Pergamon

0042-6989(95)00070-4

Vision Res. Vol. 35, No. 21, pp. 3053–3069, 1995
Copyright © 1995 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0042-6989/95 $9.50 + 0.00

Human Efficiency for Recognizing 3-D Objects in Luminance Noise

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

Received 9 May 1994; in revised form 12 January 1995

The purpose of this study was to establish how efficiently humans use visual information to recognize simple 3-D objects. The stimuli were computer-rendered images of four simple 3-D objects—wedge, cone, cylinder, and pyramid—each rendered from 8 randomly chosen viewing positions as shaded objects, line drawings, or silhouettes. The objects were presented in static, 2-D Gaussian luminance noise. The observer's task was to indicate which of the four objects had been presented. We obtained human contrast thresholds for recognition, and compared these to an ideal observer's thresholds to obtain efficiencies. In two auxiliary experiments, we measured efficiencies for object detection and letter recognition. Our results showed that human object-recognition efficiency is low (3–8%) when compared to efficiencies reported for some other visual-information processing tasks. The low efficiency means that human recognition performance is limited primarily by factors intrinsic to the observer rather than the information content of the stimuli. We found three factors that play a large role in accounting for low object-recognition efficiency: stimulus size, spatial uncertainty, and detection efficiency. Four other factors play a smaller role in limiting object-recognition efficiency: observers' internal noise, stimulus rendering condition, stimulus familiarity, and categorization across views.

Object recognition Object detection Letter recognition Efficiency Ideal observer

INTRODUCTION

Object recognition is one of the most important functions of human vision. There has been relatively little research on the role of sensory processes in object recognition. Instead, recent research in this area has focused on higher-level issues, including the computational theory associated with recognizing three-dimensional objects from two-dimensional retinal images (Marr, 1982), the nature of the perceptual and memory representations of objects (Marr, 1982; Biederman, 1987; Pentland, 1986; Brooks, 1983; Cooper & Schacter, 1992; Liu, Kersten & Knill 1995), and operations for matching memory representations to early perceptual representations (Ullman, 1989; Lowe, 1987). Others have focused on visual cues that may be used in object recognition, such as stereo, shading, texture, and motion (Bülthoff & Mallot, 1988; Todd & Bressan, 1990; Todd & Akerstrom, 1987; Pentland, 1989; Poggio, Gamble & Little, 1988; Voorhees & Poggio, 1988; Sperling & Landy, 1989). This paper is concerned primarily with characterizing limitations on object recognition imposed by the information content of stimuli and low-level sensory constraints. We used methods from signal-detection theory to ask how efficiently people use sensory signals to perform 3-D object recognition.

The term *object recognition* refers to two processes. In one, we recognize a familiar object (e.g. our favorite easy chair) by associating its visual image with a specific object we remember. In the second, we recognize a novel exemplar of a familiar category (e.g. someone else's chair) using some functional or structural criteria. In this study, we consider *recognition* only in the first sense; all of the targets are known to the observer *a priori.*

In our main experiment, we used four simple objects—wedge, cone, cylinder, and pyramid—each portrayed from eight viewpoints (Fig. 1). In each trial, one of these 32 object stimuli was selected at random and presented in noise. The subject's task was to identify the object. We chose a small number of simple and familiar objects to minimize memory demands on the subjects. We chose 3-D objects, seen from multiple viewpoints, to represent real-world object recognition. Unlike the studies by Bülthoff and Edelman (1992), Tarr and Pinker (1989) and Liu *et al.* (1995), in which subjects were trained on some views and tested on either the training or novel views, we trained our subjects on all test views in order to minimize their reliance on mid-level or high-level memory representations of objects. We wanted to focus on the low-level sensory contributions to recognition. Readers may refer to Liu *et al.* (1995) for an ideal observer approach to studying memory representations of objects.

*Department of Psychology, University of Minnesota, 75 East River Road N218, Minneapolis, MN 55455-0344, U.S.A.

†To whom all correspondence should be addressed [*Email* tjan@cs.umn.edu].

