Pergamon

0042-6989(95)00071-2

Vision Res. Vol. 35, No. 21, pp. 2955–2966, 1995

0042-6989/95 $9.50 + 0.00

Human Efficiency for Recognizing and Detecting Low-pass Filtered Objects

WENDY L. BRAJE,*† BOSCO S. TJAN,* GORDON E. LEGGE*

Received 11 July 1994; in revised form 4 January 1995; in final form 6 March 1995

Recently, Tjan, Braje, Legge and Kersten [(1995) *Vision Research, 35*, 3053–3069] found that human efficiency for object recognition was less than 10%, indicating that humans fail to use much of the information available to an ideal observer. We examine two explanations for these low efficiencies: (1) humans are inefficient in using high spatial-frequency information; and (2) humans are inefficient in detecting image samples. We tested the first possibility by measuring human efficiency for recognizing low-pass filtered objects, rendered as line drawings and silhouettes, in luminance noise. Efficiency did not improve when high frequencies were removed, and the first explanation was rejected. We tested the second explanation by comparing efficiencies for object detection and recognition. Recognition efficiency was higher than detection efficiency for silhouettes but not line drawings, showing that detection efficiency does not place a ceiling on recognition efficiency. The results indicate that human vision is designed to extract image features, such as contours, that enhance recognition. A computer simulation suggests that this can occur if the observer views the world through a band-pass spatial-frequency channel.

Efficiency Object recognition Object detection Spatial-frequency filtering

INTRODUCTION

Efficiency is a measure that compares the performance of a human observer to the performance of an ideal observer. The "ideal observer" is one that uses all of the available information to maximize performance. Recently, Tjan, Braje, Legge and Kersten (1995) measured human efficiency for recognizing simple 3-D objects in luminance noise, and found that efficiencies ranged from 2.7 to 7.8%. These values are low compared to efficiencies obtained in several other complex perceptual tasks: 25% for judging pattern symmetry (Barlow & Reeves, 1979), 60% for estimating the mean values of scatter plots (Legge, Gu & Luebker, 1989), and 42% for recognizing band-pass filtered letters (Parish & Sperling, 1991).

The low efficiencies for object recognition indicate that people fail to use much of the information available to an ideal observer. Tjan *et al.* (1995) identified three factors that play important roles in accounting for low recognition efficiencies: stimulus size, spatial uncertainty, and detection efficiency. Other factors found to play smaller roles were the observer's internal noise, rendering condition (silhouette, line drawing, or Lambertian shading), learning, and categorization across viewpoints. The highest efficiencies measured, however, were only about 13.5%—for small line drawings with spatial uncertainty (Tjan, Braje & Legge, 1994).

In this paper, we examine the possibility that humans fail to use information available in the high spatial frequencies of the images. We also explore in greater detail the possibility that low recognition efficiency is a consequence of low detection efficiency.

Recognition of low-pass filtered objects

Is human efficiency low because people do not use information contained in the high spatial frequencies? We refer here to *object spatial frequency*, rather than to retinal frequency. Object frequency is expressed in cycles/object-height and is independent of viewing distance. Evidence suggests that object frequency may be a more important determinant of performance than retinal frequency in complex tasks such as letter recognition (Parish & Sperling, 1991) and reading (Legge, Pelli, Rubin & Schleske, 1985).

Several studies have examined the relative importance of different spatial-frequency bands for recognition. Ginsburg (1980) suggested that low frequencies are sufficient for letter and face recognition, and that high frequencies are redundant. Legge *et al.* (1985) used low-pass filtered text to show that low letter frequencies (2 cycles/letter) are sufficient for rapid reading.

Although low frequencies may be sufficient for certain tasks, other studies have shown that useful information is available at higher frequencies. For face recognition, Fiorentini, Maffei and Sandini (1983) found

*Department of Psychology, University of Minnesota, 75 East River Road N218, Minneapolis, MN 55455, U.S.A.
†To whom all correspondence should be addressed [*Email* braje@eye.psych.umn.edu].

that accuracy increased when high frequencies (>5 cycles/face-width) were included, compared with low frequencies only. They also found that the removal of lower frequencies did not hurt performance, suggesting that high frequencies were sufficient for face recognition. Measuring recognition of band-pass filtered letters, Parish and Sperling (1991) found that contrast thresholds decreased as the center frequency of the band was increased up to 4.22 cycles/letter-height, and decreased even further with high-pass filtered letters. Norman and Ehrlich (1987) showed that the addition of either low or high frequencies to a stimulus reduced the error rates and shortened the reaction times for identifying pictures of toy tanks, suggesting that both low and high frequencies provide useful information.

These studies suggest that signals for recognition may be available in high-frequency bands. However, the envelope of the amplitude spectra of most objects drops with increasing spatial frequency. Thus, while these signals may be useful when presented in isolation, their signal-to-noise ratios may be too low to be useful in real, unfiltered images. Furthermore, efficiency may provide a more direct probe of the usefulness of information in different bands than other measures.

The preponderance of evidence indicates that humans rely on, and are more efficient in the use of, low object frequencies in recognition. But some evidence suggests that humans can use high frequencies, and that high-frequency information can sometimes improve recognition performance. There are two ways in which the high-frequency content of images might aid recognition. First, the high frequencies may contain non-redundant cues, and thus additional information, for recognition. In many cases, however, high-frequency features are correlated with low-frequency features, thereby providing little additional information. Second, even if the information content of the high bands is redundant, looking in several bands can improve performance by increasing the likelihood of detecting the information. This would be the case if the signals were correlated but the limiting noise was uncorrelated. For example, Watt and Morgan (1985) have proposed that the outputs of independent channels are averaged to increase signal-to-noise ratio.

If humans do not use high frequencies for object recognition, then removing these frequencies should have no effect on their recognition performance. On the other hand, if high frequencies do contain useful information for object recognition, an ideal observer's performance will decline when these frequencies are removed. The overall effect of removing high frequencies would be to reduce the difference between human and ideal performance, and hence increase human efficiency. In our first experiment, we tested this possibility by measuring efficiency for recognizing low-pass filtered objects.

Detection of low-pass filtered objects

Our second purpose was to determine if detection efficiency limits recognition efficiency. Detection efficiency can be conceptualized as a limitation on the number of signal samples encoded and used for detection. Such sampling may be done in the spatial domain (e.g. image pixels), or in any other domain (e.g. spatial frequency). An ideal observer uses all of the signal samples. If humans use only a subset of the signal samples, then their detection efficiency will be less than 100% (see Appendix B in Tjan *et al.*, 1995). For example, if humans encode and fully utilize 10 samples of a signal containing 100 equal energy samples, their detection efficiency will be no higher than 10%.* Expressing efficiency as subsampling relates closely to Fisher's original definition of statistical efficiency (Fisher, 1925).

The link between recognition efficiency and detection efficiency depends on the nature of the subsampling process. Consider, for example, two different tasks: (1) detecting a 10 by 10-pixel square; and (2) discriminating this same square from a second signal, identical to the square except for one missing pixel in a known location. All pixels are equally informative for the detection task, because each pixel discriminates the signal from a blank screen. For the recognition task, however, only the missing pixel is informative; the other 99 pixels do not distinguish the two signals. Suppose a primitive detector can encode data from just 10 samples (pixels) per trial. As described above, the detection efficiency of this device will be no more than 10%. If the detector encodes a random set of 10 samples from the target area on each trial, its recognition efficiency will be limited to 10%. Because of the random sampling, the odds of encoding the recognition feature (i.e. the missing pixel) are only one in ten. In this case, recognition efficiency is limited by detection efficiency.

Alternatively, suppose the device encodes its 10 samples strategically, always including the recognition feature. Although its detection efficiency is still limited to 10%, recognition efficiency could, in principle, be 100%. This example demonstrates that the sets of image samples (i.e. features) that determine detection and recognition efficiency are usually different. The relationship between recognition and detection efficiency will depend on the sampling "strategy" adopted by the visual system. Human vision may be designed to extract information useful for object recognition, rather than detection.

