# A Comparison of Image Quality Models and Metrics Predicting Object Detection*

A. M. Rohaly
U.S. Army Research Laboratory, Aberdeen Proving Ground, MD

A. J. Ahumada, Jr., A. B. Watson
NASA Ames Research Center, Moffett Field, CA

## Abstract

Many models and metrics for image quality predict image discriminability, the visibility of the difference between a pair of images. We compare three such methods for their ability to predict the detectability of objects in natural backgrounds: a Cortex transform model with within-channel masking, a Contrast Sensitivity filter model, and digital image difference metrics. Each method was implemented with three different summation rules: the root mean square difference, Minkowski summation with a power of 4, and maximum difference. The Cortex model with a summation exponent of 4 performed best.

## Introduction

Many models and metrics for image quality predict image discriminability, the visibility of the difference between a pair of images.[1] Some image quality applications, such as the quality of imaging radar displays, are concerned with object detection and recognition. Object detection involves looking for one of a large set of object sub-images in a large set of background images and has been approached from this general point of view.[2] We find that discrimination models and metrics can predict the relative detectability of objects in different images, suggesting that these simpler models may be useful in some object detection and recognition applications. Here we compare three types of image quality metrics that are measures of image discriminability. The first is a multiple spatial frequency channel model based on the Cortex transform with within-channel masking.[3,4,5] It is similar to the models of Lubin and Daly.[6,7] The second is a single contrast sensitivity function (CSF) filter model, and the third are metrics based on differences between the original images in the digital domain. Three different summation rules were tried in each case: the root-mean-square (RMS) difference of the outputs, the Minkowski sum of the differences with an exponent of 4, and the maximum difference. The three summation rules can all be regarded as Minkowski sums with exponents of 2, 4, and infinity, respectively. These models span the range of image quality metrics reviewed by Ahumada.[1]

## Object detection experiment

### Experimental method

Stimuli. Six images of a vehicle in an otherwise natural setting were altered by replacing the vehicle with appropriate background imagery. Two object images having lower levels of detectability were constructed from each image pair by mixing the object and non-object images. The mixing proportions were selected individually for each image and chosen to be near threshold detectability. The 510x480 pixel images were presented on a 13 inch Macintosh color monitor at a viewing distance corresponding to 95 pixels per degree of visual angle.

Observers. The observers were 19 male soldiers, aged 18 to 32 years. Their acuities were 20/20 or better, and they had normal color vision.

Procedure. Observers were asked to rate each of the 24 images on a 4 point rating scale according to the following interpretation:
1-A target was definitely in the scene.
2-There was something in the scene that probably was a target.
3-There was something in the scene but it probably was not a target.
4-There was definitely no target in the scene.

One group of 10 observers saw each image 20 times at a duration of 1 sec. A second group of 9 observers saw each image 10 times at a duration of 0.5 sec and 10 times at a duration of 2 sec. The sequence of 480 images was completely randomized separately for each observer.

*Revision of 1995 SID Digest Paper 6-2.

## Data analysis

Methods. The distance in discriminability units from each object image to its corresponding non-object image was measured in the context of a one-dimensional Thurstone scaling model.[8] The scaling model incorporated the following assumptions:

1. Internal stimulus values have a normal distribution with unit variance.
2. Distances between stimulus distributions for mixtures of stimuli are proportional to the ratios of the mixtures.
3. All subjects have the same pattern of distances between the stimulus distributions. They differ only by a multiplicative subject sensitivity factor.
4. Category boundaries vary across subjects, but are constant over stimuli.

The scaling model had one image discriminability parameter for each of the 6 image sets and one sensitivity factor and 3 category boundaries for each of the observers. Observers tested with the two different stimulus durations were allowed separate sensitivity factors. Parameters were estimated by the method of maximum likelihood.

Experimental results. Discriminability parameter estimates scaled to represent the distance (d') from the 100% vehicle image to the non-vehicle image are given in Table 1. Patterns of discriminability differences for the images were estimated separately for the 10 observers given 1.0 sec durations and the 9 observers given the 0.5 and 2.0 sec durations. The median observer sensitivity factor for each group was used to convert the sensitivity pattern to sensitivities. The values for the combined group are the geometric means of the individual group values.

