Proc. of SPIE Vol. 2434, Medical Imaging 1995: Image Processing, ed. M H Loew (May 1995) Copyright SPIE

MI POSTER AWARD
Cum Laude

Neural Network Performance for Binary Discrimination Tasks. Part II: Effect of Task, Training, and Feature Pre-Selection

K.J. Myers[1], M.P. Anderson[1], D.G. Brown[1], R.F. Wagner[1], K.M. Hanson[2]

[1]CDRH, FDA, 12720 Twinbrook Parkway, Rockville, MD 20857
[2]Los Alamos National Laboratory, Los Alamos, NM 87545

ABSTRACT

Neural networks are applied to the Rayleigh discrimination task. Network performance is compared to results obtained previously using human viewers, and to the best machine approximation to the ideal observer found in an earlier investigation. We find that simple preprocessing of the input image, in this case by projection, greatly improves network convergence and only results obtained on projections are presented here. It is shown that back propagation neural networks significantly outperform a standard nonadaptive linear machine also operating on the projections. In addition, this back propagation neural network performs competitively to a nonadaptive machine that uses the complete two-dimensional information, even though some relevant information is destroyed in the projection process. Finally, improved performance on this Rayleigh task is found for nonlinear (over linear) neural network decision strategies.

Keywords: Neural networks, reconstruction algorithms, image evaluation, back propagation, DYSTAL.

1. INTRODUCTION

We have previously described how imaging systems and image reconstruction algorithms can be evaluated based on the ability of machine and human observers to perform a binary discrimination task using the resulting images.[1-5] The performance of an ideal or Bayesian observer based on Bayesian statistical decision theory gives an upper bound on the performance of all other observers, human or machine. However, when the images being evaluated are formed using nonlinear algorithms, the Bayesian decision variable can be difficult to calculate. Sophisticated nonlinear algorithms may be required to form readable images from future imaging systems that take only a limited number of views, for example, high-speed imaging systems or novel techniques in coronary angiography. To evaluate such images, sub-optimal machine approximations to the Bayesian observer might be used instead. We used this approach in our study of images reconstructed from a limited number of tomographic views using a maximum-entropy algorithm.[2-5] While the machine approximations we considered tracked our human performance data remarkably well, a troubling question remained: How much higher would the true ideal observer's performance curve be above that of the machine approximation?

Neural networks have been shown to converge to the ideal-observer solution for signal-known-exactly/background-known-exactly tasks.[6] We are therefore studying several different neural networks in an attempt to improve on the performance of our machine observers for tasks using images reconstructed from a limited number of views. In our first investigation,[7] neural networks were applied to the task of detecting low-contrast lesions in limited-view tomographic images. The neural networks included both one- and two-layer back propagation neural networks with various interconnection architectures, and a biologically-based neural network

0-8194-1782-3/95/$6.00

called DYSTAL.[8,9] The neural networks were unable to match the disk-detection performance of the best machine observer. We postulated that this result might be due to some combination of the following three factors: 1) the task was inherently linear, 2) insufficient training had been done, and 3) preselection of features was needed for the neural networks to be successful.

To address factor (1) above, we have considered another task, the so-called "Rayleigh" task. Scenes containing bar-shaped and binary objects on a flat background were prepared. Reconstructions of the scene were generated using a maximum-entropy algorithm for a range of values of the reconstruction parameter α. Sub-images known to contain either a bar or a binary object were extracted from the scene; the discrimination task was the determination of which of two sub-images contained the binary object. While it is not clear that this task is any more nonlinear than disk detection, its resolution requirements are quite different. The dependence of the sub-optimal machine performance on α differ dramatically for the two tasks. The current investigation will determine whether the neural networks are affected similarly.

To investigate factor (2) above, the number of images was increased by a factor of 10 for this task, allowing many more to be used for both training and testing purposes.

Finally, we have investigated the preselection of features as suggested by factor (3). Specifically, for the Rayleigh task, we have studied the performance of the neural networks for projections of the sub-images orthogonal to the main axis of the object.