In the second experiment, we compared recognition and detection efficiencies for the same stimuli used in the first experiment. The key question is whether recognition efficiencies can exceed detection efficiencies. We also examined the specific possibility of edge encoding by measuring detection and recognition efficiencies for line drawings and silhouettes of objects. If the visual system uses an edge-sampling strategy, several predictions can

*Burgess and Colborne (1988) have shown that internal, signal-dependent (i.e. multiplicative) noise manifests itself as a reduction in sampling efficiency. If multiplicative noise is present, empirical estimates of sampling efficiency provide an upper bound on efficiency related to sampling.

be made. First, detection efficiency should be higher for line drawings than for silhouettes, assuming that the visual system makes equally efficient use of a step-edge (silhouette) and a line edge (line drawing). Second, for line drawings, sampling must necessarily be confined to edges, regardless of the task. Edge sampling should therefore yield nearly the same efficiencies for recognition and detection of line drawings. Finally, for silhouettes, an edge-sampling scheme should result in higher efficiencies for recognition than detection.

METHODS

Apparatus and stimuli

The apparatus and stimuli are described in detail by Tjan *et al.* (1995). Briefly, targets were presented on one Apple monochrome monitor, and noise on another, allowing for independent control of contrast. The images on the two monitors were superimposed optically, with a viewing distance of 1.72 m. Accurate contrast control was achieved with video attenuators and the Video-Toolbox software (Pelli & Zhang, 1991).

The targets were the same four 3-D objects used by Tjan *et al.* (1995), referred to as *wedge*, *cone*, *cylinder*, and *pyramid*. The objects were rendered in orthographic projection on a Stardent 2000 graphics computer, each from 8 viewpoints randomly selected from a viewing sphere. They were rendered as bright silhouettes or line drawings on a dark background. The entire image field was 452 pixels horizontally by 442 pixels vertically, corresponding respectively to 5 and 4.9 deg. The target objects occupied the central 256 by 256 pixels, subtending 2.8 deg.

The targets were digitally filtered with first-order exponential low-pass filters (see Fig. 1) using the HIPS software (Landy, Cohen & Sperling, 1984). The filters had $1/e$ bandwidths of 0.5, 1.5, 3.6, and 42 cycles per average object-height (cy/ob). Unfiltered targets, with a Nyquist frequency of 128 cy/ob, were also tested. Examples of these stimuli are shown in Fig. 2.

The targets were optically superimposed with unfiltered 2-D static Gaussian luminance noise. The two-sided vertical and horizontal bandwidths of the noise were 91.4 cy/deg (256 cy/ob). The noise had a mean luminance of 7.1 cd/m^2 and a standard deviation of 3.6 cd/m^2, measured at the subject's eye. The corresponding spectral density (noise energy per unit bandwidth) was 27.0 μ(deg^2) [$3.5*10^{-6}$ ob^2, or 3.5 μ(ob^2)].

After combination of the target and noise screens by the optical apparatus, the mean luminance of the background field was 7.5 cd/m^2, and the luminance of the brightest target pixel ranged from 7.6 to 9.0 cd/m^2, depending on signal energy. Viewing was monocular.

Procedure

There were twelve experimental conditions, each consisting of one of the two rendering conditions and one of the six filtering conditions. Each block of recognition trials was devoted to one of these twelve conditions. The blocks were presented in the same random order to each subject. Before each block of trials, observers were shown all images that would be presented in that block. The images were shown without noise, once at a high signal energy, and once at the starting signal energy of the experiment.

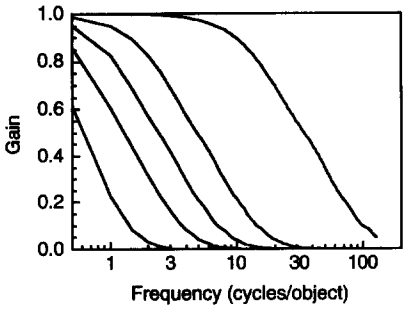

FIGURE 1. Five exponential low-pass filters were used to create the stimuli. 1/e bandwidths were 0.5, 1.5, 3.6 and 42 cy/ob. Unfiltered images have a Nyquist frequency of 128 cy/ob. The filters are of the form $e^{-f/c}$, where f is the input frequency and c is the bandwidth of the filter (i.e. the frequency at which the gain is 1/e).

On a recognition trial, one of the 32 images was randomly selected and displayed on the monitor for one second. The observer's task was to indicate which of the four objects had been presented by pressing one of four keys. The observer was not required to indicate which of the eight object views had been presented. No feedback was given. Between trials, the observer saw a uniform screen with no noise and with a luminance of 7.5 cd/m^2. The beginning of a new trial was signaled by a brief tone.

An adaptive staircase was used to estimate the threshold signal-to-noise ratio (E/N), defined as the signal-to-noise ratio at which subjects obtained 79% correct recognition (Wetherill & Levitt, 1965). The "signal" refers to signal energy (E), equal to the squared RMS contrast multiplied by the image area; the "noise" refers to noise spectral density (N), or noise energy per unit bandwidth. Signal energy was changed by increasing or decreasing the peak luminance of the object screen by 0.22 dB (5%). The noise spectral density was not changed. The staircase terminated after 14 reversals, and the threshold E/N was estimated as the mean E/N of the last 12 reversals. Each staircase contained roughly 100 trials. Three staircases were run for each experimental condition except the 0.5 cy/ob bandwidth conditions.* The thresholds obtained from each staircase in a condition were averaged.

In the detection experiment, the observer performed a "yes/no" task. A "blank" image (a uniform field of 7.5 cd/m^2) was presented in noise for one second on a random one-half of the trials, and object images, drawn at random from the set of 32, were presented on the other half of the trials. Between trials, the observer saw a uniform screen with no noise and with a luminance of 7.5 cd/m^2. The observer indicated whether or not an

*For the 0.5 cy/ob conditions, the recognition task was extremely difficult, and the thresholds were very high. Thus only one staircase was run for each rendering condition at this bandwidth.

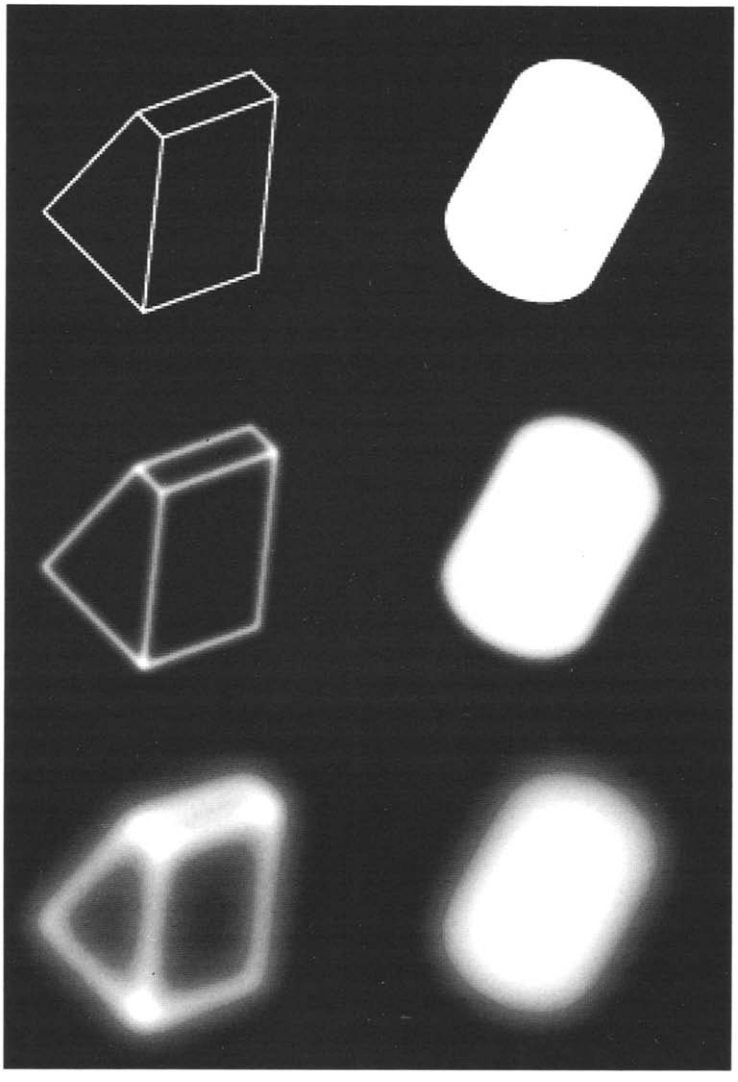

FIGURE 2. A wedge rendered as a line drawing (left column) and a cylinder rendered as a silhouette (right column), shown without noise. From top to bottom, objects are shown unfiltered, low-pass filtered with a 6 cy/ob filter, and low-pass filtered with a 1.5 cy/ob filter.

object had been presented (but not which object) by pressing one of two keys. No feedback was given. Although we did not take into account the decision criterion used by the human observers, we were able to determine that any response bias present did not greatly affect our results or alter our conclusions.* The beginning of a new trial was signaled by a brief tone. The detection procedure employed the same staircase procedure that was used in the recognition experiment. Three such staircases were run for each experimental condition, and the thresholds obtained in each condition were averaged.

