Vision Research 37, 3225-3235 (1997).

Object Detection in Natural Backgrounds Predicted by Discrimination Performance and Models

Ann Marie Rohaly*
U.S. Army Research Laboratory, Aberdeen Proving Ground, MD

Albert J. Ahumada, Jr.
NASA Ames Research Center, Moffett Field, CA

Andrew B. Watson
NASA Ames Research Center, Moffett Field, CA

---

*Author to whom correspondence should be sent:
Dr. Ann Marie Rohaly,
U.S. Army Research Laboratory,
Human Research and Engineering Directorate,
ATTN: AMSRL-HR-SC,
Aberdeen Proving Ground, MD 21005-5425.
Office phone (410) 278-5833, FAX (410) 278-8828.
E-mail: amrohaly@arl.mil

Abstract

Many models of visual performance predict image *discriminability*, the visibility of the difference between a pair of images. We compared the ability of three image discrimination models to predict the *detectability* of objects embedded in natural backgrounds. The three models were: a multiple channel Cortex transform model with within-channel masking, a single channel contrast sensitivity filter model, and a digital image difference metric. Each model used a Minkowski distance metric (generalized vector magnitude) to summate absolute differences between the background and object plus background images. For each model, this summation was implemented with three different exponents: 2, 4 and ∞. In addition, each combination of model and summation exponent was implemented with and without a simple contrast gain factor. The model outputs were compared to measures of object detectability obtained from 19 observers. Among the models without the contrast gain factor, the multiple channel model with a summation exponent of 4 performed best, predicting the pattern of observer d's with an RMS error of 2.3 dB. The contrast gain factor improved the predictions of all three models for all three exponents. With the factor, the best exponent was 4 for all three models, and their prediction errors were near 1 dB. These results demonstrate that image discrimination models can predict the relative detectability of objects in natural scenes.

Keywords: object detection, discrimination, modeling, natural images.

Introduction

One important area of applied vision research is the development of methods for assessing the quality of imaging displays for the detection and recognition of objects. We have been developing computer programs to help engineers evaluate the quality of simulated imaging displays for runway obstacle detection by the pilot. The standard modeling approach would be to construct a computer model for object detection and apply it to the simulated images. A general object detection model would simulate search and pattern recognition in the presence of noise and clutter.

Here we evaluate a simple approach that ignores all the visual issues other than masking and takes advantage of the fact that the images are simulated. Our approach is to predict object detection performance by using image discrimination models to predict the visibility of an object added to a fixed background image. Situations can be such that the ignored factors can dominate, but if the discrimination analysis says that the display will not be adequate, the other factors can only make it less adequate.

There are a number of image discriminability models for predicting the visibility of the difference between a pair of images. [For reviews and collections of such models see Ahumada (1993), Watson (1993) and Peli (1995).] We show that discrimination models 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 models that give measures of image discriminability. The first is a multiple spatial frequency channel model based on the Cortex transform with within-channel masking (Watson, 1983, 1987a,b). It is similar to the models of Lubin and Daly (Lubin, 1993; Daly, 1993). The second is a single channel contrast sensitivity function (CSF) filter model. The third model bases its predictions simply on the difference between the digital images. Each model was tested with three different Minkowski summation (generalized vector magnitude) exponents: 2, 4, and ∞. The exponent of 2 corresponds to the familiar Euclidean distance metric, the exponent of 4 to an approximation to probability summation (Quick, 1974) and the exponent of ∞ to the maximum or peak absolute difference.

The multiple channel models referred to above treat the various spatial frequency channels as operating independently of each other. Recent research investigating masking of one spatial frequency component by another has found interactions that can be explained by assuming that the contrast gain of a channel is reduced by activity in other channels (Foley, 1994). New versions of multiple channel models have incorporated lateral interactions among channels in an effort to account for this between-frequency masking (Teo & Heeger, 1994a,b, 1995; Watson & Solomon 1997). These interactions provide a contrast gain control mechanism to keep the contrast within the limited dynamic range of neural mechanisms. To reduce computational complexity, we use a simple overall contrast gain factor to account for between-frequency masking.

The outputs of the three models were compared to measures of object detectability obtained from a group of 19 observers in an earlier psychophysical experiment. Without the contrast gain factor, the multiple channel model performs much better than the other two. With the factor, the predictions of all three models improved significantly, with greater improvement for the single channel model and the digital image difference metric. As a result, with the contrast gain factor, all three models performed similarly.