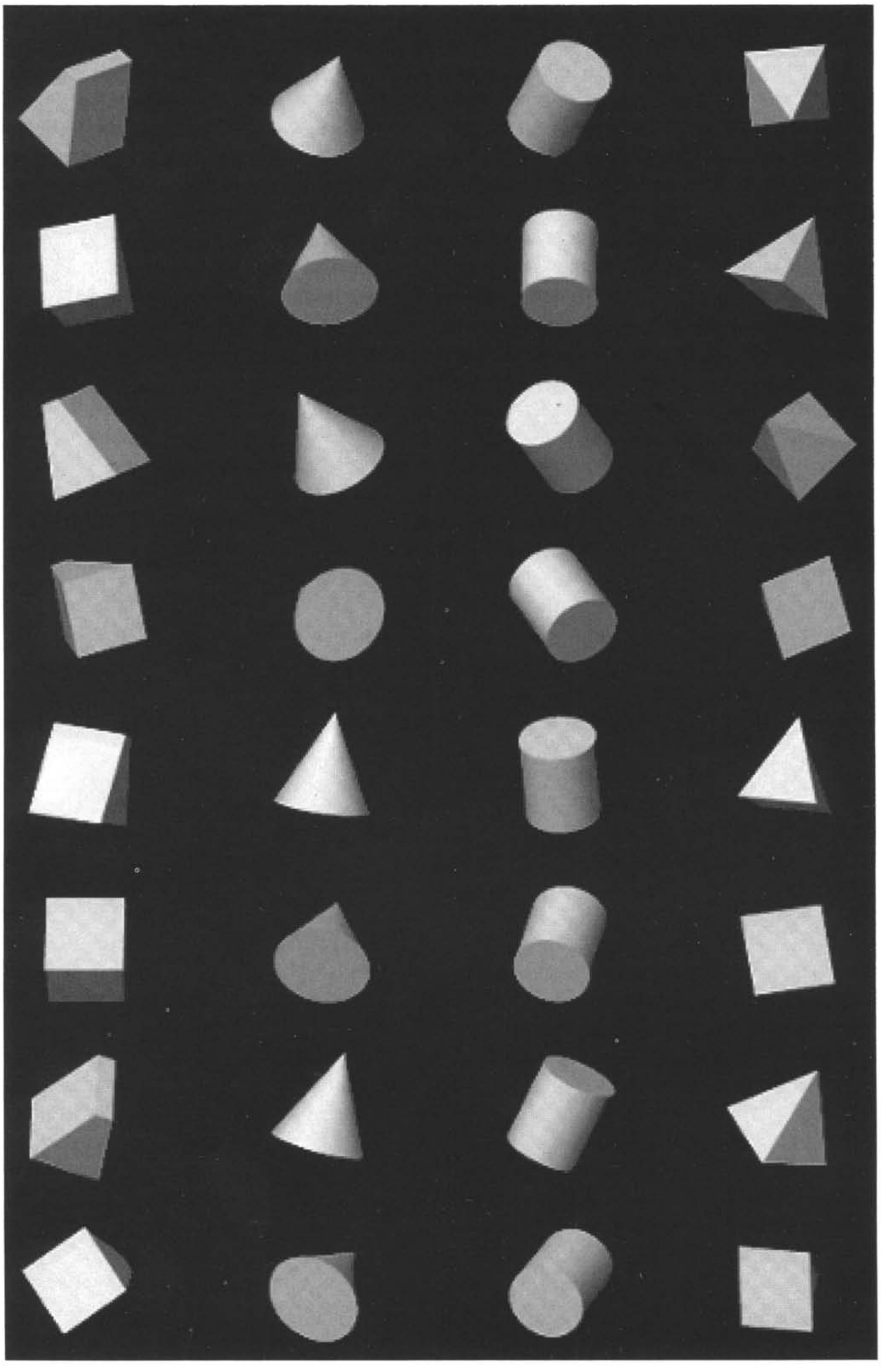

FIGURE 1. Eight views of the four objects used in the object recognition experiment. From left to right in columns: wedge, cone, cylinder, and pyramid.

Many pictorial cues (or features) are available for recognizing objects. Those that have received considerable study include texture, color, specularity, transparency, shading, shadows, spatial frequency distribution, edges, edge junctions, and occluding contours. Studying human performance for recognizing objects rendered by a single cue (e.g. shading), allows us to determine if human perception can use that cue. By itself however, an absolute performance level, such as a contrast threshold, does not indicate how well the human uses the cue. Poor performance might mean ineffective use of the cue by the observer, or it might mean that the cue contained relatively little information to begin with. To decide between these alternatives, we need a means for distinguishing the intrinsic information content of the cue from the amount of information used by the observer.

The same point also applies to performance comparisons across cues.

The notion of an *ideal observer* provides a solution to the problem of quantifying the information content of cues for a visual task. The ideal observer is a Bayesian observer that makes the best inference from the image data (Kersten, 1990). For the task of object recognition described below, the ideal observer determines the most likely object out of a known finite collection, given a noisy view of that object and knowledge of the possible views. It is ideal in the sense that it makes the least number of errors on average. Given a model for ideal performance, we can then compare it to human performance. In this paper, we use *efficiency*, as defined in signal-detection theory, to provide a measure of how effectively image information is used to infer the identity of an object.

It is only possible to measure efficiency if there is some limit on the performance of the ideal observer. In practice, performance can be limited by introducing some form of statistical uncertainty into the task. The kind of uncertainty depends on the scientific question being asked. For example, there can be inherent ambiguity in the 2-D projection of 3-D feature points. Liu *et al.* (1995) calculated the ideal observer for object recognition with viewpoint uncertainty, in addition to added structural noise (random positional placement of the 3-D feature points of the object). In our task, uncertainty was introduced in the form of superimposed luminance noise. One advantage of using luminance noise is that we are able to make direct comparisons of our results in object recognition with other studies that have used luminance noise to limit performance.

The concept of efficiency was introduced in statistics by Fisher (1925) and defined in the context of signal-detection theory by Tanner and Birdsall (1958). Hecht, Shlaer and Pirenne (1942) brought the notion of an ideal observer to vision with their work on the effects of quantum fluctuations on light sensitivity in the dark-adapted eye. Subsequently, ideal observers have been invoked in the theoretical analysis of many simple detection and discrimination tasks, culminating in the elegant sequential ideal-observer analysis of Geisler (1989). Ideal-observer analysis has also been used in studying more complex visual-information processing tasks, including detection of mirror symmetry (Barlow & Reeves, 1979), discrimination of dot density (Barlow, 1978), detection of modulation of dot density (van Meeteren & Barlow, 1981), discrimination of the number of dots in displays (Burgess & Barlow, 1983), estimation of means and variances of scatter plots and other graphical displays (Legge, Gu & Luebker, 1989), location of the centroid of dot clusters (Morgan & Glennerster, 1991), letter recognition (Parish & Sperling, 1991; Solomon & Pelli, 1994), and object recognition with structural noise (Liu *et al.*, 1995). We continue this trend in complexity by using ideal-observer analysis as a theoretical framework for studying object recognition.