*If we make the simplifying assumption that the task is to discriminate between two equal-variance Gaussian distributions, then an unbiased observer's proportions of "yes" and "no" responses will equal the *a priori* probabilities of the respective presence and absence of the signal, 0.5 in our case. This was the case for observer BT but not WB. WB responded "yes" 37% of the time and "no" 63% of the time on average, indicating a conservative bias. This bias had only a very small effect on efficiency. First, the efficiencies of the two observers were very similar. Second, given our assumption, the ratio of the unbiased efficiency of WB to that of BT is equal to the squared ratio of d' (i.e. the difference between the z-scores of hit and false-alarm rates, which provides a criterion-free measure of sensitivity) for each subject. This squared d' ratio was 1.01. Knowing that BT was unbiased, we can compute the unbiased efficiency of WB by multiplying BT's measured efficiency by this squared d' ratio. When we do this, we find that WB's unbiased efficiency is only 1.02 times the measured efficiency, which is not large enough to alter any of our conclusions.

Human observers

Two of the authors served as subjects. Both had normal or corrected vision with Snellen acuity in the tested eye of 20/20. Both were very familiar with the stimuli and were highly practiced on the tasks.

Ideal observer

The measurement of statistical efficiency involves the comparison of human performance to ideal performance. The ideal observer uses an algorithm that is optimal in the sense that it uses all of the available information to maximize performance on a particular task. The comparison is usually meaningful only when the ideal observer's performance is held below 100% correct by some source of uncertainty in the stimulus. In

our experiments, uncertainty was introduced by adding luminance noise to the stimuli.

We used the same ideal observer described in detail by Tjan *et al.* (1995). The ideal observer was formulated as a modified template matcher. For the recognition task, it stores the set of 32 2-D orthographic image projections used in the experiment (4 objects, each with 8 views). The appropriate filtered templates were used in our experiments. In a simulated trial, the ideal observer is presented with one of the images plus luminance noise, and it compares this noisy image with each of the 32 templates. For each object, the ideal observer computes the *a posteriori* probability of that object by summing the *a posteriori* probabilities of its eight views. It then responds with the object with maximum probability. Note that this is different from selecting the object that corresponds to the single template providing the best match, which would be a sub-ideal decision rule. Mathematically, the ideal observer chooses the object that maximizes the following function, which is monotonic to a likelihood function:

$$L'(i) = \sum_{j=1}^{8} \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{p}[R_k - T_{ij_k}]^2\right];$$

where

R = input image as an array of contrast values,
T_{ij} = template of object i at view j,
σ = standard deviation of noise contrast,
p = number of pixels per image.

The detection experiment is formulated as a special kind of recognition experiment. There are two target categories, an *object* and a *blank*. The *object* has 32 "views," corresponding to the 32 object images used in the recognition experiment. The *blank* has one image, a uniform gray screen, which occurs on roughly half of the trials. In a simulated trial, the ideal observer is presented with either an object or blank in luminance noise. It computes the *a posteriori* probability of an object by summing the *a posteriori* probabilities of its 32 views, and, likewise, the *a posteriori* probability of a blank. In the above equation, i ranges from 1 to 2 (*object* or *blank*), and j ranges from 1 to 32. The ideal observer's task is to choose the category (*object* or *blank*) that has the higher *a posteriori* probability.

The ideal observer used a binary search algorithm (see Tjan *et al.*, 1995), adjusting the signal energy iteratively, to find its threshold for a given task. As with the human observers, the ideal observer's threshold is the signal-to-noise ratio (E/N) at which the algorithm yields 79% correct.

Efficiency

The threshold E/N was determined for human and ideal observers. Because the noise spectral density is the same for both human and ideal observers, efficiency is simply the ratio of the threshold signal energies, that is, the ideal observer's signal energy (E_I) divided by the human's signal energy (E_H):

$$\text{efficiency} = (E_I)/(E_H).$$

A ratio of 1 (or 100%) represents ideal performance.

RESULTS AND DISCUSSION

Recognition of low-pass filtered objects

The purpose of the recognition experiment was to determine whether the low efficiency for object recognition was due to the human observers' failure to use high-frequency information used by the ideal observer. The results of the recognition experiment suggest that this does not account for the inefficiency.

Figure 3 shows human and ideal observer threshold signal-to-noise ratios (E/N) for recognition of low-pass filtered stimuli. E/N is plotted as a function of filter bandwidth for line drawings and silhouettes.

We first consider the performance of the ideal observer. In both rendering conditions, the ideal observer's thresholds increased as object bandwidth decreased from 6 to 0.5 cy/ob. This rise reflects a decline in the "quality" of information from 6 to 0.5 cy/ob. By "quality" of

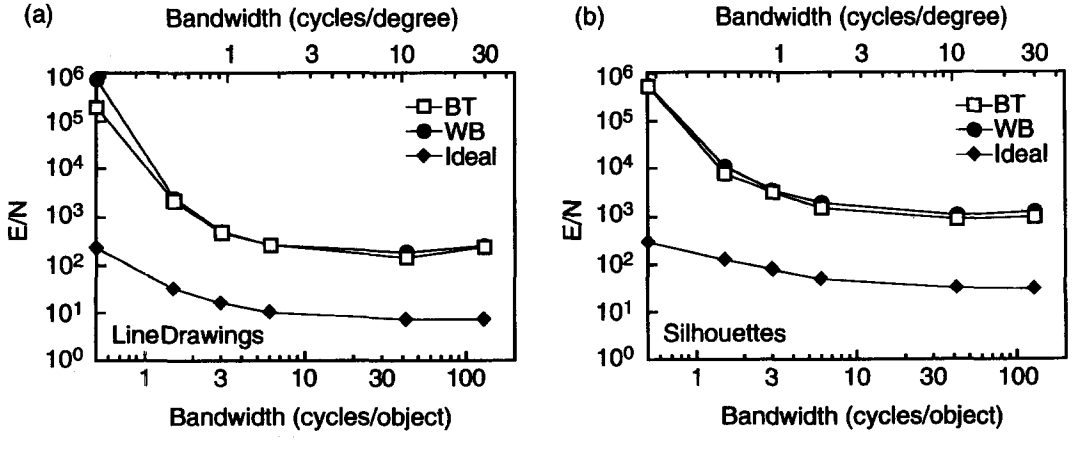

FIGURE 3. Threshold signal-to-noise ratio (E/N) for recognizing (a) line drawings and (b) silhouettes, plotted as a function of filter bandwidth for 2 human subjects and the ideal observer. Each data point for the human observers shows the mean of 3 thresholds. Standard errors are plotted but are smaller than the plot symbols.