Experiment 1: Object detection

Methods

Stimuli. Six original digital color images (O_i, i=1,6) of a vehicle in a natural setting (Fig. 1) were altered to form background images (B_i, i=1,6) by replacing the vehicle with appropriate background imagery from elsewhere in the image. Figure 1 shows a black and white version of an original image. Figure 2 shows cropped versions of all six original and background image pairs. Test images were constructed from each image pair by adding a proportion p of the difference between the original and background images to the background image.

$$T_{p,i} = B_i + p\,(O_i - B_i),\; i = 1, 6.$$

Two of the mixing proportions p were 0 and 1, giving the background and original images, respectively. For each image pair, two more proportions were selected to give test images with moderately detectable vehicles. For image pair 1 (Fig. 2), these proportions (p) were 0.6 and 0.8; for pairs 2, 3 and 4, they were 0.4 and 0.6; and for pairs 5 and 6 they were 0.375 and 0.5. The 510×480 pixel images were presented on a 13 inch Macintosh color monitor at a viewing distance giving 95 pixels per degree of visual angle and an image size of 5.33×5.05 deg. The mean luminance of the images was about 10 cd/m^2. When an image was not present, the screen was filled with random amplitude gray scale pixels, uniformly distributed over the digital domain interval [0, 255].

----Insert Figure 1 about here----

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 (six original images at four levels of object detectability each) on a four-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 1.0 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 2.0 sec.* The total sequence of 480 images was completely randomized separately for each observer.

---

*This experiment was conducted at an earlier date as part of another research project with different objectives. The different stimulus durations represent manipulations necessary to test hypotheses specific to the previous investigation and are not directly relevant to the goals of the present study.

Data Analysis. The distance (d'_i) in discriminability units from each object image to its non-object image was measured in the context of a one-dimensional Thurstone scaling model (Torgerson, 1958). The scaling model had the following assumptions:

1. The presentation of an image generates an internal value that is a sample from a normal distribution with unit variance.
2a. The mean of the distribution generated by a background image $T_{0,i}$ is zero.
2b. The mean of the distribution generated by an original object image $T_{1,i}$ is d'_i.
2c. The mean of the distribution generated by an image $T_{p,i}$ is $p\ d'_i$.
3. Means for different images are the same for different observers except for a multiplicative observer sensitivity factor.
4. Each observer has three fixed criteria that are used to categorize an internal value to one of the four rating responses.

The scaling model for this experiment had five d' parameters (the ratios among the six d' values) for each group of observers, plus one sensitivity factor and three category boundaries for each observer. Each observer in the group tested with the two different stimulus durations was given two sensitivity factors. Parameters were estimated by the method of maximum likelihood. Estimates of d' for the six image sets were computed 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 arbitrarily assigned to be unity.

Results and Discussion

Estimates of d' for the six image sets are shown in Fig. 3 for both groups of observers. For the 10 observer group (open squares), the ratio of the highest observer sensitivity factor to the median factor was 1.5 and to the lowest factor was 3.3. For the 9 observer group (open circles), these ratios were 1.9 and 4.1, respectively. For this group, the sensitivity factors estimated for the two stimulus durations were neither appreciably nor significantly different. In addition, the d' values for the two groups of observers were very similar, both in pattern and average level. The filled circles in Fig. 3 are the geometric means of the two group values and the error bars indicate 95% confidence intervals based on the group by image interaction with five degrees of freedom. Strictly speaking, these are confidence intervals for the difference between the mean for an image and the overall mean, a pattern difference appropriate for comparisons with model predictions of the d' pattern when the average model d' has been forced to fit. Based on this interaction, the estimated standard deviation of the geometric means is a factor of 1.06 or 0.50 decibels (dB; 20 dB = 1 log unit).

----- Insert Figure 3 about here ----

Modeling

Stimuli

Although the observers were presented with color images, the models could only be 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 and were cropped around the central target area to 128×128 pixels. The resulting six 2.7×2.7 deg image pairs are shown in Fig. 2.