Table 1 - Experimental discriminability indices (d')

| image pair | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| n=10 | 4.1 | 10.3 | 3.7 | 6.7 | 4.5 | 3.7 |
| n=9 | 5.5 | 10.3 | 4.7 | 8.5 | 4.9 | 4.7 |
| n=19 | 4.8 | 10.3 | 4.2 | 7.6 | 4.7 | 4.2 |

For the 10 observer group, the ratio of the best observer sensitivity factor to the median observer sensitivity factor was 1.5 and to the worst observer sensitivity factor was 3.3. For the 9 observer group these ratios were 1.9 and 4.1, respectively. The sensitivities measured for the 0.5 and 2.0 sec durations were neither appreciably nor significantly different from each other.

## Models and metrics

Although the observers were presented with color images, the models were only presented with gray scale images. The RGB color images were converted to gray scale using the coefficients 87/253, 127/253, and 39/253 for the respective color planes. Also, these gray scale images were pixel-averaged by factors of two in the horizontal and vertical dimensions.

### Algorithms

Multi-channel model. The channel model calculations had the following steps: The images were converted to luminance contrast based on the mean luminance of the non-object image. A CSF filter followed by the cortex transform was then applied. The differences between the transforms of the object and non-object images were then masked by the transform values of the non-object image. A masking exponent of 0.7 was used and the output scaled to predict the detectability of grating patches on a uniform background. Finally, these JND (just-noticeable-difference) images were summed using a Minkowski metric with exponents of 2, 4, and infinity.

CSF filter model. For the filter model, the images were also converted to luminance contrast based on the mean luminance of the non-object image. A CSF filter was then applied and the difference image formed and summed using a Minkowski metric with exponents of 2, 4, and infinity.

The CSF filters were calibrated separately for each of the 6 combinations of channel or filter model and exponent. They were designed to fit the prediction of Barten's CSF formula for 1.33 deg square grating patches over frequencies ranging from 1.125 to 18 cycles per degree in octave steps.[9]

Digital image metrics. The difference between the gray scale images was formed and summed using the same three Minkowski metrics. The digital metrics have two implicit parameters, the pixel density Nyquist frequency of 24 cycles/deg, a high frequency cutoff, and the display gamma of 1.5, which controls the relative weighting of differences as a function of luminance level.

### Model results

Least squares predictions of the observer discriminabilities from the model predictions were computed in the log domain, assuming only an additive constant, because analyses in the

Table 2 - Prediction errors in percent of discriminability indices (d')

| | channel | | | CSF filter | | | image | | |
|---|---|---|---|---|---|---|---|---|---|
| exponent | 2 | 4 | ∞ | 2 | 4 | ∞ | 2 | 4 | ∞ |
| n=10 | 51 | 33 | 41 | 55 | 55 | 86 | 49 | 43 | 38 |
| n=9 | 47 | 28 | 30 | 53 | 50 | 79 | 59 | 51 | 46 |
| n=19 | 48 | 30 | 35 | 54 | 52 | 82 | 53 | 47 | 42 |

Table 3 - Prediction errors for image contrast corrected models

| | channel | | | CSF filter | | | image | | |
|---|---|---|---|---|---|---|---|---|---|
| exponent | 2 | 4 | ∞ | 2 | 4 | ∞ | 2 | 4 | ∞ |
| *a* | 7 | 14 | 9 | 0 | 0 | 0 | 0 | 0 | 0 |
| error | 25 | 16 | 26 | 14 | 13 | 33 | 17 | 14 | 18 |

discriminability domain showed neither constant terms nor squared terms significantly improved the fits. The standard errors of the predictions converted to percentage error in discriminability units are shown in Table 2. The best metric is the channel model with a summation exponent of 4. The image difference metric did well with the maximum rule while the CSF filter model did very poorly with the maximum rule. Figure 1 shows plots of the predictions of the average subject detectabilities for the 6 image pairs for each algorithm using the best of the three summation rules for that algorithm (exponents of 4, 4, and infinity, for the channel, filter, and image difference rules, respectively). The error bars represent 95% confidence intervals for the mean of the two groups of subjects based on the variance between the two groups.

## Contrast normalization

Recent data have shown that contrast energy at other spatial frequencies raises the threshold of grating increments and models have been developed to account for this effect.[10] A simple way of allowing for such an effect is to multiply the above predictions by $a/\sqrt{a^2 + c^2}$, where $c$ is the RMS background image contrast passed by the contrast sensitivity function, and $a$ is a parameter estimated from the data. For the filter models the best estimate of $a$ was close to zero, so in this case we simply divided the predicted discriminability by the RMS contrast of the filtered background image. For the image difference model, we divided the image difference measure by the standard deviation of the background image values.