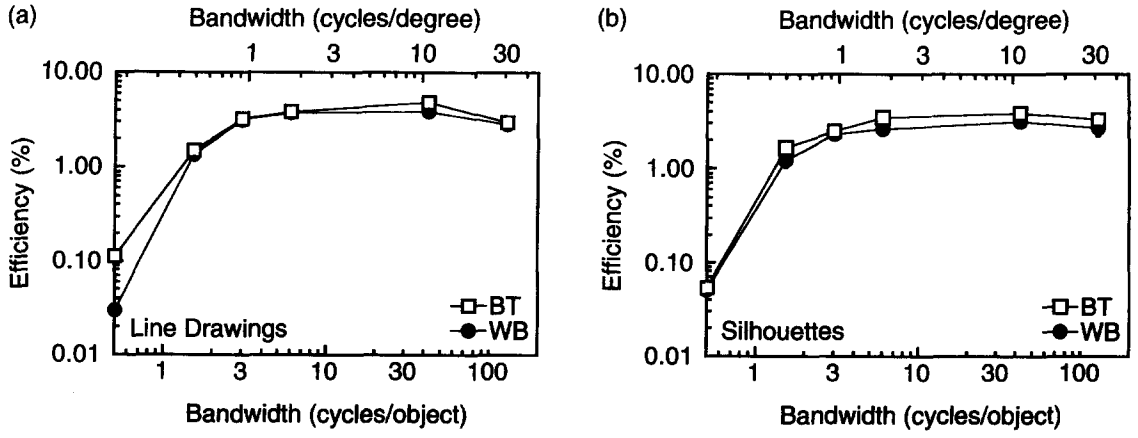

FIGURE 4. Human efficiencies for recognizing (a) line drawings and (b) silhouettes, plotted as a function of filter bandwidth. Each data point shows the mean of 3 observations. Standard errors are plotted but are often smaller than the plot symbols.

information, we mean how easily the objects can be distinguished with a given amount of signal energy. For example, our ideal observer's recognition thresholds show that a unit of signal energy in the band of frequencies between 3 and 6 cy/ob is more useful for distinguishing the objects than is a unit of signal energy at lower frequencies. At higher frequencies (6–128 cy/ob), the ideal observer's thresholds remained roughly constant across frequency. One possible interpretation of this result is that information quality is constant within this range of frequencies. However, the average amplitude spectrum of the set of objects falls off as $1/f^{1.46}$ for silhouettes and $1/f^{1.14}$ for line drawings. Because the signal energy is so low at high frequencies, changes in information quality may not be noticeable. The high frequencies therefore make a small contribution to the ideal observer's overall performance, regardless of how useful they are for distinguishing the objects.

Human thresholds follow trends similar to those of the ideal observer. Human thresholds were nearly constant for bandwidths above 6 cy/ob, with signal-to-noise ratios averaging about $10^{2.36}$ for line drawings and $10^{3.11}$ for silhouettes. Decreasing the bandwidth below 6 cy/ob increased human thresholds.

Efficiency measures can tell us whether the human performance in Fig. 3 is explained by informational constraints in the stimuli or processing limitations within the human. Recognition efficiencies (ideal thresholds divided by human thresholds from the curves in Fig. 3) are plotted in Fig. 4 as a function of filter bandwidth. For bandwidths of 6 cy/ob or greater, efficiency was constant for line drawings (3.75%) and silhouettes (3.23%). This indicates that, contrary to our prediction, efficiency did not increase as high frequencies were removed.

For line drawings, there was a very small but statistically significant increase in efficiency between unfiltered objects and objects filtered at 42 cy/ob (confirmed by a Tukey HSD test, $\alpha = 0.01$). This increase is in agreement with our prediction, but is not large enough to account for the generally low efficiencies.

A sharp decrease in efficiency occurred as filter bandwidth decreased from 6 to 1.5 cy/ob. This drop reflects the fact that human thresholds rose more rapidly than ideal observer thresholds as bandwidth decreased in this frequency range. The rapid change in efficiency suggests that humans make more effective use of information in the range from 1.5 to 6 cy/ob. At the very lowest bandwidth studied (0.5 cy/ob), efficiency was nearly 0%. This means that the ideal observer could make use of very coarse luminance cues that humans do not use in recognition. An analysis of variance performed on the efficiencies (excluding the 0.5 cy/ob filter condition) supports the finding that filter bandwidth affected efficiency [$F(4,20) = 65.06$, $P < 0.01$].

Efficiencies (averaged over all filter bandwidths) were higher for recognizing line drawings than silhouettes [$F(1,20) = 19.83$, $P < 0.01$]. A Tukey HSD test ($\alpha = 0.05$) revealed that this difference existed for filtered (3, 6 and 42 cy/ob) but not unfiltered (128 cy/ob) stimuli,* showing that human recognition of blurred stimuli is better for line drawings than silhouettes. One possible explanation for better performance with line drawings is that they contain fewer pixels than silhouettes. If humans can encode only a limited number of image samples, then reducing the number of samples in the stimulus should increase human efficiency. This is consistent with the present results for filtered stimuli, as well as with the findings of Tjan *et al.* (1995), who obtained higher efficiencies for recognizing small silhouettes (0.7 deg) than large silhouettes (2.8 deg). However, Tjan *et al.* (1995) found that recognition efficiency was substantially higher for small silhouettes than for line drawings containing approximately the same number of pixels. Furthermore, the present results showed no difference between line drawing and silhouette recognition with unfiltered stimuli. Thus, the higher efficiency for recognizing line drawings is not simply due to the reduced number of image samples.

*Tjan *et al.* (1995) found a small but statistically significant difference in efficiency for recognizing silhouettes vs line drawings. The reason for this discrepancy is unknown.

The data indicate that human recognition efficiency is quite low, even for low-pass filtered stimuli, when compared to efficiencies obtained in other complex perceptual tasks (see the Introduction). There is little evidence that the low efficiency is a consequence of a failure to use high-frequency information. Coupled with the findings on ideal observer performance at high frequencies, the human results show that the failure of humans to improve recognition performance at high frequencies is due to properties of the stimuli, not limitations of human processing. In other words, object recognition (at least our version of it) is a low-frequency task.

Recognition of band-pass filtered objects

The decline in efficiency at low bandwidths means that the ideal observer used some low-frequency information not used by humans. This raises the possibility of increasing human recognition efficiency by filtering out very low frequencies. To test this possibility, we measured human efficiency for recognizing band-pass filtered objects. The band-pass filtered images were constructed by (1) low-pass filtering our objects with the 6 cy/ob exponential filter, and then (2) high-pass filtering the resulting image with a first-order exponential filter. We used two high-pass filters, with cut-offs of 0.5 and 1.5 cy/ob. The resulting band-pass filters had center frequencies of 1.7 and 3.0 cy/ob, respectively. These are plotted in Fig. 5.

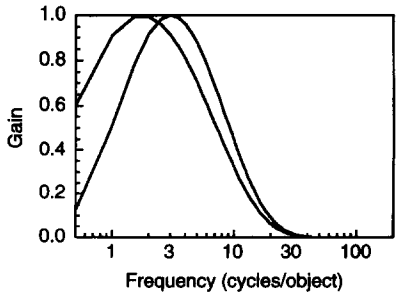

FIGURE 5. Two band-pass filters, centered at 1.7 and 3.0 cy/ob. The filters are of the form $Ae^{(-f/6)}e^{(-c/f)}$. A is a factor (different for each filter) that scales the luminances of the resulting images to fill the entire 0–255 range, and f is the input frequency. The first exponential term is the 6 cy/ob low-pass filter used in the low-pass recognition experiment. This is multiplied by a high-pass filter of bandwidth c (the frequency at which the gain is 1/e). c is 0.5 for the first filter and 1.5 for the second filter. A small amount of DC is added to each image such that no negative luminance is present.

Recognition efficiencies for the band-pass filtered stimuli are shown in Fig. 6, along with data for 6 cy/ob low-pass filtered objects. There was a statistically significant increase in efficiency for recognizing the band-pass filtered objects compared with the low-pass filtered objects, but the effect was very small [$F(2,12) = 11.39$, $P < 0.01$]. Clearly, people's failure to use information in very low frequencies does not account for the overall low recognition efficiency.

Detection of low-pass filtered objects

In the second experiment, we compared object recognition and detection efficiencies. Fig. 7 plots human and ideal observer threshold signal-to-noise ratios (E/N) for detection as a function of filter bandwidth. Human detection thresholds were nearly constant for bandwidths above 3 cy/ob. The ideal observer's thresholds were approximately constant across all frequencies. Its performance reflects the fact that all frequencies are equally informative for this detection task. Because the task is to determine whether or not a signal is present, the ideal observer's performance depends only on the total amount of signal energy, regardless of how this energy is distributed across frequency.

