion that an algorithm can only be properly evaluated by trying it out on a statistically meaningful sample of trials. A major advantage of the Monte Carlo technique is that new effects may be easily added. On the other hand, only the overall effect of all the conditions is observed. It is difficult to determine the relative contributions of individual effects. The Monte Carlo simulaton technique is particularly useful in situations that do not lend themselves to analytic analysis. It can provide a good statistical sampling over all the uncontrollable variables in the problem. An example of this is the typical problem of the effect of discrete sampling on signal analysis as, for example, in the problem of the detection of small objects. In such a case, what makes sense is to average the detectability over all possible positions of the object relative to the discrete measurements and the reconstruction grid [9]. The Monte Carlo method is perfect for this.

We have seen that the nonnegativity constraint is generally useful for the specific problem addressed here – detection of low-contrast discs in the presence of high-contrast discs using CT reconstructions. This constraint is particularly helpful when the data consist of a limited number of noiseless projections. When the data are complete but degraded by additive noise, the nonnegativity constraint does not improve detectability. Some improvement is obtained when the data are both incomplete and noisy. The effectiveness of the nonnegativity constraint is found to depend on the choice of relaxation parameters used in the ART algorithm. The optimal selection needs investigation.

There are many possible extensions to this preliminary effort. Alternative choices for the decision variables could be pursued to obtain improved performance. For example, a weighted sum of the reconstruction values over a local region could be used. The optimal weights might be determined by using half the simulated reconstructions as a training set and the second half to estimate the task performance index. This would probably be too difficult to handle in general, but, with suitable restrictions on the number of variables in the weights, it might be possible. The optimal choice of decision variable could be dependent upon the reconstruction procedure. If this line of research were pursued, it would be reasonable to compare the performance of one algorithm against another only on the basis of the best decision procedure that could be achieved with each. As the detection task specified in the present example is truly simple and not very closely related to most real problems, another worthwhile extension would be to explore more complex and interesting tasks.

Clearly, this approach of random simulation is generally applicable to test any or all aspects of the entire imaging chain from scene generation to the final method of task performance. Possibly a very fruitful line of research that can be addressed using this approach is the optimization of the imaging system, either in terms of its individual parts or in its entirety. If many parameters are to be varied in the opimization, one must be concerned about the stability of the optimization process. Regularization may be required to stabilize the search for the optimum.

References

[1] H. C. Andrews and B. R. Hunt. *Digital Image Restoration.* Prentice-Hall, Englewood Cliffs, New Jersey, 1977.

[2] 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.

[3] K. M. Hanson. Variations in task and the ideal observer. *Proc. SPIE*, 419:60–67, 1983.

[4] A. D. Whalen. *Detection of Signals.* Academic, New York, 1971.

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

[6] Y. Censor, P. P. B. Eggermont, and D. Gordon. Strong underrelaxation in Kaczmarz's method for inconsistent systems. *Numer. Math.*, 41:83–92, 1983.

[7] K. M. Hanson. Detectability in computed tomographic images. *Med. Phys.*, 6:441–451, 1979.

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

[9] K. M. Hanson. The detective quantum efficiency of CT reconstruction: the detection of small objects. *Proc. SPIE*, 173:291–298, 1979.