2. EXPERIMENTAL METHODS

In the following sections we describe the manner in which images were generated and evaluated, including scene and data generation, image reconstruction, and evaluation by human, machine, and neural network observers.

2.1 The Scene and the Data

A set of 1250 scenes was generated using a Monte Carlo procedure for random scene generation.[1] Each scene contained 8 Gaussian doublets and 8 Gaussian bars. Thus 10,000 of each object were generated in total. The objects were randomly placed and oriented in a circle of reconstruction inscribed in a 128 x 128 array. The doublets were pairs of points separated by 6 pixels and convolved with a 2D Gaussian function with a 4 pixel FWHM. The bars were formed by convolving the same Gaussian function with a bar of length 10.4 pixels. An example scene is shown in Figure 1. This same scene description was used previously in our investigation of machine and human observer performance for the Rayleigh task.[3,5] The purpose of the current investigation is to compare neural network performance with the earlier results.

Data consisted of just 8 parallel projections equally spaced over 180°, each containing 128 samples. Additive, zero-mean Gaussian noise with $\sigma=1.0$ was added to each projection. The noise in the data was 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.

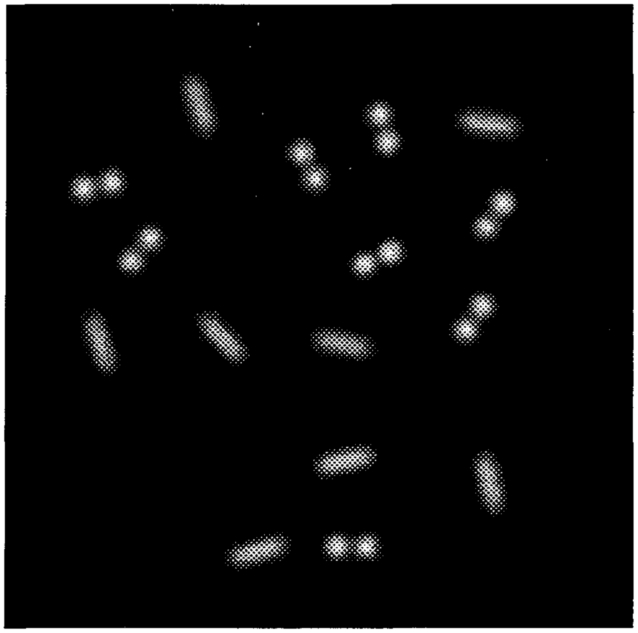

Figure 1. Sample Rayleigh scene.

2.2 The Reconstruction Algorithm

With just eight views forming the data set, image reconstruction by standard filtered back-projection would result in severe streak artifacts. A number of nonlinear algorithms have been developed that estimate the underlying scene using some form of prior knowledge to augment the limited data set. Typically these algorithms have one or more adjustable parameters that determine the relative weight of the information from the prior and the data that are combined to estimate an image. The optimal choice for these weights is observer- and task-dependent.

In the current work, image reconstruction was performed using the commercially available code MEMSYS 3 to provide Bayesian reconstructions based on an entropy prior.[10] This maximum-entropy algorithm has a variable reconstruction parameter α that controls the degree of smoothing of the images. Low values of α result in sharp, high-resolution images and high values lead to smooth, low-resolution ones. Sample reconstructions are shown in Figure 2 for the range of α values used in this study.

The artifacts visible in the reconstructions of Figure 2 are deterministic but object-dependent. From Figure 2 it is also clear that the character of the artifacts is affected by the reconstruction parameter α. Because the objects in the scenes are randomly placed, the artifacts supply an additional stochastic component to the randomness of the images beyond that which results from the Gaussian additive noise.

2.3 The Task

For the Rayleigh task posed here, observers were evaluated for their ability to discriminate between the Gaussian bar and the Gaussian doublet. Evaluation of human and neural network performance followed a two-alternative forced-choice (2AFC) paradigm: observers were presented with two image subregions and asked to choose which contained the bar object and which contained the doublet. The subregion was a 19 x 19 pixel region centered on the object. Although the objects were oriented randomly in the scenes prior to data-taking, the subregions were rotated to align the long axis of the object horizontally prior to presentation to the observers. It was not our purpose to investigate a task with object orientation unknown; rather, the random orientation served to probe the artifact-generating characteristics of the imaging system. The task was therefore signal non-random (shape, position, and orientation known) but the background statistics were unknown to the observers (because of the random artifacts and the unknown nature of the random noise after processing by the nonlinear reconstruction algorithm).