Table 3 has the resulting errors shown as a percentage of the pooled group discriminabilities. Values of $a$ are in percent contrast. With contrast normalization all three metrics performed much better and essentially all did equally well at their optimal summation exponent of 4. Figure 2 shows these results plotted as in Figure 1. If the value of $a$ for the filter model is set to about half that for the channel model, the filter model performs well using our relative error measure and also accurately predicts the absolute level of performance when calibrated for threshold contrast detection. An implicit parameter of the contrast gain correction is the size of the region over which the contrast is computed. In this case it was a 2.7 deg square, not out of line with psychophysical measurements of the width of the contrast gain control region.[11,12]

## Conclusions

Discrimination models designed to answer, "Are these two images different?" can predict answers to the question, "Is there an object in this image?" Without contrast gain factors, the multiple channel model performs better than the simpler models, and it is the only model that comes close to predicting the absolute level of performance when calibrated for threshold contrast detection. However, a contrast gain term allows much better prediction and obviates the need for complex models in this particular situation. CSF weighting does not necessarily improve measures of image quality.[13,14]

## Acknowledgements

R. Horng wrote and ran the MatLab program that generated the model and metric predictions. Parts of this work have been reported previously.[15,16,17] This work was supported in part by NASA RTOPs 505-64-53 and 537-08-20.

## References

1. A.J. Ahumada, Jr. (1993) Computational image quality metrics: a review. *SID Digest, 24*, 305-308.
2. H.H. Barrett (1992) Evaluation of image quality through linear discriminant models. *SID Digest, 23*, 871-873.
3. A.B. Watson (1983) Detection and recognition of simple spatial forms, in O. J. Braddick and A. C. Sleigh, eds., *Physical and biological processing of images*, Springer-Verlag, Berlin.
4. A.B. Watson (1987) The Cortex transform: rapid computation of simulated neural images, *Computer Vision, Graphics, and Image Processing, 39*, 311-327.
5. A.B. Watson (1987) Efficiency of an image code based on human vision. *JOSA A, 4*, 2401-2417. 7. S. Daly (1993) The visible differences predictor: an algorithm for the assessment of image fidelity, in Watson, ed. *Digital Images and Human Vision*. MIT Press, Cambridge, MA.

7. J. Lubin (1993) The use of psychophysical data and models in the analysis of display system performance, in Watson, ed. *Digital Images and Human Vision*. MIT Press, Cambridge, MA.
8. W.S. Torgerson (1958) *Theory and Methods of Scaling*, Wiley, New York.
9. P.G.J. Barten (1993) Spatiotemporal model for the contrast sensitivity of the human eye and its temporal aspects, in B. Rogowitz and J. Allebach, eds., *Human Vision, Visual Processing, and Digital Display IV*, Proc. Vol. 1913, SPIE, Bellingham, WA, pp. 2-14.
10. J.M. Foley (1994) Human luminance pattern-vision mechanisms: masking experiments require a new model, *Journal of the Optical Society of America A,* vol. 11, pp. 1710-1719
11. J.S. DeBonet, Q. Zaidi, (1994) Weighted spatial integration of induced contrast-contrast, *Investigative Ophthalmology and Visual Science,* vol. 35 (ARVO Suppl.), p. 1667.
12. M. D'Zmura, B. Singer, L. Dinh, J. Kim, J. Lewis, (1994) Spatial sensitivity of contrast induction mechanisms, *Optics and Photonics News,* vol. 5, no. 8 (suppl), p. 48 (abs).
13. J. Farrell, H. Trontelj, C. Rosenberg, and J. Wiseman (1991) Perceptual metrics for monochrome image compression. *SID Digest, 22*, 631-634.
14. B. Girod (1993) What's wrong with mean squared error? in Watson, ed. *Digital Images and Human Vision*. MIT Press, Cambridge, MA.
15. A.B. Watson and A.J. Ahumada, Jr. (1994) A modular, portable model of image fidelity, *Perception*, Vol. 23, ECVP Suppl., p. 95 (Abs.).
16. A.M. Rohaly, A.J. Ahumada, Jr., and A.B. Watson (1994) Visual detection in natural backgrounds, *Optics and Photonics News, 5*, (OSA Annual Meeting Suppl.), 48 (Abs.).
17. A.J. Ahumada, Jr., A.B. Watson, A.M. Rohaly (1995) Models of human image discrimination predict object detection in natural backgrounds, in B. Rogowitz and J. Allebach, eds., *Human Vision, Visual Processing, and Digital Display IV*, Proc. Vol. 2411, SPIE, Bellingham, WA, paper 34.