Algorithms

Multiple channel model. The multiple channel model calculation for a pair of images had the following steps, approximating first the image display and then the early visual system image processing (Watson, 1983, 1987a,b; Lubin, 1993; Daly, 1993). First the digital images I_0, the background image, and I_1, the object image, were converted to luminance (cd/m^2) images using the monitor calibration function

$$I_j \leftarrow a + b\, I_j^{\gamma}, \quad j = 0, 1\,, \tag{1}$$

with $a = 1.00$, $b = 0.0208$ and $\gamma = 1.5$. In the above calculation, the exponentiation and other operations are applied separately to each pixel of the image. Next the images were converted to luminance contrast by subtracting and then dividing by the background image mean luminance $\bar{I}_0$,

$$I_j \leftarrow (I_j - \bar{I}_0) / \bar{I}_0\,. \tag{2}$$

A contrast sensitivity filter S was then applied to the two contrast images.

$$I_j \leftarrow F^{-1}[S\, F[I_j]]\,, \tag{3}$$

where F and F^{-1} are the forward and inverse Fourier transforms, respectively. Next, the Cortex transform (Watson, 1987a,b) was applied to the images, resulting in 20 images of cortex coefficients, corresponding to the combination of five spatial frequency channels, each spanning 1.0 octave of spatial frequency, and four orientation channels, each spanning 45 deg. Each spatial frequency channel below the highest one was subsampled by a factor of two in each spatial dimension. Here we represent the coefficients for image I_j as $c_{j,k}$, where the index k ranges over four dimensions, one for the spatial frequency octave, one for the orientation, and two for the spatial position within the filtered image. The detectability d_k contributed by each coefficient was then computed by taking the absolute differences of the background and object image coefficients reduced by the background image coefficient,

$$d_k = \frac{|c_{1,k} - c_{0,k}|}{\max(1, |c_{0,k}|^{0.7})}\,. \tag{4}$$

This function accounts for the discriminability of increments in suprathreshold grating contrast, ignoring the "dipper" effect (Legge & Foley, 1980). Finally, d' was given by a Minkowski sum of the individual contributions with summation exponent β,

$$d' = (\sum_k d_k^{\beta})^{1/\beta}. \tag{5}$$

For the case $\beta = \infty$, d' was computed as the largest of the d_k.

Single channel model. For the single channel model, the steps were the same through the image filtering [Eqn. (3)], then the filtered image values were used to compute

$$d_k = | I_{1,k} - I_{0,k} |, \tag{6}$$

where the index k now refers to image pixels. Equation (5) was then used to obtain d'.

Image difference metric. For the image difference metric, Eqn. (6) was applied directly to the digital images and then Eqn. (5) was applied.

Contrast gain factor. Because our multiple channel model is already computationally intensive, we used a relatively simple contrast gain factor to account for masking effects among different spatial frequencies. Our method was to multiply the d' predictions by

$$\frac{1}{\sqrt{1 + (c/c_0)^2}}, \tag{7}$$

where c is the RMS background image contrast passed by the contrast sensitivity filter and c_0 is a parameter estimated from the data. For the digital image difference metric, c is the standard deviation of the background image pixel values.

For each model and summation exponent β, a best fitting c_0 was chosen based on the standard error of prediction in the log sensitivity domain, allowing an arbitrary scale factor. As c_0 becomes small, Eqn. (7) approaches c_0/c. When the best estimate of c_0 was zero, we divided d' by c to obtain a contrast normalized prediction. Also, to compute c, the contrast sensitivity filter was normalized to unity at its peak. Note that an implicit parameter of this contrast gain factor is the size of the region over which contrast is computed. In this case it was a 2.7 deg square (i.e., the entire image). However, recent attempts to measure the spread of masking from background regions to a target found no measurable spread (Snowden & Hammett, 1995; Solomon & Watson, 1995). If masking is truly local, the estimation of the background masking parameter from the entire image would be appropriate only when the background is spatially homogeneous.

Linearization. These models are linearized versions of more general models in which the contrast and masking calculations are done for each image separately (Ahumada, 1987; Girod, 1989). These simplified versions have the property that discriminability is linear in the amount of the difference image that is added to the background. The linearized models thus satisfy the second assumption of the above observer response scaling model (see Experiment 1, Data Analysis). Linearization is accomplished by using the background image luminance to convert from luminance to contrast [Eqn. (2)] and by using the background image for the masking calculations [Eqn. (4)]. The model predictions thus need to be computed for only one level of object detectability.