Machine observers were evaluated using standard receiver-operating-characteristic-curve (ROC) methodology as described in Section 2.5 below.

Sample sub-images for each type of object are shown in Figure 3.

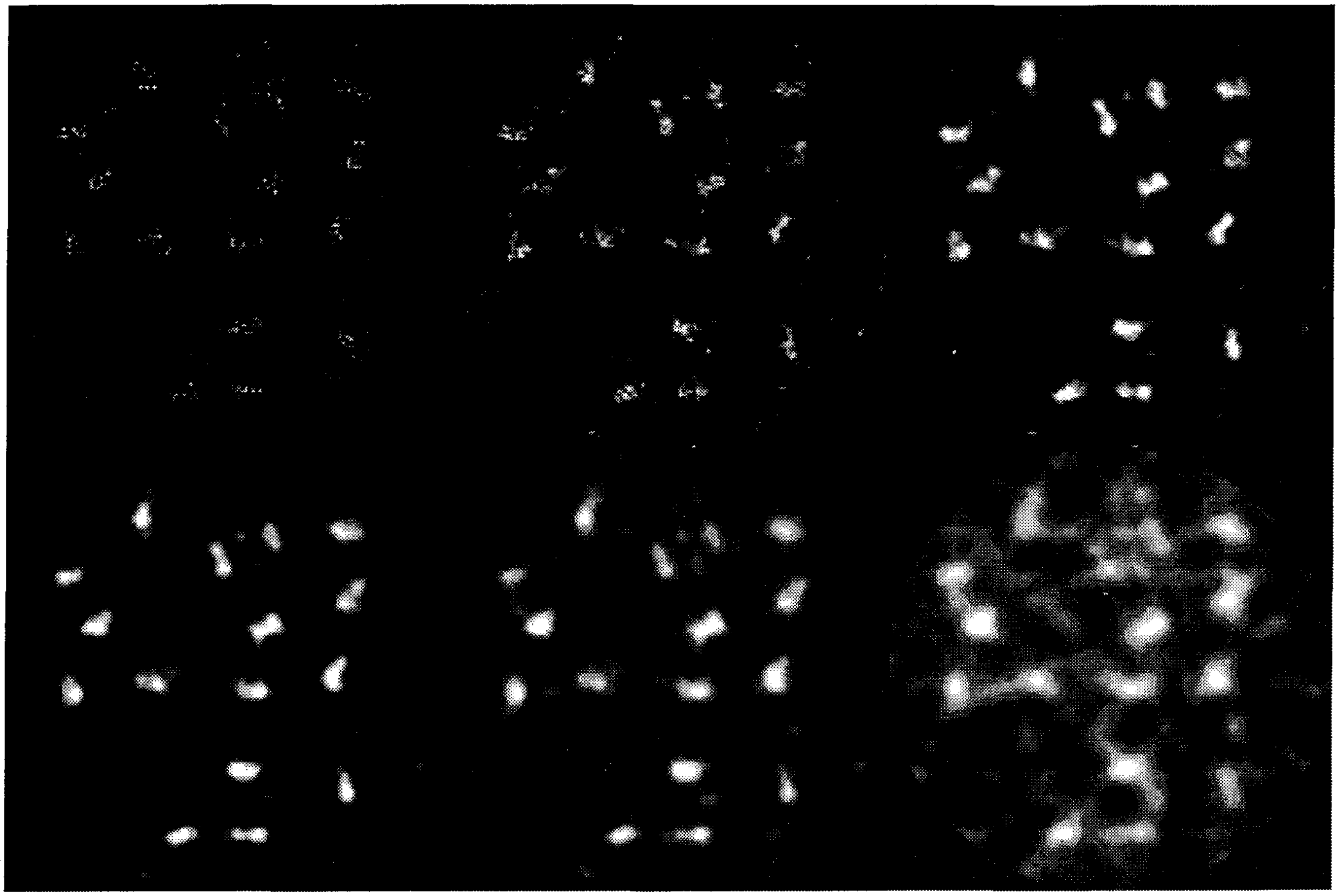

Figure 2. Reconstructions of the scene shown in Figure 1 for a set of values of the reconstruction parameter α. Across the top, α = 0.005, 0.05, and 0.6. Across the bottom, α = 4, 19.5, and 100.