Recognition and detection efficiencies are plotted in Fig. 8. The results are consistent with the use of an edge-sampling strategy, as outlined in the Introduction. First, detection efficiency was higher for line drawings than for silhouettes (Fig. 8a vs b, and c vs d). Second, Figs 8a and c show that, for bandwidths above 1.5 cy/ob, line drawing recognition efficiencies are very similar to detection efficiencies. Although there was a statistically significant difference between these efficiencies (paired one-tailed t-test, $\alpha < 0.05$, 29 d.f.), the difference was quite small, with line drawing detection efficiencies about

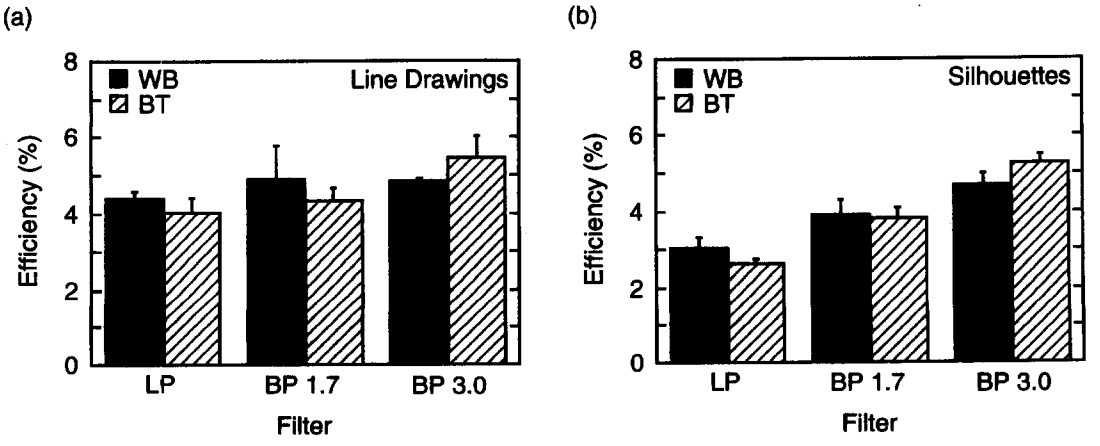

FIGURE 6. Efficiency for recognizing low-pass (LP) and band-pass (BP) filtered objects rendered as (a) line drawings and (b) silhouettes. The low-pass filter had a 1/e bandwidth of 6 cy/ob; the two band-pass filters were centered respectively at 1.7 and 3.0 cy/ob. Each data point shows the mean of 3 observations. Standard errors are plotted.

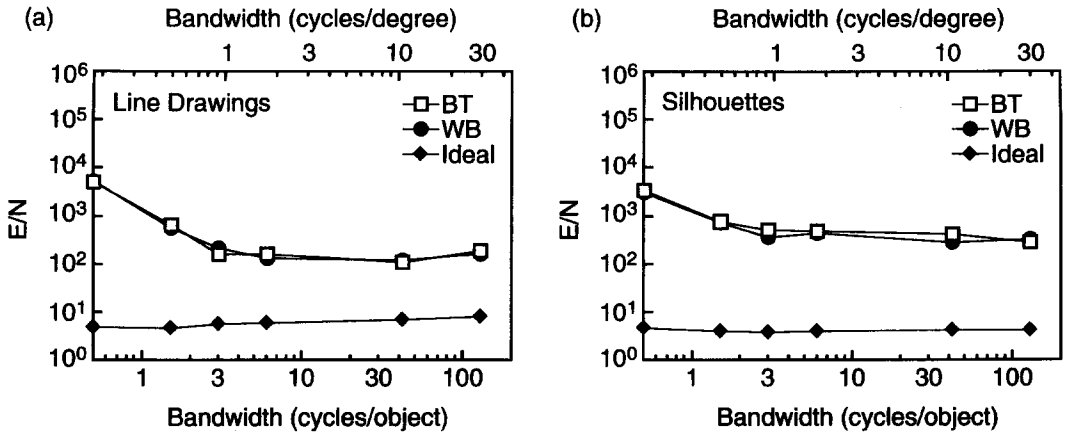

FIGURE 7. Threshold signal-to-noise ratio (E/N) for detecting (a) line drawings and (b) silhouettes, plotted as a function of filter bandwidth for 2 human subjects and the ideal observer. Each data point for the human observers shows the mean of 3 thresholds. Standard errors are plotted but are smaller than the plot symbols.

1.2 times higher than recognition efficiencies. Finally, for silhouettes (8b and d), recognition efficiencies were substantially higher (2.6 times on average) than detection efficiencies for bandwidths above 1.5 cy/ob (paired one-tailed t-test, $\alpha < 0.001$, 29 d.f.). These results demonstrate that detection efficiency does not place a ceiling on recognition efficiency. They are consistent with a visual-processing strategy that samples recognition-relevant features, such as bounding contours, rather than sampling image data across relatively uniform regions.

SIMULATION

The comparison of recognition and detection efficiencies suggests that the human visual system tends to sample image features that are relevant to recognition, such as edges. How can this selective sampling be

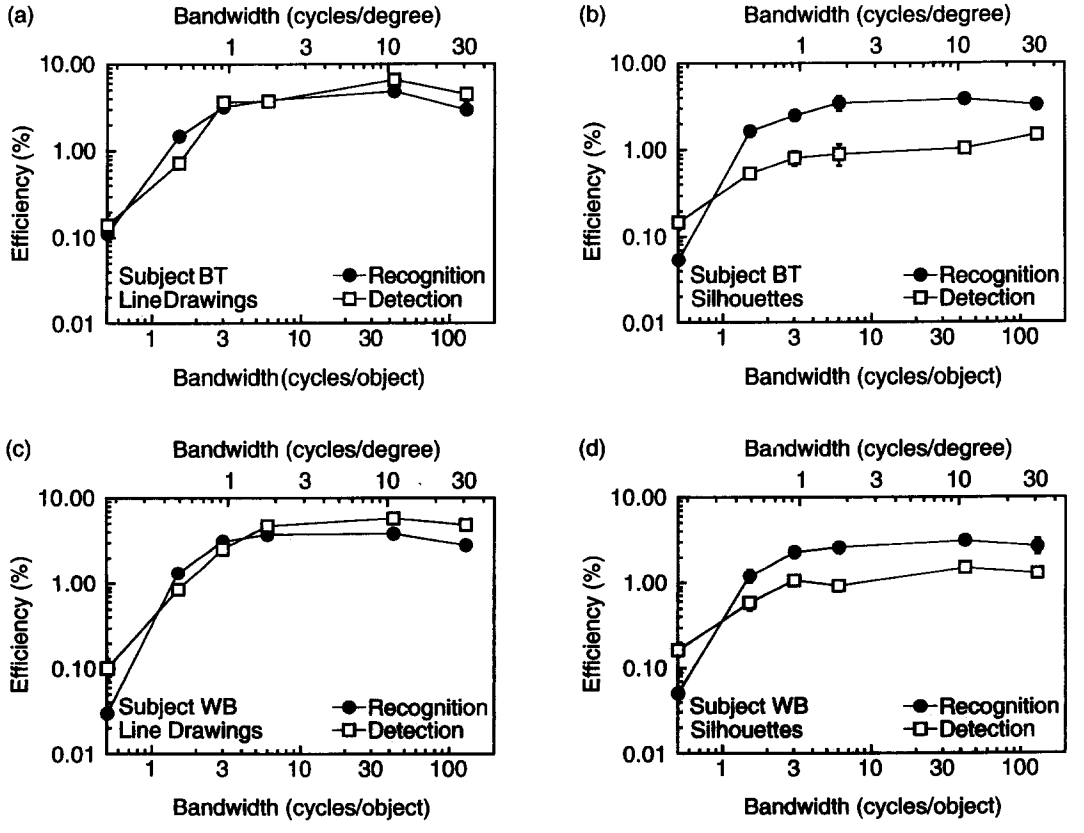

FIGURE 8. Human efficiencies for detecting (a, c) line drawings and (b, d) silhouettes, plotted as a function of filter bandwidth. Recognition efficiencies are replotted for comparison. Each data point shows the mean of 3 observations. Standard errors are plotted but are often smaller than the plot symbols.