Figure 1. Detection data predictions by three image quality metrics.

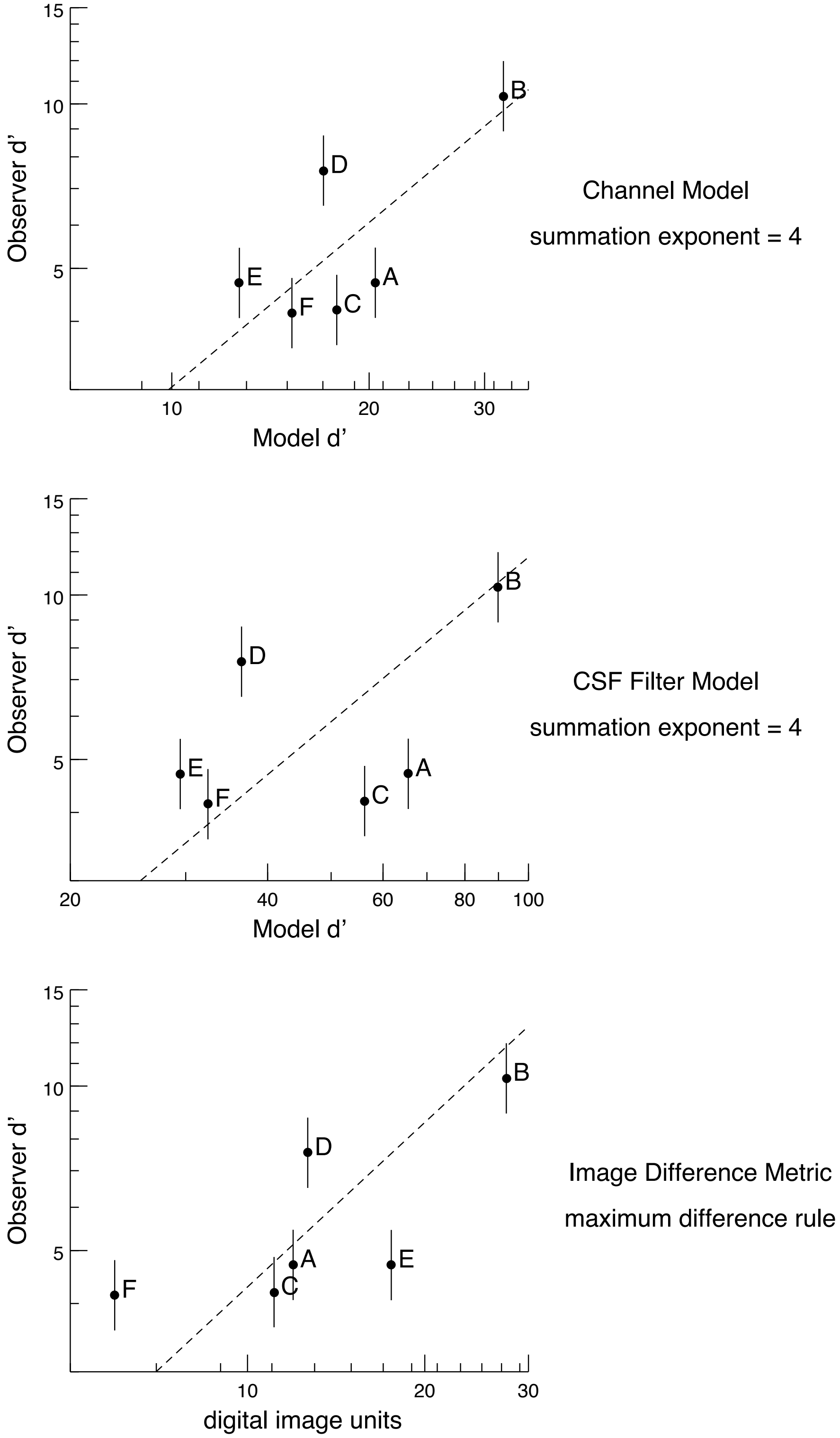
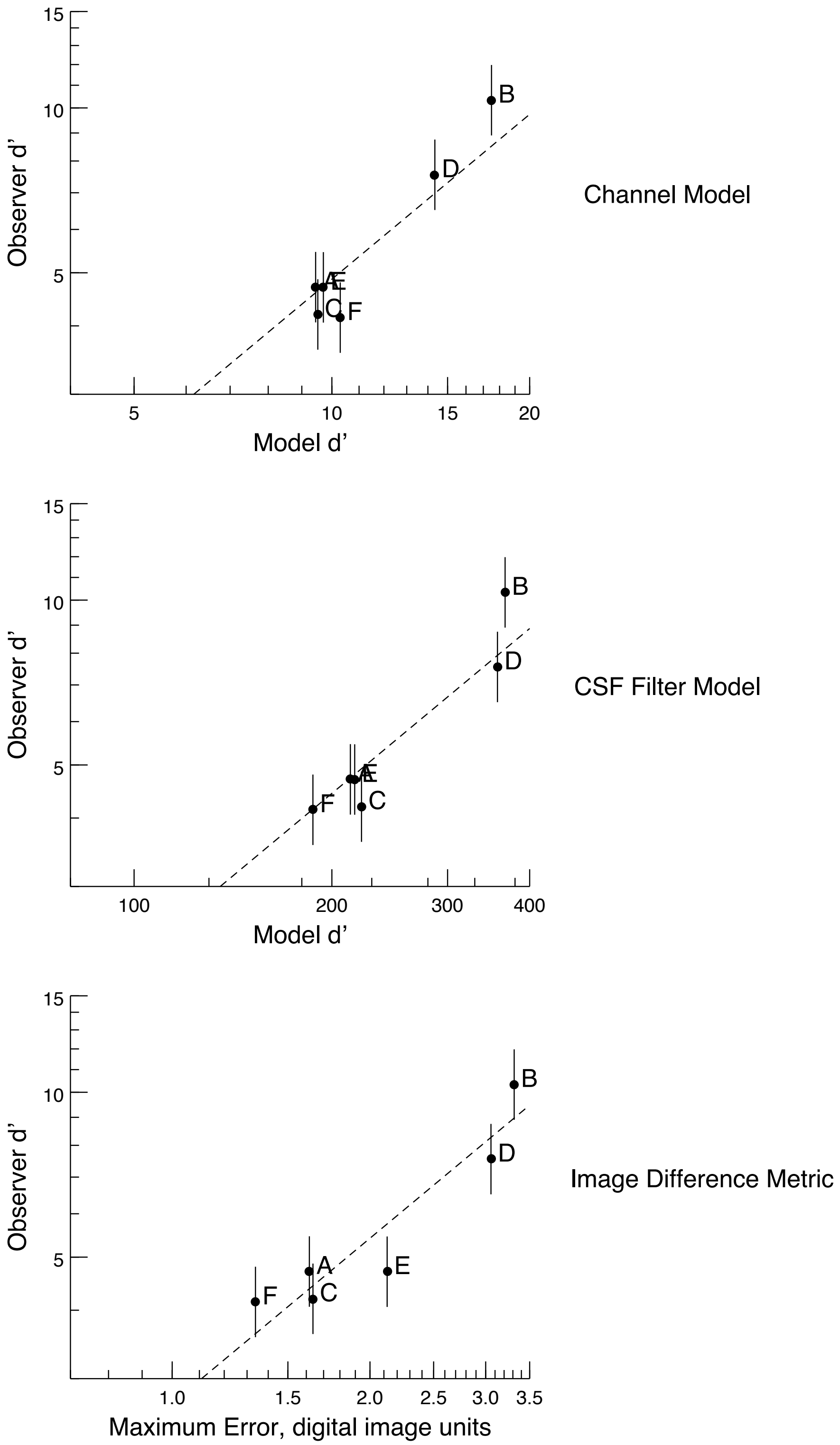


Figure 2. Detection data predictions by three image quality metrics  normalized by contrast.


Observer d'
B
D
A
E
C
F
Model d'
Channel Model
Observer d'
B
D
A
E
F
C
Model d'
CSF Filter Model
Observer d'
B
D
A
E
F
C
Maximum Error, digital image units
Image Difference Metric