Our initial goal was to find out if efficiency for object recognition is very high, 30–60%, as in some of the information processing tasks listed above. If so, we can conclude that recognition performance is limited almost entirely by the information content of the stimulus (at least of the simple stimuli we studied), and not by human information processing.

We also wanted to find out if human vision is particularly efficient in processing specific types of luminance or contour information in objects. We measured efficiency for recognizing objects rendered in four cue conditions: Lambertian shading, line drawings, large silhouettes, and small silhouettes. Lambertian shading provided the greatest visual detail, including bounding contours, internal contours, and luminance gradients. Line drawings retained bounding and internal contours but no luminance gradient information. The silhouettes were uniform in luminance and contained only bounding contours. Comparison of performance with large and small silhouettes permitted evaluation of the effects of spatial summation on efficiency. Figure 2 shows examples of the four rendering conditions.

When it transpired that efficiency was quite low for all rendering conditions, we conducted two auxiliary experiments. In one, we measured efficiency for detecting (rather than recognizing) objects to see if recognition efficiency is limited by the ability to detect the presence of an object. In a second auxiliary experiment, we measured efficiency for recognizing letters. Our purpose was to confirm previous reports of higher efficiencies [as high as 42% in Parish and Sperling (1991)] and to try to account for the gap in efficiency between object recognition and letter recognition.

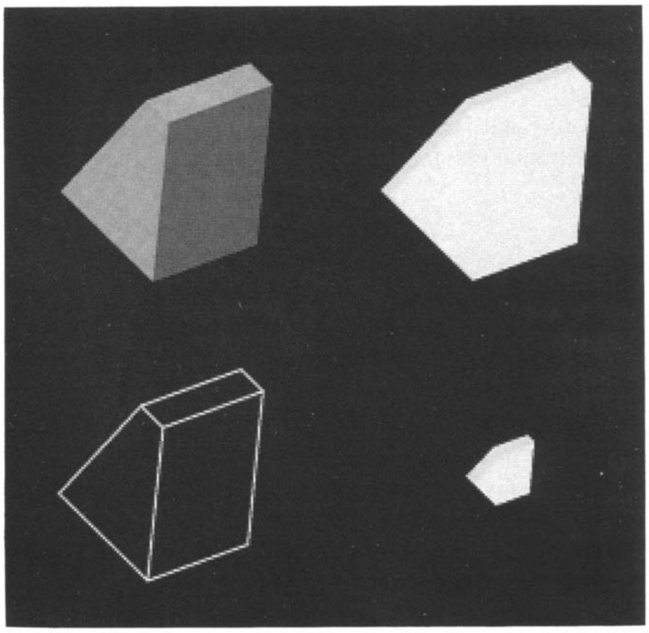

FIGURE 2. Four rendering conditions used in the object recognition experiment. Clockwise, beginning at the upper left corner: shaded object, large silhouette, small silhouette, and line drawing. All except the shaded object condition were also used in the object detection experiment.

THEORY

Ideal observer analysis

We assume that all of the targets are known to the observer *a priori*. With this requirement, we can formulate an ideal observer for the task of 3-D object recognition. A 3-D object, illuminated by a light source, can be rendered as an image on a raster display. The image is composed of a finite number of pixels, each with a finite number of luminance levels (and possible colors). If the intensity of the light source is known, and its position is fixed relative to the observer,* the image of a 3-D object will depend only on its pose (i.e. its 3-D position and orientation with respect to the observer). Although there is no limit on the number of possible poses, the discrete nature of the imaging system (and the observer's finite field size and spatial resolution) means that there is a finite number of points in the pose space that can be rendered by the imaging system. As a result, a 3-D object can be completely represented, at least in theory, by a large but finite set of all of its 2-D projections onto the image plane. Each of these projections can be thought of as a template, corresponding to the noise-free image of one view of one object.

Suppose that O_i is the ith object and T_{ij} is the jth view (projection or template) of the ith object, where $i = 1 \ldots n$ objects, and $j = 1 \ldots v$ discrete views. Given an image R, the probability of object O_i being present is the sum of the probabilities of each of the individual views being presented, and these probabilities can be expressed using Bayes rule (cf. Duda & Hart, 1973):

$$P(O_i|R) = \sum_j P(T_{ij}|R) = \sum_j \frac{P(R\,|\,T_{ij})P(T_{ij})}{P(R)}. \qquad (1)$$

The ideal observer's task is to select the object i that maximizes the quantity in equation (1), thus giving it the best chance of being correct. This is an *a posteriori* maximization rule. Because the probability $P(R)$ of seeing image R is independent of the observer's choice of i, the optimal strategy can be restated as choosing the object i that maximizes the following function:

$$L(i) = \sum_j P(RT_{ij})P(T_{ij}). \qquad (2)$$

$L(i)$ is a sum of the product of two probabilities: the probability (or the likelihood) of producing the observed image R from a given template T_{ij}, and the prior probability of the template. Calculating the likelihood term involves comparing the image with a stored template. The prior probability term takes into account the possibility that some views are more likely than others.