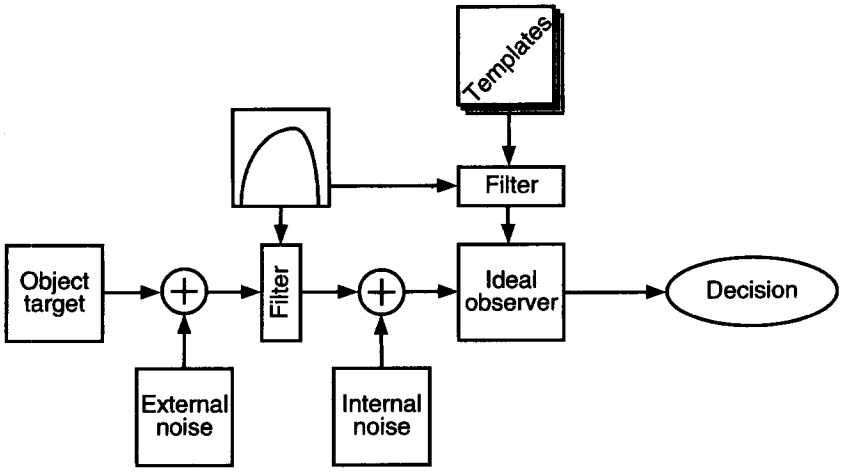

FIGURE 9. The narrow-band observer. The object target and luminance noise pass through a linear band-pass filter. Internal noise is then added. The observer compares the stimulus to its filtered templates and determines which object provides the best match.

implemented before the object's identity is known? One possibility is a bottom-up strategy, in which an appropriate spatial-frequency channel is used to pick out informative features. For example, recent results of Solomon and Pelli (1994) suggest that a band-pass channel mediates letter recognition.

To test the plausibility of this idea for object recognition, we programmed a simulated observer that views our low-pass filtered stimuli through a noisy band-pass filter, as diagrammed in Fig. 9. This observer is not intended as a quantitative model of human performance, but is instead used to demonstrate that the qualitative characteristics of our human data can be obtained using a band-pass channel. The input stimulus, comprised of external Gaussian white noise added to an object target, first passes through a linear band-pass filter. Internal Gaussian white noise is then added after filtering. This non-zero level of internal noise is necessary for the filter to have an effect. If there were no noise beyond the filter, the filter would not change the signal-to-noise ratio (both target and external noise would be attenuated by the same factor at every frequency), and performance would be unaffected.* The observer matches the filtered target plus internal noise to its stored templates, which are filtered with the same band-pass filter.

When white noise, consisting of uncorrelated image samples in the space domain, is passed through a band-pass filter, nearby pixel values in the filtered image have correlated values. In this case, the formulation of an ideal observer can not be done easily in the space domain. A nearly-ideal observer (see Appendix), however, can be formulated in the Fourier domain. In an unpublished manuscript, Chubb, Sperling and Parish (1988) derived a nearly-ideal observer for one-dimensional filtered signals in filtered noise. In the Appendix, we summarize this derivation, extend it to two-dimensional images, and formulate our narrow-band 2-D observer.

*Note that, if the filter is rectangular, there is no need for internal noise.

If the human visual system also uses this type of band-pass mechanism for object recognition and detection, then the narrow-band observer's pattern of efficiencies should be qualitatively similar to the humans': its recognition efficiency should exceed its detection efficiency in the case of silhouettes but not line drawings.

We programmed three different narrow-band observers, each of which used a different band-pass filter. The three filters were filters A, B, and C proposed by Wilson, McFarlane and Phillips (1983). These filters (plotted in Fig. 10) are centered at (A) 1.9 cy/ob (0.8 cy/deg), (B) 4.1 cy/ob (1.7 cy/deg), and (C) 6.7 cy/ob (2.8 cy/deg). The filters were chosen because they are based on human data, and because they cover the range of frequencies we found to be important for recognition. We set the internal noise spectral density to 0.2 μ (deg^2) (determined from human data in Tjan *et al.*, 1995). Each of the three narrow-band observers performed the tasks of recognizing and detecting our low-pass filtered objects (see the Methods for a description of the stimuli and procedures used).

The results are plotted in Fig. 11. The recognition efficiency of all three narrow-band observers was equal to or higher than the detection efficiency in both rendering conditions. The most notable similarity to our human efficiencies is that, for the filter-C observer, this

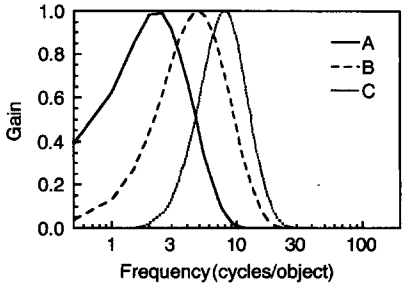

FIGURE 10. Band-pass filters used for the narrow-band observers. The filters are taken from Wilson *et al.* (1983). They are centered at (A) 1.9 cy/ob (0.8 cy/deg), (B) 4.1 cy/ob (1.7 cy/deg), and (C) 6.7 cy/ob (2.8 cy/deg).

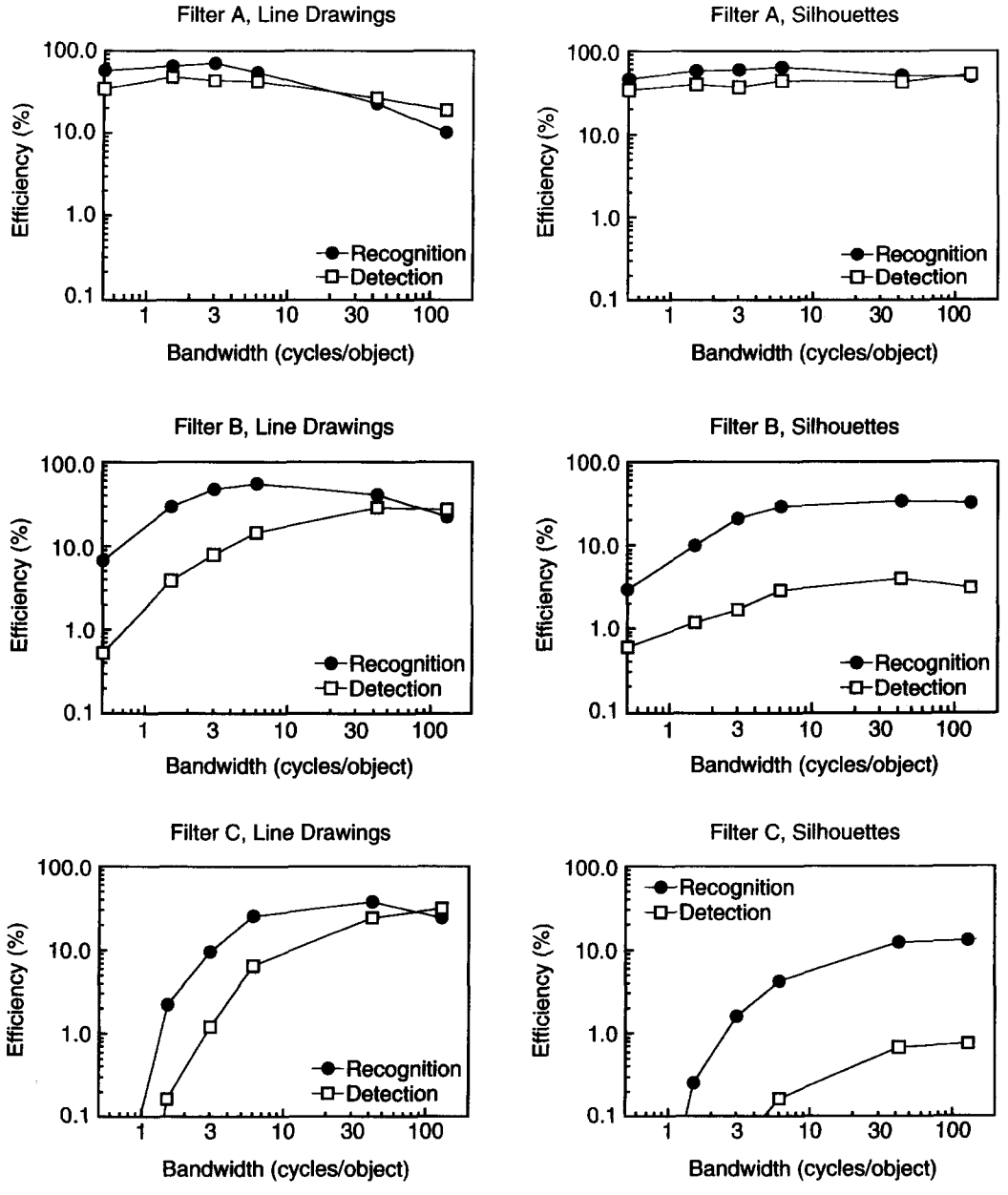

FIGURE 11. Efficiencies of 3 different narrow-band observers, relative to the broad-band ideal observer, for recognizing and detecting low-pass filtered silhouettes and line drawings.

difference was much larger for silhouettes than for line drawings. Thus, an observer viewing the world through a band-pass filter centered at 6.7 cy/ob does perform similarly to humans in one important way: recognition performance is enhanced at the expense of detection performance. Our simulation shows that this can be accomplished by the use of a simple linear band-pass mechanism.