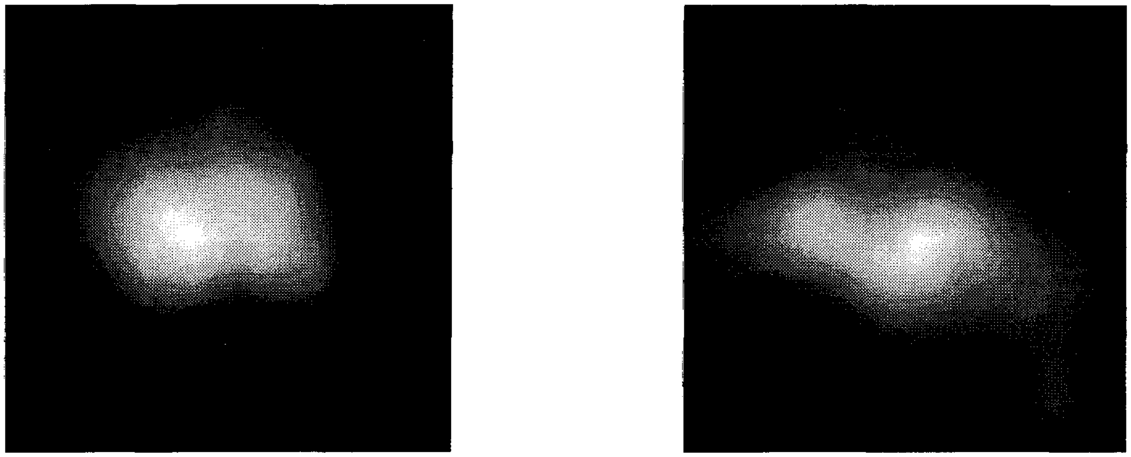

Figure 3. Sample sub-images: 19 x 19 regions extracted from a reconstructed image (interpolated for display purposes) showing a Gaussian doublet on the left and a Gaussian bar on the right.

2.4 Human Observers

Details of the human observer study were reported previously by Myers, Wagner, and Hanson.[5] That investigation was carried out on a set of 320 sub-images, half containing the bar and half containing the doublet. Two human observers performed a 2AFC study in which they viewed pairs of sub-images, one from each class, drawn randomly from the population of sub-images. Also displayed were samples of a bar and doublet object from the original scene, so that the observers had full knowledge of the objects' size and shape. Feedback on the correctness of individual decisions was given to the observers after each decision. The percentage correct determined at the end of the experiment corresponds to the area under the receiver operating characteristic curve.[11] This area measure was converted to a detectability measure according to the method given in the Evaluation Methodology section that follows.

Human data have not been collected on the full data set of 20,000 objects obtained for this study. Thus, human performance will be reported based on the earlier study on 320 sub-images.

2.5 Machine Observers

Several machine approximations to the Bayesian observer were considered in our earlier work, ranging from linear non-adaptive ones to nonlinear adaptive ones that were based on the posterior probability function.[3] The highest performance was found for a machine that was assumed to know the exact form of the two alternative signals, and who calculated the mean-squared difference between the reconstruction and each of the known signals. The difference between the values obtained for each of the expected signals (the bar and the doublet) formed the machine's decision variable. Values of the decision variable for each of the 320 sub-images in the first study were histogrammed separately for the known bar and doublet objects. Then, by varying the decision-function threshold, the receiver operating characteristic (ROC) curve was generated. The area under the curve was used to obtain the detectability index d_a as indicated in the Evaluation Methodology section.