The details of the conditional probabilities in equation (2) depend on the complexity of the world and the imaging process. If we assume that an image R contains the projection of a *single* object, perturbed by the addition of static Gaussian luminance noise with zero mean contrast and standard deviation σ, the conditional probability can be expressed as

$$P(R\,|\,T_{ij}) = \frac{1}{(\sigma\sqrt{2\pi})^M}\exp\left(-\frac{1}{2\sigma^2}\left\|R - T_{ij}\right\|^2\right), \qquad (3)$$

where M is the number of pixels in the image, and $\| R - T_{ij} \|^2$ is the Euclidean distance between the image R and the template T_{ij}. Since the term before the exponential function is independent of i, maximizing equation (2) is equivalent to maximizing the following:

$$L'(i)\sum_j \exp\left(-\frac{1}{2\sigma^2}\left\|R - T_{ij}\right\|^2\right)P(T_{ij}). \qquad (4)$$

Equation (4) shows that the ideal strategy for recognizing 3-D objects in Gaussian noise is to compute a weighted sum of a similarity measure between the 2-D noisy image (i.e. stimulus) and each possible 2-D projection of an object. With Gaussian noise, the similarity measure is monotonic to the negative of the Euclidean distance between the image and a template. Calculation of this Euclidean distance is usually known as doing *template matching*.

Notice that if there is only one view for each object, the summation sign of equation (4) used for grouping views into objects can be dropped, and since the exponential function is monotonic, maximizing equation (4) is the same as minimizing the Euclidean distance $\| R - T_{ij} \|^2$ between the image and the template. Furthermore, $\| R - T_{ij} \|^2$ equals $\| R \|^2 - 2RT_{ij} + \| T_{ij} \|^2$. Notice that $\| R \|^2$ is a property of the image and independent of i or j. If the stimuli are constructed from templates of equal energy, i.e. $\| T_{ij} \|^2$ is constant for all i and j, then minimizing the Euclidean distance is the same as maximizing the cross correlation RT_{ij} between the image and a template. The signal-known-exactly ideal observer (cf. Green & Swets, 1974), which uses the strategy of maximizing the cross correlation, is therefore a special case of the ideal observer formulated here.

Definition and interpretation of efficiency

Burgess and Barlow (1983) described two generic ways in which human performance can be suboptimal. Observers might simply fail to use some of the information available to them, but optimally process the remaining information. On the other hand, observers might use all the information, but contribute imprecision (intrinsic noise) due to errors of internal representation. Burgess and Barlow (1983) showed how these two factors—incomplete sampling and internal (equivalent) noise—can be teased apart by measuring thresholds as a function of the level of externally added visual noise [see also Barlow (1977), Pelli (1981, 1990), Burgess, Wagner, Jennings and Barlow (1981) and Legge, Kersten and Burgess (1987)]. Adopting this approach, we can treat the human observer as equivalent to an ideal observer "sitting behind" a noisy and subsampled information channel.

We can apply this approach by measuring observers' contrast thresholds for recognizing objects in noise. The

*The observer does not need to explicitly know the position of the light source, only that it is unchanged with respect to the observer for all images. This assumption is chosen to simplify the analysis but can be relaxed by adding more templates to the ideal observer formulation.

contrast thresholds are then converted to signal energy (E), which is plotted against the noise spectral density (N) (noise energy per unit bandwidth). We call this graph an E–N plot. Appendix A provides the definitions of contrast, signal energy, and noise spectral density. In Appendix B, we show that the E–N plot is a straight line passing through the origin for the ideal observer in our recognition and detection tasks. The slope of this line is the signal-to-noise ratio required for the threshold performance level. Empirically, we found that the E–N plots of our human observers were also close to straight lines, but differed in two important ways from the ideal observer's plot: the human E–N plots had higher slopes, indicating that humans require a higher signal-to-noise ratio to reach threshold performance, and they had a negative x-intercept, indicating the presence of equivalent noise.

We define the sampling efficiency (Eff_s) to be the ratio of slopes of the E–N plots, i.e. $Eff_s =$ Slope_ideal/Slope_human. As shown in Appendix B, the sampling efficiency for simple identification tasks is the same as the proportion of the total number of samples (which can be pixels or other "features") that the ideal observer would use to achieve the same performance level as the human observer. For example, a sampling efficiency of 10% indicates the human observer is performing at a level which an ideal observer can achieve by using only 10% of the available samples.

Measuring sampling efficiency requires at least two thresholds to establish the slope of an E–N plot. Alternatively, a single threshold can be used to measure *total efficiency*. At a given noise level, total efficiency (Eff) is defined simply as the threshold signal energy of the ideal observer divided by that of the human observer, that is, $Eff = E$_ideal/E_human. This is equivalent to the squared ratio of RMS contrast threshold of the ideal observer to that of the human observer, i.e. $Eff = (C$_ideal/C_human$)^2$. If human observers had no equivalent noise, their total efficiency would be equal to their sampling efficiency. When equivalent noise is nonzero, total efficiency is lower than sampling efficiency. However, when the noise added to the stimuli (external noise) is much greater than the noise internal to human observers, the effect due to equivalent noise becomes insignificant, and total efficiency approaches sampling efficiency (see Appendix B).

METHOD

Apparatus

Targets and noise were generated separately on two Apple monochrome monitors, allowing for independent control of luminance. The monitors were controlled by a Macintosh IIx computer through two 8-bit Apple video boards. The monitors had a luminance range of 0–90 cd/m^2. Each monitor was 640 pixels horizontally by 480 pixels vertically. At the viewing distance of 1.72 m, each pixel subtended 0.66 min-arc. Stimuli on each monitor were restricted to a centered region of 452 pixels horizontally by 442 pixels vertically, subtending 5.0 by 4.9 deg, respectively. Accurate contrast control was achieved with video attenuators and the associated Video Toolbox software described by Pelli and Zhang (1991).

The object display contained a bright target on a dark surround. The noise display contained luminance variations around a mean level. The images on the two monitors were superimposed optically (Fig. 3) using mirrors, a beam-splitter, and a neutral density filter. The neutral density filter attenuated the luminance of the object display relative to the noise display, thereby reducing the contrast of the object when superimposed with the noise.