Contrast sensitivity filter calibration. In general, if the same contrast sensitivity filter is used in different models or in the same model with different summation exponents, different predictions will result for the same input images. We arbitrarily decided to calibrate the models to predict contrast thresholds for 1.33 deg square grating patches at five spatial frequencies centered in each of the five bandpass channels of the multiple channel model.

Instead of using the results of a single contrast sensitivity measurement, we calibrated to the predictions of Barten's CSF formula, whose parameters were adjusted to fit the data from a number of experiments (Barten, 1993).

The contrast sensitivity filters were restricted to have a difference of Gaussian form,

$$S(f) = a_c \exp^{-(f/f_c)^2} - a_s \exp^{-(f/f_s)^2}, \tag{8}$$

where a_c and a_s are the center and surround amplitudes and f_c and f_s are the center and surround high frequency cutoffs. Parameters were estimated by least squares fits to simulation outputs in the log threshold domain ignoring quantization and windowing effects. The contrast sensitivity filters were calibrated separately for each of the six combinations of multiple or single channel model and summation exponent of 2, 4, and ∞. The resulting filters appear in Fig. 4.

In order for a given model to predict the same contrast sensitivity as the summation exponent changes, the contrast gain of the filter must change. The larger the exponent, the larger the gain required, because the summation over space and/or spatial frequency is reduced. Because the calibration gratings had constant area, the single channel model filters have the same shape. The multiple channel model filters, however, have different shapes for the different summation exponents. As the summation exponent increases, the multiple channel model needs more gain at high spatial frequencies because of the increase in the number of channels with spatial frequency.

-----Insert Figure 4 about here-----

Results and Discussion

Without the contrast gain factor. Predictions of the three models for the discriminability in d' units of the object image from the background image for each of the three summation exponents are plotted in the left-hand column of Fig. 5. These least squares predictions of the relative observer discriminabilities were computed in the log domain from the model predictions, assuming only an additive constant (discriminability domain multiplicative factor). Including either constant terms or squared terms in the discriminability domain did not significantly improve the fits.

-----Insert Figure 5 about here-----

The model predictions in Fig. 5 have been shifted vertically by the multiplicative factors needed to predict the average observer discriminabilities. The factors for the exponents 2, 4 and ∞, respectively, are 0.21, 0.30 and 0.27 for the multiple channel model, 0.19, 0.12 and 0.059 for the single channel model and 0.037, 0.17 and 0.43 for the digital image difference metric. The multiple channel model has correction factors closer to 1.0, indicating that its within-channel masking allows it to better predict our observer detection performance when calibrated for contrast detection on a uniform background.

To compare the models' abilities to predict the relative detectability of the different targets, the standard errors of the log predictions shown in Fig. 5 were converted to decibels. The prediction errors in decibels for the exponents 2, 4 and ∞, respectively, are 3.4, 2.3 and

2.6 dB for the multiple channel model; 3.8, 3.6 and 5.2 dB for the single channel model; and 3.7, 3.3 and 3.0 dB for the digital image difference metric. The lack of fit is statistically significant at the 0.05 level by an F test (df = 5, 5) if the prediction error exceeds 1.12 dB. The best performance on this error measure is achieved by the multiple channel model with a summation exponent of $\beta = 4$. The exponent of 4 was also best for the single channel model, which did very poorly with the maximum rule ($\beta = \infty$). The digital image difference rule performed best with the maximum rule.

With the contrast gain factor. To account for general contrast masking effects, the d' predictions were multiplied by the correction factor given by Eqn. (7). For each model and summation exponent β, a best fitting c_0 was estimated by minimizing the standard error of prediction in the log sensitivity domain, allowing an arbitrary multiplicative factor. Thus, the estimated c_0 was not constrained by the average detectability of the targets. The values of c_0 in percent contrast for the exponents 2, 4 and ∞, respectively, were 6.7, 14.2 and 9.2% for the multiple channel model, 0.0, 4.3 and 0.0% for the single channel model and 0.0, 0.0 and 0.0% for the digital image difference metric. As the amount of contrast masking varies inversely with the magnitude of c_0 [Eqn. (7)], the larger values of c_0 for the multiple channel model are expected because it already incorporates within-channel contrast masking.