The mean-squared-difference strategy is approximately equivalent to a non-prewhitening matched filter (NPWMF). NPWMF results were calculated on the full set of 20,000 sub-images. Since neural network performance was determined for sub-images preprocessed by projection, NPWMF performance was also calculated on the set of 20,000 projected sub-images.

2.6 Neural Network Observers

Two types of neural networks were investigated. The neural networks included both one- and two-layer back propagation neural networks with various interconnection architectures, and a biologically-based neural network called DYSTAL. Adaptive neural network performance was computed on the same sub-images viewed by the non-adaptive (NPWMF) machine reader.

Preprocessing can increase neural network performance by improving network convergence for finite computing times, and by reducing the number of needed nodes or layers. We therefore investigated the performance of the neural networks for images preprocessed by a projection operation. Each sub-image was projected perpendicular to the known long axis of the object. This operation reduced the dimensionality of the input to the neural network from 361 to 19.

DYSTAL

The DYnamically STable Associative Learning (DYSTAL) neural network was developed from an examination of biological neural systems.[8,9] It may be thought of as a vector quantization scheme. During the one-pass training, a set of vectors was generated corresponding to each of the classes to be discriminated. During testing, the test vector was assigned to the class of the training vector to which it was closest. The degree of closeness was determined by a similarity measure -- of which there are a number of plausible choices. The original DYSTAL similarity measure is the dot product between the normalized training ("patch") vector and the normalized input feature vector. Our choice of similarity measure S was simply the square of the element-by-element difference between the training ("patch") vector P and the input feature vector I:

$$S = \sum_{j} [P_j - I_j]^2$$

Back Propagation Neural Network

The back propagation neural networks used in this study were feed-forward, completely connected networks with the usual logistic "squashing" function output from each node. Two distinct network architectures were used on the sub-images: a simple two-layer network with 361 input nodes and a single output node; and a three-layer network with 361 input, 16 hidden layer, and again a single output node. For projected sub-images the two networks were: 19 input, 1 output; and 19 input, 4 hidden, 1 output node. The target outputs for the neural networks were set to 0.1 and 0.9, rather than 0.0 and 1.0, in order to facilitate fluidity of training. "Error" contributions from exceeding the target values were ignored.

2.7 Evaluation Methodology

Performance evaluation for all the observers included in this study was based on the area under the receiver operating characteristic (ROC) curve, A_z. For human and neural network observers, a 2AFC paradigm was followed. The percentage of correct decisions determined by each 2AFC study is equivalent to A_z.[11] For the machine observer, A_Z was determined from the decision variable histograms by calculating the true-positive and false-positive fractions obtained as a decision threshold was varied.

Values for A_z were translated to a signal-to-noise-ratio (SNR) scale using the concept of a detectability index, d_a The detectability is the SNR of an equivalent (equal A_z) Gaussian-distributed decision function. The value for d_a was obtained from A_z by:

$$d_a = \sqrt{2}\ [\phi^{-1}(A_z)]\ \ ,$$

where Φ^{-1} is the inverse Gaussian distribution function. Our results are presented in terms of d_a^2 because it is this quantity that is relevant to discussions of observer efficiency.

3. RESULTS

Results for the initial set of 320 objects are shown in Figure 4 for human observers, three types of neural networks, and the best machine observer found in our earlier study. The machine observer was a nonadaptive observer, approximately equivalent to the non-prewhitening matched filter (NPWMF). The networks were trained on 80 sub-images of each type (Gaussian bar versus Gaussian doublet) and tested on the remaining 80 sub-images of each type. Clearly the performance of the back propagation neural networks is superior to human performance and approximates or betters that of the machine observer. For all observers there is a broad peak in performance vs. α, and an apparent fall-off in both directions away from that value. The original DYSTAL appears to best approximate human performance, giving the worst performance of all the computational observers.