Stimuli

Object targets and noise. The targets were four simple geometric 3-D objects, given the names "wedge," "cone," "cylinder," and "pyramid" (Fig. 4). The objects were rendered white-on-black in orthographic projection on a Stellar GS2000 graphics computer. After attenuation by the optical apparatus, the luminance of the black background on the target screen was 0.38 cd/m^2, and the luminance of the brightest pixels of the target images ranged from 0.47 to 1.88 cd/m^2, depending on the contrast. Each object was rendered from 8 different viewpoints, randomly selected for each object from a viewing sphere (Fig. 1). In order to avoid clustering of viewpoints at the poles, the range was limited to latitudes between +85 and −85 deg. The entire 360 deg of longitudinal range was included. The objects were placed in such a way that their axes of symmetry did not coincide with the axis joining the poles of the viewing sphere.

The objects were rendered in four conditions: Lambertian shading, large and small silhouettes, and line drawings. In all but the small silhouette condition, the objects subtended 2.8 deg on average. Lambertian shading is the simplest shading model, in which the luminance at each point of an object surface is proportional to the cosine of the angle between the direction of the light source and the surface normal. This model produces no specularity and corresponds to the appearance of a matte surface. Images with Lambertian shading were produced with 256 gray levels, assuming a point light source situated at infinity, 21 deg up and 15 deg left from the line of sight measured at the center of the viewing sphere.

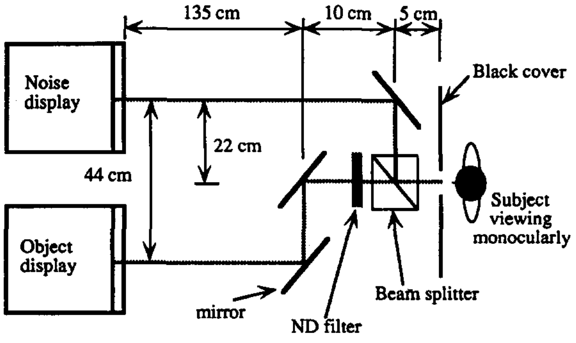

FIGURE 3. The apparatus used for the experiments.

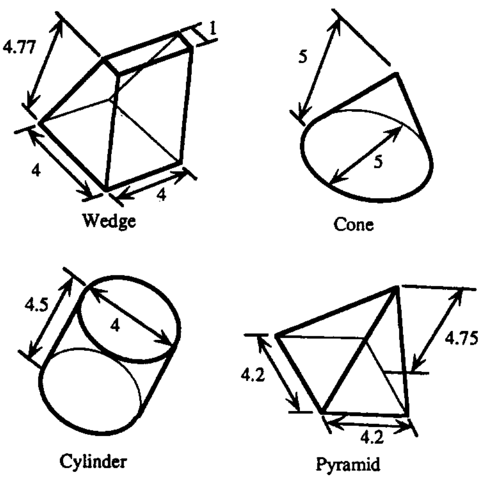

FIGURE 4. Relative dimensions of the four objects used in the object recognition and detection experiments.

A silhouette was produced by setting all pixels inside the object's bounding contour to a uniform white. Silhouette objects had two sizes: "large" silhouettes subtending 2.8 deg, and "small" silhouettes subtending 0.9 deg. The small silhouettes had the same shape as their larger counterparts and differed only in size.

A line drawing was produced by marking all of the luminance discontinuities in a shaded image by one-pixel-wide lines, and then thickening the lines to about 4 pixels by blurring and thresholding. As with the silhouettes, only two gray levels were required to render the line drawings.

To produce the static Gaussian luminance noise field, we used a pseudo-random number generator (Pelli Video Toolbox) running on a Macintosh IIx computer. After attenuation by the optical apparatus, the noise had a mean luminance of 7.12 cd/m^2. Three noise levels were used, with the standard deviations set at 0, 2.67, and 3.56 cd/m^2, resulting in noise spectral densities of 0, 15.2, and 27.0 μ(deg^2) respectively. The two-sided vertical and horizontal bandwidths of the static noise were 91.4 c/deg (see Appendix A for definitions).

Our computer was too slow to generate an entirely fresh field of noise on every trial. For each block of trials at a new noise level, the computer generated an array of 904 by 442 noise samples, exactly twice as large as the noise field. On each trial, a new noise field was derived from this array using two uniform random variables, d and x. d specified a starting position into the noise array from which the noise field for the trial was to be extracted, and x was used to ensure that accidental local spatial features would not survive from trial to trial. Specifically, a noise pattern of size $(452 + x)$ by 442 was copied to the noise display starting at the location d of the noise array, with a vertical strip of size x by 442 clipped away from the right. x was uniformly distributed between 0 and 20, and d between 0 and $442 * (452 - x)$ and truncated to a multiple of four to speed up the noise pattern transfer from the noise array to the screen.

When the noise and object displays were optically superimposed, the observer saw a noisy target on a noisy background of mean luminance 7.50 cd/m^2. The peak luminance of the target ranged from 0.09 to 1.50 cd/m^2 above the background depending on the contrast setting.