The filter-B observer, which is centered within the range we found to be most important to human observers (1.5–6 cy/ob), also shows evidence of this difference between line drawings and silhouettes, but the difference is not as striking. The filter-B observer's graph of efficiency vs stimulus bandwidth is very similar in shape to human efficiencies (Fig. 8), rising at low frequencies and leveling off at high frequencies, but its efficiencies are generally higher than human efficiencies.

Our simulation of an ideal observer who views the world through a single band-pass spatial-frequency filter, though not capturing all of the quantitative details of our human data, does account for some of their qualitative features. In particular, it illustrates that band-pass mechanisms enhance recognition performance at the expense of detection performance.

CONCLUSIONS

We examined two possible explanations for the low object recognition efficiencies observed by Tjan *et al.* (1995) by low-pass filtering our stimuli and measuring recognition and detection efficiencies. Recognition efficiency failed to increase when high frequencies were removed from the stimuli, showing that the explanation does not lie in humans' failure to use these

frequencies. We also found that efficiency for object recognition can be higher than efficiency for object detection, implying that recognition was not limited by the detectability of the stimulus. Rather, it suggests a sampling scheme favorable for recognition, such as contour sampling. Finally, our simulation of an ideal observer with a narrow-band front end argues that spatial-frequency channels function to extract features relevant to recognition, not detection.

REFERENCES

Barlow, H. B. & Reeves, B. C. (1979). The versatility and absolute efficiency of detecting mirror symmetry in random-dot display. *Vision Research, 19*, 783–793.

Burgess, A. E. & Colborne, B. (1988). Visual signal detection. IV. Observer inconsistency. *Journal of the Optical Society of America A, 5(4)*, 617–627.

Chubb, C., Sperling, G. & Parish, D. H. (1988). Designing psychophysical discrimination tasks for which ideal performance is computationally tractable. Unpublished manuscript.

Fiorentini, A., Maffei, L. & Sandini, G. (1983). The role of high spatial frequencies in face perception. *Perception, 12*, 195–201.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Ginsburg, A. P. (1980). Specifying relevant spatial information for image evaluation and display design: an explanation of how we see certain objects. In *Proceedings of the Society for Information Display, 21*, 219–227.

Landy, M. S., Cohen, Y. & Sperling, G. (1984). HIPS: image processing under UNIX. Software and applications. *Behavior, Research Methods, Instruments and Computers, 16*, 199–216.

Legge, G. E., Gu, Y. & Luebker, A. (1989). Efficiency of graphical perception. *Perception & Psychophysics, 46(4)*, 365–374.

Legge, G. E., Pelli, D. G., Rubin, G. S. & Schleske, M. M. (1985). Psychophysics of reading I. Normal vision. *Vision Research, 25(2)*, 239–252.

Norman, J. & Ehrlich, S. (1987). Spatial frequency filtering and target identification. *Vision Research, 27(1)*, 87–96.

Parish, D. H. & Sperling, G. (1991). Object spatial frequencies, retinal spatial frequencies, noise, and the efficiency of letter discrimination. *Vision Research, 31(7–8)*, 1399–1415.

Pelli, D. G. & Zhang, L. (1991). Accurate control of contrast on microcomputer displays. *Vision Research, 31(7–8)*, 1337–1350.

Solomon, J. A. & Pelli, D. G. (1994). The visual channel that mediates letter identification. *Nature, 369*, 395–397.

Tjan, B. S., Braje, W. L. & Legge, G. E. (1994). Spatial uncertainty in human object recognition. *Investigative Ophthalmology and Visual Science, 35(4)*, 1626.

Tjan, B. S., Braje, W. L., Legge, G. E. & Kersten, D. J. (1995). Human efficiency for recognizing 3-D objects in luminance noise. *Vision Research, 35*, 3053–3069.

Watt, R. J. & Morgan, M. J. (1985). A theory of the primitive spatial code in human vision. *Vision Research, 25(11)*, 1661–1674.

Wetherill, G. B. & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *The British Journal of Mathematical and Statistical Psychology, 18(1)*, 1–10.

Wilson, H. R., McFarlane, D. K. & Phillips, G. C. (1983). Spatial tuning of orientation selective units estimated by oblique masking. *Vision Research, 23*, 873–882.

Acknowledgements—This work was supported by NIH Grant EY02857 to Gordon E. Legge. Wendy Braje was supported by NIH training grant EY07133. Preliminary versions of this paper were presented at the Association for Research in Vision and Ophthalmology meeting (1992 and 1993). We would also like to thank Charlie Chubb for helpful discussions of the Fourier-based narrow-band observer.

APPENDIX

Our simulated narrow-band observer is modeled as an ideal observer viewing the world through a linear band-pass filter. This Appendix gives the formulation of such a model. We restate Chubb *et al.*'s (1988) key result and extend it to our two-dimensional application. We then formulate our narrow-band model. We assume readers are familiar with the space-domain formulation of the ideal observer (see Tjan *et al.*, 1995).

Fourier transform of Gaussian noise

Chubb *et al.* (1988) proved that in the limit, Gaussian noise in the space domain has Gaussian coefficients in the frequency domain. Consider discrete one-dimensional noise represented as an array of real numbers $\langle X_k \rangle$, $k = 1,2,3, \ldots, D$. The discrete Fourier transform $\Phi[\omega]$ of this noise is

$$\Phi[\omega] = \frac{1}{D}\sum_{k=1}^{D} X_k \exp(-i2\pi\omega k/D),\ \omega\epsilon\{0,1,2,\ldots,D/2\} \tag{1}$$

$$= \underbrace{\frac{1}{D}\sum_{k=1}^{D} X_k \cos(2\pi\omega k/D)}_{R[\omega]} + i\underbrace{\frac{-1}{D}\sum_{k=1}^{D} X_k \sin(2\pi\omega k/D)}_{I[\omega]}.$$

We deliberately omit the negative frequencies from $-D/2$ to -1 in our formulation. Since our signals are real values, the Fourier coefficient at a negative frequency is just the complex conjugate of its positive counterpart, and thus carries no extra information.

Chubb *et al.* showed that, if $\langle X_k \rangle$ are jointly independent, identically distributed normal random variables,* each with mean 0 and variance σ^2, then all the real parts ($R[\omega]$) and imaginary parts ($I[\omega]$) of $\Phi[\omega]$, for $\omega = 0,1, \ldots, D/2$, will also be jointly independent. When D is sufficiently large, $\Phi[\omega]$ will tend towards the following distribution:

$$(R\omega],\ I[\omega]) \xrightarrow[D\to\infty]{} \begin{cases} \frac{\sigma}{\sqrt{2D}}(\varepsilon,\varepsilon') \text{ for positive integers } \omega < D/2 \\ \frac{\sigma}{\sqrt{D}}(\varepsilon,0) \text{ for } \omega = 0 \text{ or (with } D \text{ even) for } D/2 \end{cases} \tag{2}$$

where ε and ε' are jointly independent standard normal random variables.