Results for the complete data set of 20,000 sub-images are shown in Figure 5. For both DYSTAL and back propagation a subclassification has been made in the figure between "linear" and "nonlinear." "Linear" for DYSTAL implies that the patch creation threshold was set so high that only one patch per class was created. "Linear" back propagation means only a simple perceptron architecture was used with no hidden layer of nodes. "Machine 1" and "Machine 2" are the one- and two-dimensional NPWMF results, where the filter was determined using the known profiles of the Gaussian bar and doublet used in the scene generation. That is, the filter function was based on the object profiles prior to the data acquisition and image reconstruction process. DYSTAL results were also calculated on both the 2D sub-images and the 1D inputs obtained by projection; similar results were found in each case. However, for the 2D sub-images with 361 input values, back propagation convergence was too slow, especially for the larger values of α, for results to be given at this time. Thus, for consistency, all neural network results in Figure 5 are presented for the sub-images which were preprocessed by projection. Neural network training was accomplished on the first 5000 sub-images of each type, and testing was carried out on the remaining 5000 sub-images of each type.

We see from Figure 5 that performance improves significantly in going from DYSTAL, to nonlinear DYSTAL, to the NPWMF operating on the projections, to the two-dimensional NPWMF and the back

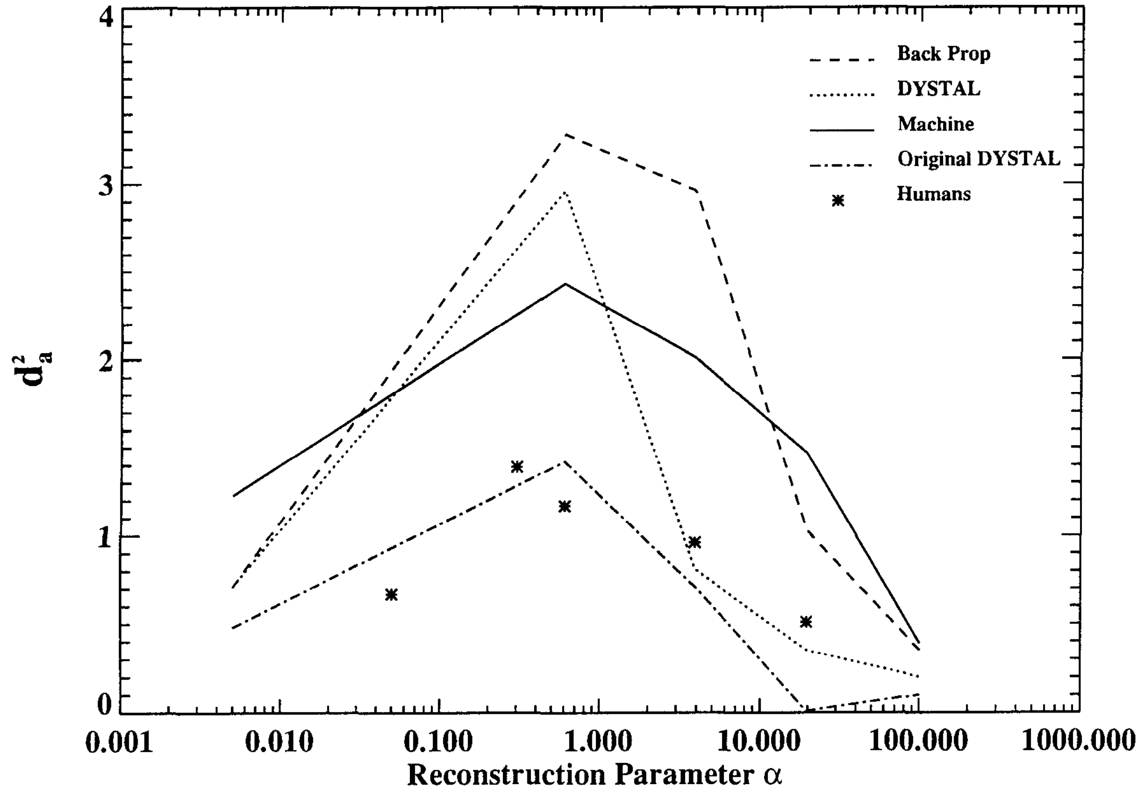

Figure 4. Performance in terms of d_a^2 as a function of the reconstruction parameter α for the pilot set of 360 objects.