The predictions of the three models with the contrast gain factor are plotted in the right-hand column of Fig. 5 for each of the three summation exponents. Comparing these predictions to those in the left-hand column reveals that the contrast gain factor improved the relative predictions of all the models for all exponents. Again, the lines representing the model predictions in Fig. 5 have been shifted vertically by the multiplicative factors needed to predict the average observer d'. In this case, the multiplicative factors for the exponents 2, 4 and ∞, respectively, are 0.63, 0.49 and 0.40 for the multiple channel model, 0.036, 0.54 and 0.011 for the single channel model and 0.60, 2.7 and 6.9 for the digital image difference metric. Comparing these scale factors to those corresponding to the un-normalized versions of the models shows that when c_0 was zero, the division by c (background contrast), a number less than one, shifted the scale factor farther from unity. However, when c_0 was nonzero, the scale factors were all closer to unity than they were prior to contrast gain normalization, indicating better prediction of the average d'.

Once again, in order to compare the performance of the various models, the standard errors of the log predictions shown in Fig. 5 were converted to decibels. Note that because a degree of freedom was removed for the estimation of c_0, the errors for the normalized models are not forced to be smaller than those for the un-normalized models. The normalized prediction errors in dB for the exponents 2, 4 and ∞, respectively, are 1.9, 1.3 and 2.0 for the multiple channel model, 1.1, 1.1 and 2.5 for the single channel model and 1.4, 1.1 and 1.4 for the digital image difference metric. The lack of fit is statistically significant at the 0.05 level by an F test (df = 4, 5) if the prediction error exceeds 1.14 dB. Thus, with the addition of the contrast gain factor, all three models provided better predictions of the relative detectability of the targets and the models were essentially equivalent with their best summation exponent, $\beta = 4$.

For an exponent of 4, the multiplicative factors given above indicate that the normalized models still mispredict the observer data by a factor near 2.0. A possible explanation lies in the fact that the observer data were obtained in a detection experiment while the models being tested are models of image discrimination. To examine this possibility, additional data were collected in a discrimination experiment.

Experiment 2: Object discrimination

Methods

Stimuli. The stimuli for this experiment were the lower resolution gray scale images that were used as input to the models (Fig. 2). Test images were constructed from each image pair as in Experiment 1, with the difference that mixing proportions of 0.0, 0.1, 0.2 and 0.4 were used for all six image pairs. The images were presented on a 15 inch Sony monitor using a look-up table to match the luminance and gamma of the monitor used in the object detection experiment. Because the resolution of the images used in this experiment was lower by a factor of two than that of the previous experiment, the viewing distance was set to give 47.5 pixels per degree to equate the spatial frequency content of the two sets of images.

Observers. Three non-military observers participated in this experiment. They were all near 30 years of age and had been refracted within two months of the experiment to normal acuity.

Procedure. As in the detection experiment, the observers were asked to rate each of the 24 images (six original images at four levels of object detectability each) on a four-point rating scale. For this experiment, the ratings were as follows:

1-Definitely the non-vehicle image.
2-Probably the non-vehicle image.
3-Probably the vehicle image.
4-Definitely the vehicle image.

Instead of presenting all the images in one completely randomized sequence (as in Experiment 1), mixture sets based on each of the six original images were presented in separate blocks so the observers could respond to any visible difference and not only rely on those that contributed to the detection of the vehicle.

Trials were run in blocks of 60, using one vehicle and its background. Each of the four images composing a mixture set (no vehicle, 10%, 20% and 40% vehicle) was presented with probability 0.25. Before each set of 10 trials, the 100% vehicle image was shown to the observer as a memory aid. Six blocks of trials were run for each of the six mixture sets in a 6×6 Latin square design, randomized separately for each observer. The image duration was 1.0 sec.

Data Analysis. The data were analyzed in the context of the same Thurstone scaling model used to analyze the data of Experiment 1, except that in the parameter estimation, each observer had separate d's and criterion values were estimated separately for different blocks of trials. The resulting discriminability parameter estimates were scaled to represent the distance (d') from the 100% vehicle image to the non-vehicle image.

Results and Discussion

The geometric means of the three observers' d' estimates are shown in Fig. 6 along with the mean d' estimates obtained in Experiment 1. The ordering of d's with respect to the six image pairs is similar for both experiments. The standard deviation of prediction of the detection d's from the discrimination d's in the log domain gives a prediction error of 1.6 dB.