We can extend this result to noise in two dimensions. Let $\langle X_{mn} \rangle$, $m = 1, \ldots, D_x$, and $n = 1, \ldots, D_y$ be a two-dimensional array of jointly independent and identically distributed normal random variables with mean 0 and variance σ^2. The discrete Fourier transform of $\langle X_{mn} \rangle$ is given by:

$$\Phi[\omega_x,\omega_y] = \frac{1}{D_y}\sum_{n=1}^{D_y}\left[\exp(-i2\pi\omega_y n/D_y)\underbrace{\frac{1}{D_x}\sum_{m=1}^{D_x} X_{mn}\exp(-i2\pi\omega_x m/D_x)}_{A_n}\right]$$

$$\text{for integers } \omega_x\epsilon\{0,1, \ldots, D_x/2\} \tag{3}$$

$$\text{and } \omega_y\epsilon\{-(D_y/2)+1, \ldots, -1,0,1, \ldots, D_y/2\}.$$

As in the one-dimensional case, only one half of the frequency domain is considered; the other half is simply its complex conjugate. The inner summation term $A_n[\omega_x]$ is the one-dimensional discrete Fourier transform of $\langle X_{mn} \rangle$, identical to the expression in equation (1). Therefore, if we let $RA_n[\omega_x]$ and $IA_n[\omega_x]$ be the real and imaginary parts of $A_n[\omega_x]$

*Chubb *et al.*'s result is actually more general and allows any distribution whose third moment is zero.

respectively, from equation (2) we know that they are jointly independent with the following normal distributions:

$$(RA_n[\omega_x],\ IA_n[\omega_x]) \xrightarrow[D_x\to\infty]{} \begin{cases} \dfrac{\sigma}{\sqrt{2D_x}}(\varepsilon,\varepsilon') \text{ for positive integers } \omega_x < D_x/2 \\ \dfrac{\sigma}{\sqrt{D_x}}(\varepsilon,0) \text{ for } \omega_x = 0 \text{ or (with } D_x \text{ even) for } D_x/2. \end{cases} \quad (4)$$

Expanding equation (3) in terms of $RA_n[\omega_x]$ and $IA_n[\omega_x]$, and regrouping its real and imaginary terms, we have:

$$\Phi[\omega_x,\ \omega_y] = \underbrace{\frac{1}{D_y}\sum_{n=1}^{D_y} RA_n[\omega_x]\cos(2\pi\omega_y n/D_y)}_{S[\omega_x,\omega_y]} + i\underbrace{\frac{1}{D_y}\sum_{n=1}^{D_y} IA_n[\omega_x]\cos(2\pi\omega_y n/D_y)}_{T[\omega_x,\omega_y]} - i\underbrace{\frac{1}{D_y}\sum_{n=1}^{D_y} RA_n[\omega_x]\sin(2\pi\omega_y n/D_y)}_{U[\omega_x,\omega_y]} + \underbrace{\frac{1}{D_y}\sum_{n=1}^{D_y} IA_n[\omega_x]\sin(2\pi\omega_y n/D_y)}_{V[\omega_x,\omega_y]}$$

$$= \underbrace{S[\omega_x,\omega_y] + V[\omega_x,\omega_y]}_{R[\omega_x,\omega_y]} + i\underbrace{(T[\omega_x,\omega_y] - U[\omega_x,\omega_y])}_{I[\omega_x,\omega_y]}. \quad (5)$$

Note that the summation terms S, T, U, and V share the same form as R and I in equation (1). By applying equation (2) to these terms and noting that RA_n and IA_n are jointly independent (equation 4), we can conclude that these summation terms are themselves jointly independent random normal variables. Their distributions for non-negative ω_x and ω_y can be determined by a straightforward application of equation (2), with σ replaced by the standard deviations given in equation (4) and considering all distinct cases of ω_x and ω_y.

Also note that the real ($R[\omega_x,\omega_y]$) and imaginary ($I[\omega_x,\omega_y]$) parts of $\Phi[\omega_x,\omega_y]$ are respectively the sum and difference of jointly independent normal variables: $R = S + V$ and $I = T - U$. Therefore, they are also jointly independent and normal, with a variance equal to the sum of the variances of their constituency.

Lastly, we want to show that $R[\omega_x,\omega_y]$ and $I[\omega_x,\omega_y]$ for negative ω_y are also jointly independent with the rest. This can be done by observing for $1 \leqslant \omega_y \leqslant D_y/2$,

$$R[\omega_x,\omega_y] = S[\omega_x,\omega_y] + V[\omega_x,\omega_y] \text{ and}$$

$$R[\omega_x,-\omega_y] = S[\omega_x,\omega_y] - V[\omega_x,\omega_y].$$

Since S and V are jointly independent and identically distributed in the range, so are $S + V$ and $S - V$. This is because $E[(S + V)(S - V)] = E[S^2] - E[V^2] = 0$.

Finally, we can summarize the distribution of the real and imaginary parts of the discrete Fourier transform of two-dimensional Gaussian noise with standard derivation σ as:

For integers ω_x and ω_y,

$$\Phi[\omega_x,\omega_y] \xrightarrow[D_y\to\infty]{D_x\to\infty} \begin{cases} \dfrac{\sigma}{\sqrt{2D_xD_y}}(\varepsilon,\varepsilon') & \text{if } 0 < \omega_x \lneq D_x/2 \text{ or } 0 < \omega_y < D_y/2 \\ \dfrac{\sigma}{\sqrt{D_xD_y}}(\varepsilon,0) & \text{if } (\omega_x,\omega_y) = (0,0),\ (0,D_y/2),\ (D_x/2,0) \text{ or } (D_x/2,D_y/2). \end{cases} \quad (6)$$

Narrow-band observer

The result of equation (6) allows us to formulate an ideal observer in the Fourier domain, which is needed for our narrow-band observer. In the space domain, the signal-plus-noise seen by the ideal observer can be expressed as

$$(s + n_e)*b + n_i = \underbrace{s*b}_{\text{filtered signal}} + \underbrace{n_e*b + n_i}_{\text{noise}} \quad (7)$$

where

s = signal (a view of an object),
n_e = external Gaussian noise added to the display,
b = band-pass filter kernel used by the observer,
n_i = Gaussian noise internal to the observer.

Using upper-case letters to represent the discrete Fourier transform of their lower-case counterparts, we can express the same signal-plus-noise in the Fourier domain as

$$(S + N_e)B + N_i = \underbrace{SB}_{\text{filtered signal}} + \underbrace{N_eB + N_i}_{\text{noise}}. \quad (8)$$

Convolution in the space domain is a simple pointwise multiplication in the Fourier domain. Therefore, unlike in the space domain, the neighboring pixels of the external noise expressed in the Fourier domain are not correlated after filtering.

From equation (6), we know that both N_e and N_i are independent Gaussian random variables, so the combined noise term is also Gaussian and independent for all frequency points. If we take the filtered signal (SB) as the "signal," we are back to the familiar case of signal discrimination in Gaussian noise, and the formulation by Tjan *et al.* (1995) applies. The ideal decision rule is similar to that stated in the Methods section with two exceptions: (1) the variance of the noise changes from frequency point to frequency point according to the MTF of the band-pass filter, and (2) half of the frequency "image" consists of the real parts of the Fourier coefficients, and the other half consists of the imaginary parts. This is because the real and imaginary parts are jointly independent (equation 6). To be precise, the ideal decision rule for the narrow-band observer is to choose the object i that maximizes the expression

$$L'(i) = \sum_{j=1}^{\#\text{views}} \exp\left[-\frac{1}{2} \sum_{(\omega_x,\omega_y)=\left(0,\frac{-D_y}{2}+1\right)}^{\left(\frac{D_x}{2},\frac{D_y}{2}\right)} \left[\frac{\left\| X[\omega_x,\omega_y] - T_{ij}B[\omega_x,\omega_y] \right\|^2}{\left(R\sigma_{N_e}B[\omega_x,\omega_y]\right)^2 + \left(R\sigma_{N_i}[\omega_x,\omega_y]\right)^2} \right] \right] \quad (9)$$

where

views = number of views per object (8 for recognition, 32 for detection),
X = input image ($X[\omega_x,\omega_y] = (S + N_e)B + N_i$),
$T_{ij}B$ = filtered templates of object i view j,
$\|....\|$ = modulus of a complex number,
R = real part of a term,
$\sigma_{N_e}B$ = SD of the filtered external noise in the Fourier domain ($\sigma_{N_e}B[\omega_x,\omega_y] = \sigma_{N_e}[\omega_x,\omega_y]B[\omega_x,\omega_y]$),
σ_{N_i} = SD of the internal noise in the Fourier domain.

Only the real parts of the noise are used in the expression. This is because (1) the imaginary parts have the same distributions as do the real parts, except at the corners of the DC and Nyquist frequencies (see equation 6) where they diminish to zero; and (2) at those corners, the imaginary parts of both the signal and template are zero, thus removing the imaginary parts from the summation.