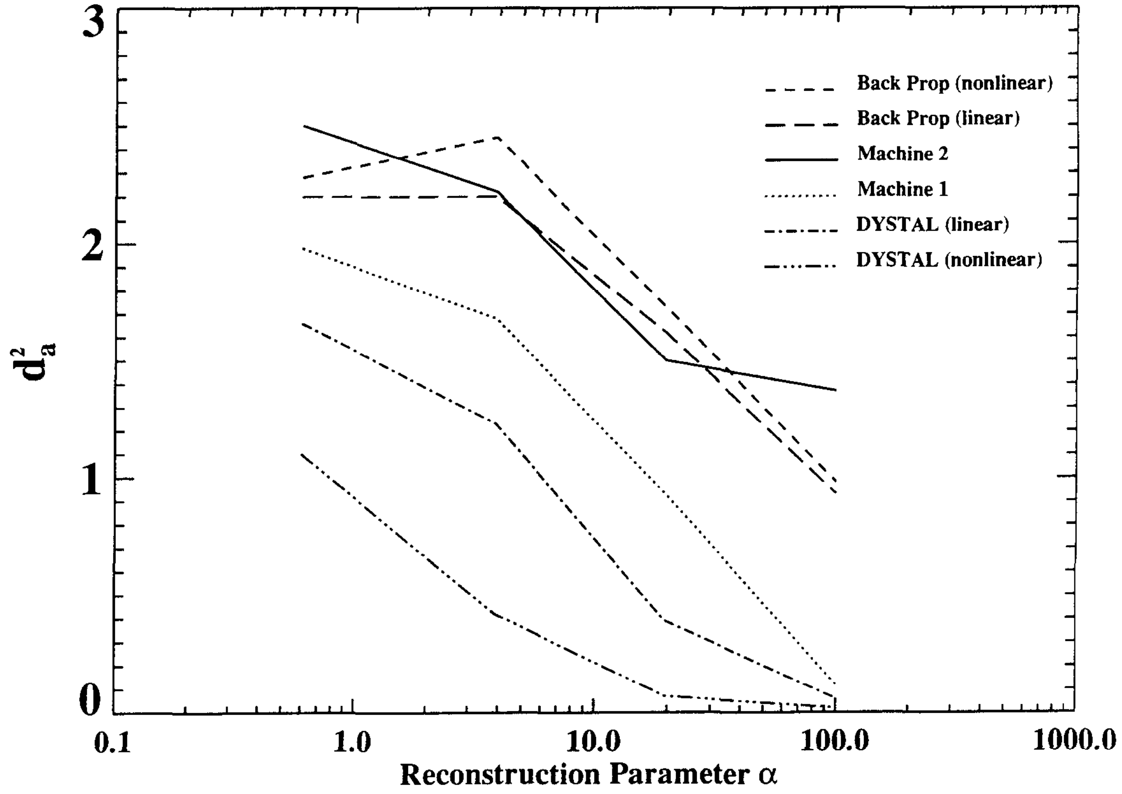

Figure 5. Performance on the full set of 20,000 objects.

propagation neural network. The nonlinear back propagation neural network for $\alpha=100$ had not had a sufficient number of training iterations to converge, leading us to suspect that more training would pull the back propagation curves further apart at the largest value of α. The present task appears to require a nonlinear observer, unlike the results of the previous paper for which nonlinearity was found to provide no additional benefit.

The NPWMF observers use *a priori* information about the objects in the scene to form a template with which to match the observed image or projection. In comparison, the neural network must use the sub-images (or the projections in the 1D case) themselves to estimate the form of such a template. One common assumption regarding this estimation process is that the network uses (or should use) the difference between the average inputs of the two classes as a template. We therefore computed the difference between the average network inputs for each of the classes using the first 5000 of each class, and determined the performance of this filter on the remaining 5000 members of each class. This was done on the sub-images preprocessed by projection to compare to the results described above for the back propagation neural network. We found that the performance of this observer was below that of the projection-data machine observer, and therefore well below that of the back propagation neural network operating on the projections. This observer corresponds to the neural network at a very early stage of training. Almost the entire training period of the back propagation neural network is spent "fine tuning" the difference image template to obtain improved performance.