The multiplicative factor for predicting the detection results from the discrimination results is 0.52. This factor is close to that needed to scale the contrast-gain-normalized multiple and single channel model predictions to fit the observer data. Thus, it can be regarded as the factor needed to correct for the difference between detection and discrimination in this situation.

----Insert Figure 6 about here----

General Discussion

The detectability of targets in natural scenes was measured and compared to the predictions of three image discrimination models. The models were able to predict the psychophysical measures of target detectability for our 6 target/background combinations. This result suggests that when search is removed from the detection task, performance is limited by those target properties and background masking properties that are accounted for by these models. Note also that the models were able to predict the psychophysical data despite the fact that they were presented with gray scale images while the human observers were presented with color images. This implies that, for our image set, chromatic information was not a major factor in the detection of the target.

Models without contrast gain. The simplest discrimination metric considered was the Minkowski distance metric applied to the images in the digital domain. For a summation index of 2, this metric is the RMS error metric used by the digital image processing community. The best exponent for this metric was ∞, turning the metric into the maximum absolute difference between the digital image pixels. This result, suggesting no spatial summation, was probably helped by our coarse sampling of exponent values and by the low pass filtering done in the course of lowering the resolution of the images presented to the models. This metric outperformed the single channel model, which takes into account the contrast sensitivity of the observer. Although contrast sensitivity must in general be taken into account (Girod, 1993), others have also found no advantage for adding the complexity of a contrast sensitivity filter (Farrell *et al.*, 1991). Here, the digital image difference was better than the single channel model at predicting the relative detectability of our 6 target/background combinations, regardless of the summation exponent. This version of the single channel model can be regarded as a visible luminance contrast difference metric. The advantage of the digital value metric over the visible contrast metric was probably abetted by the aforementioned low pass filtering of the images and might reflect that, because of the gamma function of the display, the digital values are closer to a JND intensity scale than are the luminance contrast values.

The multiple channel model with contrast gain may have outperformed these simpler models because the within-channel intensity scale is closer to a discriminability scale. Another possibility is that the channels are the appropriate domain for summing the differences generated by the target. Evidence for this being the larger effect can be found in the pattern of errors for the summation exponents 2 and 4. For the exponent of 2, summation is based on the Euclidean distance and is relatively unaffected by the channel representation. The intensity scaling is present in the multiple channel model and there is only a 0.4 dB difference in favor of the multiple channel over the single channel model. When the summation exponent was changed to 4 so that the channel representation matters, the single channel model improved by only 0.2 dB while the multiple channel model improved by 1.1 dB.

The digital difference metric does not predict the average level of detection. Without the contrast gain factor, the single channel model predicts that the detectability of a target depends only on its visible contrast and not on the contrast in the background. In these high contrast backgrounds, it badly overpredicts the average target detectability. Contrast within a channel of the multiple channel model does reduce the sensitivity for differences within that channel, so the background contrast does reduce the predicted detectability of a target. The multiple channel model thus does better than the single channel model at predicting the average detectability. However, this effect is not strong enough to predict the masking of our natural backgrounds and the model overpredicts the average detectability.

Models with contrast gain. To say that channel models need to include a contrast gain adjustment that depends on input from other channels is to say that the JND scale for a channel depends on the activity in other channels as well. If the contribution from other spatial frequency channels is relatively independent of spatial frequency, and the background is spatially homogeneous, interchannel interactions can be approximated by a simple contrast gain factor. The addition of a contrast gain factor improved the predictions of all three models, masking the advantages of both contrast sensitivity filtering and the spatial frequency channel representation. As a result, the best predictor of the relative detectability of our 6 target/background combinations was the generalized vector length of the difference image divided by the background image standard deviation. This simple measure, however, does not predict the average level of target detectability. The gain control parameter estimated to optimize the prediction of the pattern of detection differences also allowed the single and multiple channel models to accurately predict the average level of target discriminability in the second experiment, and thus overpredict the average target detectabilities by a factor of two.

Conclusions

Discrimination models designed to answer, "Are these two images different?" can predict the answer to the question, "Is there an object in this image?" When the effects of general contrast masking were not taken into account, a multiple channel model performed better than either a single channel model or a digital image difference metric at predicting both the relative (between the six images) and average levels of target detectability. When general contrast masking effects were included, however, the relative predictions of all three models improved to the same level. Visual transformations of the digital images were not needed to predict the relative detectability or discriminability. The two visual models were calibrated to predict grating detection on a uniform background. With general contrast masking, both the single and multiple channel models predicted the average discriminability.