Letter targets and noise. We generated 26 uppercase letters—Geneva font with a size of 100 points—on a Macintosh computer using the System 7.0 True-Type font manager. We studied efficiency for recognition of letters rendered in two resolutions: *fine* resolution of 90 pixels per deg, matching the resolution of our objects, and *coarse* resolution of 30 pixels per deg to match conditions of studies by Pelli and colleagues (Burns & Pelli, 1991; J. A. Solomon & D. G. Pelli, private communication, 1993). Coarse letters were generated by uniformly sub-sampling fine letters from 90 by 90 pixels to 30 by 30 pixels, and then replacing each pixel with 3 by 3 pixels of the same luminance. Both the fine and coarse letters were rendered as uniform white pixels on a black background and had an average size of 1 deg. Figure 5 shows these stimuli.

The Gaussian noise field for the letter recognition experiment differed from the noise field in the object experiment in two ways. First, the displayed noise field was smaller, measuring 2.8 by 2.8 deg (252 by 252 pixels). Second, the noise, like the letter targets, was rendered to two resolutions: 90 pixels per deg (*fine* resolution) and 30 pixels per deg (*coarse* resolution). In the latter, each noise sample consisted of 3 by 3 pixels of equal luminance. For both resolutions, the standard deviation of the noise was 3.56 cd/m^2. The noise spectral densities were 27.0 μ(deg^2) at the fine resolution and 243.0 μ(deg^2) at the coarse resolution.

Procedure

Object-recognition experiment. We took three steps to ensure that subjects were very familiar with the objects. Prior to testing, subjects viewed and handled 3-D cardboard models of the four objects. The subjects then previewed all the computer-rendered test views of the objects at high contrast, stepping through the sequence at their own pace. Finally, the test views of all objects were again displayed in a slide-show format, each for one second at the starting contrast for that block of trials. After the slide-show, the subjects spent at least 3 min adapting to the mean luminance of the noise field.

There were twelve experimental conditions, each consisting of one of the four rendering conditions and one of the three noise levels. Each block of trials contained one experimental condition. In a trial, one of the 32 target images was selected at random with equal

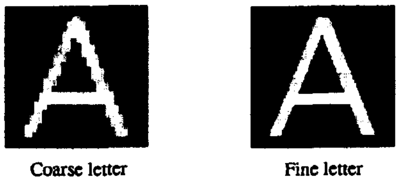

FIGURE 5. "Coarse" and "fine" resolutions for the letter recognition experiment.

probability and presented with added noise for 1 sec. The observer had to identify the target as a "wedge", "cone", "cylinder", or "pyramid" by pressing one of four buttons. No feedback was given. Between trials, the subjects saw a uniform display of 7.50 cd/m^2, equal to the mean luminance of the noisy background.

Contrast thresholds for object recognition were obtained in two phases. First, we used a one-up, three-down staircase to provide an initial estimate of the contrast level yielding 79% correct (Wetherill & Levitt, 1965). The step size was 10% of the current contrast value. The staircase terminated after fifteen reversals (in about 70 trials), and the initial estimate of threshold contrast was taken as the mean of the contrasts at the last 12 reversals.

In the second phase, we used the method of constant stimuli to obtain a more accurate threshold estimate. For each experimental condition, there were six blocks of trials at fixed contrasts. In principle, the observer had access to information specifying the contrast level being tested, consistent with assumptions underlying computation of ideal performance. Two blocks of 100 trials each were run at contrasts 30% above and 30% below the staircase threshold contrast. A new estimated contrast threshold (79% criterion) was calculated from the resulting data using a linear interpolation on a graph of percent correct vs log contrast. This process was repeated twice. In all, six blocks of trials (600 trials) using fixed contrast stimuli were run, and three estimates of contrast threshold were obtained for each condition. The average of these three estimates was taken as the final estimate of the threshold for a given experimental condition.

Object-detection experiment. The apparatus and stimuli were identical to those in the object recognition experiment. Three rendering conditions were tested: large silhouettes, small silhouettes, and line drawings. Only the high noise level was used [27.0 μ(deg^2)]. We omitted the shaded rendering because the recognition efficiency was similar for shaded objects and large silhouettes. We used only the high noise level since the recognition results indicated that this noise level was high enough for the measurement of *total efficiency* to be a good approximation to *sampling efficiency* (see Theory section above).

Each block of trials consisted of one rendering condition. On half of the trials, one of the 32 images was randomly selected and presented with noise; on the other half, only the noise field was presented. The subject indicated whether an object was present or not, with the goal of maximizing percent correct. No feedback was given.

The same one-up, three-down staircase was used to estimate the contrast threshold (79% correct criterion), but there was no second phase involving the method of constant stimuli. Thresholds were obtained from three staircases, and the results were averaged.

Letter-recognition experiment. Three of the four possible conditions were tested involving fine and coarse resolutions of letters and noise: fine letters in fine noise (flfn), coarse letters in fine noise (clfn), and coarse letters in coarse noise (clcn). The fourth condition, i.e. fine letters in coarse noise (flcn), was not tested because each coarse noise pixel would cover 3 by 3 fine-letter pixels, and the values of the noise samples would no longer be independent at each pixel. This situation would require a more complex ideal observer.

In a trial, one of the 26 letters was selected at random with equal probability and presented in noise. The letter and noise remained on the screen until the subject responded by naming one of the 26 letters.

The same one-up, three-down staircase was used to estimate contrast threshold for letter recognition (79% correct criterion). Thresholds from three staircases were averaged for each subject and test condition.

Human subjects

Authors WB and BT were subjects for all experiments. Both had normal vision with Snellen acuity in the tested eye of 20/20. They were well practiced on all three tasks. All experiments reported here were undertaken with the understanding and consent of each subject.

Ideal-observer simulation

Because there is no closed-form solution for equation (4), we simulated the ideal observer on a Stellar GS2000 workstation, equipped with four floating point vector processors. Fresh noise patterns were generated for each trial (i.e. no reliance on a fixed noise pool as in the human experiments).