We were unable to obtain results at the lowest values of α considered in our pilot study because of the tremendous computing time it takes to obtain reconstructions that fit the data to that degree. (In general, the amount of time it takes to form a reconstruction increases as α decreases.) We intend to extend this work to determine whether the downturn in performance observed in our pilot study for smaller values of α will be found for much larger numbers of samples. It is interesting to note, though, that the number of iterations required to train the neural network decreases as α decreases. The increased resolution in the images at low α means less time is required by the neural network to optimize its weights. The two factors together -- time to form a reconstruction and time to train the neural network -- work together to make it difficult to obtain precise data for either very low or very high values of α for large numbers of images.

4. CONCLUSIONS

We have shown that back propagation neural networks significantly outperform a standard nonadaptive linear machine when both are operating on the projections. In addition, the back propagation neural network performs competitively to a nonadaptive machine that uses the complete two-dimensional information, even though some relevant information is destroyed in the projection process. It remains to be seen how well the neural network will perform when it is presented with the complete two-dimensional sub-image as input. Finally, improved performance on this Rayleigh task was found for nonlinear (over linear) neural network decision strategies.

5. ACKNOWLEDGMENTS

The authors wish to thank Jonathan Boswell, Jackie Deming, Bob Gagne, and Phil Quinn for computer, secretarial, and graphics support.

6. 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. Myers, K.J., K.M. Hanson, R.F. Wagner, "Task Performance Based on the Posterior Probability of Maximum-Entropy Reconstructions Obtained with MEMSYS 3," Proc. of the SPIE **1443**, 172-182 (1991).
3. K.M. Hanson, K.J. Myers, "Rayleigh Task Performance as a Method to Evaluate Image Reconstruction Algorithms," Tenth International Workshop on Maximum Entropy and Bayesian Methods in Science and Engineering, Laramie, WY (August 1990). In Maximum Entropy and Bayesian Methods, W.T. Grandy, Jr. and L.H. Schick, eds. (Kluwer, Netherlands: 303-312, 1991).
4. R.F. Wagner, K.J. Myers, and K.M. Hanson, "Task Performance on Constrained Reconstructions: Human Observer Performance Compared with Sub-Optimal Bayesian Performance," Proc. SPIE **1652**, 352-362 (1992).
5. K.J. Myers, R.F. Wagner, K.M. Hanson, "Rayleigh Task Performance in Tomographic Reconstructions: Comparison of Human and Machine Performance," Proc. of the SPIE **1898**, 628-637 (1993).
6. D.W. Ruck, S.K. Rogers, M. Kabrisky, M.E. Oxley, B.W. Suter, "The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function," IEEE Trans. Neural Networks **1**, 296-300 (1990).
7. D.G. Brown, M.P. Anderson, K.J. Myers, R.F. Wagner, "Comparison of Neural Network, Human, and Sub-Optimal Bayesian Performance on a Constrained Reconstruction Detection Task," Proc. of the SPIE **2167**, 614 (1994).
8. K.T. Blackwell, T.P. Vogl, S.D. Hyman, G.S. Barbour, D.L. Alkon, "A New Approach to Hand-Written Character Recognition," Pattern Recognition **25**, 655-666 (1992).
9. T.P. Vogl, K.T. Blackwell, J.M. Irvine, G.S. Barbour, S.D. Hyman, and D.L. Alkon, "DYSTAL: A Neural Network Architecture Based On Biological Associative Learning," *Progress in Neural Networks*, Vol. III, C.L. Wilson, and O.M. Ovidvar eds., Ablex Publishing Corp., Norwood, NJ (1992).
10. S.F. Gull and J. Skilling, *Quantified Maximum Entropy - MEMSYS 3 Users' Manual*, Maximum Entropy Data Consultants Ltd., Royston, England (1989).
11. D.M. Green, J.A. Swets, *Signal Detection Theory and Psychophysics*, Krieger, NY (1974).