Acknowledgements

R. Horng wrote and ran the MatLab program that generated the model and metric predictions. Parts of this work have been reported previously (Ahumada, Watson & Rohaly, 1995a,b; Ahumada, Rohaly & Watson, 1995; Rohaly, Ahumada & Watson, 1995). This work was supported in part by NASA RTOPs 505-64-53 and 537-08-20 and NASA Cooperative Agreement NCC 2-307 with Stanford University.

References

Ahumada, A.J., Jr. (1987). Putting the noise of the visual system back in the picture. *Journal of the Optical Society of America A, 4,* 2372-2378.

Ahumada, A.J., Jr. (1993). Computational image quality metrics: a review. *SID Digest, 24*, 305-308.

Ahumada, A.J., Jr., Rohaly, A.M. & Watson, A.B. (1995). Image discrimination models predict object detection in natural backgrounds. *Investigative Ophthalmology and Visual Science, 36 (Suppl.),* 439.

Ahumada, A.J., Jr., Watson, A.B. & Rohaly, A.M. (1995a). Models of human image discrimination predict object detection in natural backgrounds. In Rogowitz, B. & Allebach, J. (Eds), *Human Vision, Visual Processing, and Digital Display VI, SPIE 2411*, 355-362.

Ahumada, A.J., Jr., Watson, A.B. & Rohaly, A.M. (1995b). Object detection in natural backgrounds predicted by discrimination performance and models. *Perception, 24 (Suppl.),* 7.

Barten, P.G.J. (1993). Spatiotemporal model for the contrast sensitivity of the human eye and its temporal aspects. In Rogowitz, B. & Allebach, J. (Eds), *Human Vision, Visual Processing, and Digital Display IV, SPIE 1913,* 2-14.

Daly, S. (1993). The visible differences predictor: an algorithm for the assessment of image fidelity. In Watson, A.B. (Ed.), *Digital Images and Human Vision* (pp. 179-206). Cambridge, Mass.: MIT Press.

Farrell, J.E., Trontelj, H., Rosenberg, C. & Wiseman, J. (1991). Perceptual metrics for monochrome image compression. *Society for Information Display Digest of Technical Papers, 22,* 631-634.

Foley, J.M. (1994). Human luminance pattern-vision mechanisms: masking experiments require a new model. *Journal of the Optical Society of America A, 11,* 1710-1719.

Girod, B. (1989). The information theoretical significance of spatial and temporal masking in video signals. In Rogowitz, B.E. (Ed.), *Human Vision, Visual Processing, and Digital Display, SPIE 1077,* 178-187.

Legge, G.E. & Foley, J.M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America, 70,* 1458-1471.

Lubin, J. (1993). The use of psychophysical data and models in the analysis of display system performance. In Watson, A.B. (Ed.), *Digital Images and Human Vision* (pp. 163-178). Cambridge, Mass.: MIT Press.

Peli, E. (1995). *Vision Models for Target Detection and Recognition*. New Jersey: World Scientific Publishing.

Quick, R.F. (1974). A vector magnitude model of contrast detection. *Kybernetik, 16,* 65-67.

Rohaly, A.M., Ahumada, A.J., Jr. & Watson, A.B. (1995). A comparison of image quality models and metrics predicting object detection. *Society for Information Display Digest of Technical Papers, 26,* 45-48.

Snowden, R.J. & Hammett, S.T. (1995). The effect of contrast surrounds on contrast centres. *Investigative Ophthalmology and Visual Science, 36 (Suppl.),* 438.

Solomon, J.A. & Watson, A.B. (1995). Spatial and spatial frequency spreads of masking: measurements and a contrast-gain-control model. *Perception, 24 (Suppl.),* 37.

Teo, P.C. & Heeger, D.J. (1994a) Perceptual image distortion. In Rogowitz, B. & Allebach, J. (Eds), *Human Vision, Visual Processing, and Digital Display V, SPIE 2179,* 127-141.