In a trial, the program based its decision on the likelihood function, equation (4). For object recognition, i ranged from 1 to 4, and j ranged from 1 to 8 (i.e. four objects and eight views). For detection i ranged from 1 to 2 (object present or absent), and j ranged from 1 to 32. In both cases, the template size was 452 by 442 pixels. For letter recognition, i ranged from 1 to 26, and j was equal to 1. For the two fine noise conditions in letter recognition, the template size was 252 by 252 pixels, which was the same size as the noise field, with the central 90 by 90 pixels containing the letter. For the coarse noise condition, the template size was 84 by 84 pixels, with the central 30 by 30 pixels containing the letter. This is because the screen pixels in each 3 by 3 group were identical and can be represented by just a single sample for the ideal observer.

The ideal observer used a binary-search algorithm to find its 79%-correct contrast threshold (with $\pm 0.3\%$ tolerance). The search began with a contrast range having a minimum of zero and a maximum equal to the human threshold value. The program first ran 600 stimulated trials (1000 for letter recognition) at the contrast value midway between the minimum and maximum. If performance was within $79 \pm 0.3\%$ correct, then the process terminated, and this value was taken as the contrast threshold. Otherwise, if the performance was above 79% correct, then the maximum contrast value was set to the current contrast, and if it was below, then the minimum contrast value was set to the current contrast. The

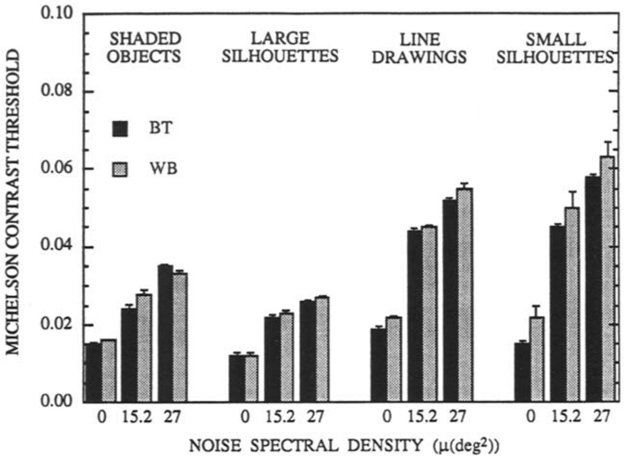

FIGURE 6. Human Michelson contrast thresholds for object recognition at 79% correct performance criterion. Error bars indicate 1 SE (some may be too small to be visible). See Appendix A for definition of Michelson contrast for images of objects.)

process repeated itself. Normally the program converged in less than ten iterations.

Statistical analysis

For each experiment, we used an ANOVA to determine if a particular factor (e.g. rendering condition) had an effect on the threshold or efficiency results. A Tukey HSD (honest significant difference) test was used to make comparisons between conditions within an experiment. To compare thresholds and efficiencies across experiments, we used a paired two-tailed t-test. We used $\alpha = 0.01$ as the criterion for an effect to be significant for all three tests.

RESULTS

Object recognition

Figure 6 shows the contrast thresholds (Michelson definition, Appendix A) for recognizing objects at the three noise levels and four rendering conditions for the two human subjects. There were slight but statistically significant differences in contrast thresholds between subjects. We found a large effect of noise level, indicating that noise significantly elevated thresholds, and a large effect of rendering condition. Across noise levels, thresholds for the line drawings and small silhouettes were similar, and about a factor of 2 higher than those for the shaded images and large silhouettes. We also observed a significant interaction between rendering condition and noise level.

As a first step in computing efficiency, it is informative to replot these data as threshold signal energy E versus noise spectral density N (E–N plots), as shown in Fig. 7 (see Appendix A for definitions of these quantities). Straight lines provided good fits to the E–N plots for the two human subjects ($r > 0.974$) and the simulation data for the ideal observer ($r > 0.997$). These data can be used to estimate subjects' *equivalent noise*, *sampling efficiency*, and *total efficiency* (see the subsection above on Definition and Interpretation of Efficiency). Table 1 displays the sampling efficiencies and total efficiencies for both subjects and shows that the two measures of efficiency were very similar. Henceforth, we will use the term "efficiency" and cite figures for total efficiency unless otherwise specified.

Because efficiency describes human performance relative to ideal performance, it can sometimes give a different picture from threshold measures. For example, although small silhouettes have high thresholds (Fig. 6) indicative of low contrast sensitivity, they have the highest efficiencies (Table 1). Overall, efficiencies for recognizing objects were quite low, ranging from 2.57 to 8.38% across rendering conditions and subjects. Because of the square-law relationship between efficiency and RMS contrast (see Appendix A), these values correspond to human RMS contrast thresholds that were about three to six times higher than ideal thresholds in the same task.

The highest efficiency was for small silhouettes, averaging 7.84% for the two subjects. The lowest efficiency was for line drawings, averaging 2.69%. The average efficiencies for shaded objects and large silhouettes were 3.28 and 4.51%, respectively. Rendering condition significantly affected recognition efficiencies, but there was no significant difference between subjects.