Teo, P.C. & Heeger, D.J. (1994b) Perceptual image distortion. *Proceedings of ICIP-94,*

Volume II (pp. 982-986). Los Alamitos, Calif.: IEEE Computer Society Press.
Teo, P.C. & Heeger, D.J. (1995) A general mechanistic model of spatial pattern detection. *Investigative Ophthalmology and Visual Science*, *36 (Suppl.)*, 438.
Torgerson, W.S. (1958). *Theory and Methods of Scaling*. New York: Wiley.
Watson, A.B. (1983). Detection and recognition of simple spatial forms. In Braddick, O.J. & Sleigh, A.C. (Eds), *Physical and biological processing of images* (pp. 100-114). Berlin: Springer-Verlag.
Watson, A.B. (1987a). The Cortex transform: rapid computation of simulated neural images, *Computer Vision, Graphics, and Image Processing*, *39*, 311-327.
Watson, A.B. (1987b). Efficiency of an image code based on human vision. *Journal of the Optical Society of America A, 4*, 2401-2417.
Watson, A.B. (1993). *Digital Images and Human Vision*. Cambridge, Mass.: MIT Press.
Watson, A.B. & Solomon, J.A. (1995). Contrast gain control model fits masking data. *Investigative Ophthalmology and Visual Science, 36 (Suppl.)*, 438.

Figure Captions

FIGURE 1. A black and white version of one of the six original color images used as a test image in Experiment 1. The image shows a vehicle in a natural setting.

FIGURE 2. The set of six image pairs presented to the discrimination models. Each image pair comprises an object image (vehicle plus background) and a non-object image (background only). These 128×128 pixel, gray scale images were obtained from the 510×480 pixel, RGB color images used in the detection experiment (Experiment 1) as explained in the text. The images are numbered in order of increasing object detectability (d') based on the results of Experiment 1 (Fig. 3).

FIGURE 3. Discriminability indices (d') for the six image pairs of Fig. 2 estimated from the data of Experiment 1. Values estimated for the 10 observers given 1.0 sec durations are plotted as open squares while those estimated for the 9 observers given the 0.5 and 2.0 sec durations are plotted as open circles. Error bars represent 95% confidence intervals for the mean of the two groups of observers based on the variance between the groups. The geometric means of the two group values, shown as filled circles, were used to gauge model performance.

FIGURE 4. Contrast sensitivity filters for the multiple channel (solid lines) and single channel (dotted lines) models obtained from the calibration to Barten's (1993) contrast sensitivity equation. The uppermost pair of curves are the filters for a summation exponent of ∞, the middle pair are for an exponent of 4 and the lowest pair for an exponent of 2. Note that with the weaker summation rules (larger exponents), more gain is required for the models to predict a given contrast sensitivity.

FIGURE 5. Model predictions. Mean observer detectabilities (d's) for the six image pairs (see Fig. 2) are plotted as points with error bars corresponding to 95% confidence intervals based on the variance between the two groups of observers. The same data are plotted in each panel to allow comparison with the predictions of each model. The multiple channel predictions are shown in the top panel, the single channel predictions in the middle panel and the digital image difference predictions in the bottom panel. The left column contains the predictions obtained when the contrast gain factor was not used while the right column contains the predictions including the factor. For each model, predictions are shown for each of the summation exponents tested: 2 (solid lines), 4 (dashed lines) and ∞ (dotted lines). The lines representing the model predictions have been shifted vertically by multiplicative factors (given in the text) needed to correctly predict the average of the 6 detectabilities.
Without the contrast gain factor (left column), the multiple channel model with an exponent of 4 (top panel, dashed line) provides the best fit to the detection data. When the contrast gain factor is included (right column), however, the predictions of all three models improve for all exponents. In this case, the predictions of three models fit the detection data equally well for an exponent of 4 (dashed lines).

FIGURE 6. Comparison of d' values obtained from the detection and discrimination experiments. Average d' values for the 19 observers who participated in the detection experiment (Experiment 1) are plotted against the average d' values for the three observers

who participated in the discrimination experiment (Experiment 2). Error bars represent 95% confidence regions based on the pooled standard deviations. The number next to each data point denotes the corresponding image pair (see Fig. 2). The dashed line in the graph is a best-fitting line of unit slope in log-log coordinates. The y-intercept of this line, 0.52, is the multiplicative factor for predicting the detection data from the discrimination data. This factor is close to the multiplicative factors needed to allow the contrast-gain-normalized multiple and single channel model to predict the average detectability of the six target/background combinations.