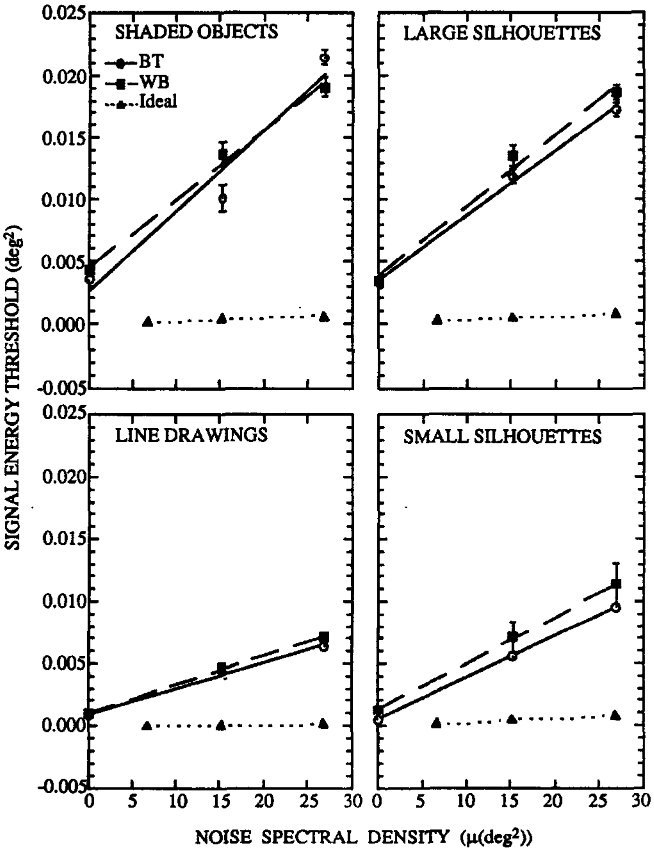

FIGURE 7. Signal energy threshold as a function of noise spectral density (E–N plot) for the object recognition task in four rendering conditions. Lines represent linear least-square fits to the data. Error bars are ± 1 SE in length, and some are smaller than the plot symbols.

TABLE 1. Human object-recognition efficiency under different rendering conditions for both subjects and their average

	Subject BT		Subject WB		Average	
	Sampling efficiency (%)	Total efficiency (%)	Sampling efficiency (%)	Total efficiency (%)	Sampling efficiency (%)	Total efficiency (%)
Shaded objects	3.33 ± 0.07	3.09 ± 0.09	3.96 ± 0.24	3.47 ± 0.16	3.65 ± 0.18	3.28 ± 0.12
Large silhouettes	5.15 ± 0.32	4.68 ± 0.17	4.71 ± 0.08	4.33 ± 0.12	4.93 ± 0.18	4.51 ± 0.12
Line drawings	3.03 ± 0.06	2.82 ± 0.06	2.98 ± 0.06	2.57 ± 0.12	3.01 ± 0.04	2.69 ± 0.08
Small silhouettes	8.47 ± 0.16	8.38 ± 0.16	7.99 ± 2.06	7.30 ± 1.39	8.23 ± 1.15	7.84 ± 1.09

± intervals indicate ± 1 SE.

Object detection

One key feature of our object recognition results is that recognition efficiency is low. Intuition suggests that object-recognition efficiency may be limited by detection efficiency. If detection efficiency is low, say 10%, then it is as if only 10% of the stimulus samples are used effectively in detection (see Appendix B). Since additional information loss would probably be incurred in using these samples in recognition, we might expect recognition efficiency to be lower than 10%. As described in the next two paragraphs, recognition efficiency was lower than detection efficiency for line drawings but not for silhouettes.

Figure 8 shows the detection data in three formats: contrast thresholds in panel (a), signal energy thresholds in panel (b), and efficiencies in panel (c). The large and small silhouettes had the lowest threshold contrasts, averaging 1.85 and 2.32% respectively. However, when measured in terms of signal energy, the threshold for large silhouettes was more than twice that for small silhouettes and was the highest of the three conditions. The effect of rendering on contrast thresholds was significant, and there was no significant difference between subjects. The highest efficiencies were for detection of small silhouettes and line drawings, averaging 4.74 and 4.62% respectively. The efficiency for detection of large silhouettes was much lower, 1.53% on average. The effect of rendering on efficiencies was also significant.

Averaged across subjects, the detection efficiencies were significantly lower than recognition efficiencies for both silhouette conditions (1.53 vs 4.51% for large silhouettes; 4.74 vs 7.84% for small silhouettes). However, efficiency for detecting line drawings was significantly higher than efficiency for recognizing them. These results showed that detection efficiency does not impose a strict upper limit on recognition efficiency. We shall address the implications of these results in the Discussion.

Letter recognition

The highest efficiency we obtained for object recognition was about 7.8%, substantially lower than values reported in the literature for letter recognition. Is this gap a result of methodological differences or are there intrinsic differences between letter recognition and object recognition?

Parish and Sperling (1991) obtained their highest efficiency (42%) for spatially band-pass filtered letters with a center frequency of 1.5 cycles per letter. Our object stimuli were unfiltered. It is possible that humans rely selectively on a preferred band of spatial frequencies in object recognition, and fail to use information in other bands. We address the issue of the role of spatial frequency in object recognition in a separate paper (Braje, Tjan & Legge, 1995).

Pelli and colleagues have measured efficiency for recognizing unfiltered letters, reporting values between 12% and 20% (Burns & Pelli, 1991; Solomon & Pelli, 1994). In their experiments, Burns and Pelli (1991) used letters and noise with a coarser pixel resolution than the objects and noise in our main experiment.

In an effort to bridge the gap between our measurements of efficiency for object recognition and Pelli *et al.*'s measurements of efficiency for letter recognition we (i) attempted to replicate Pelli *et al.*'s results for letter recognition, and (ii) investigated the effects of pixel resolution on efficiency for letter recognition. The flfn condition (see Methods) matched the conditions in our object recognition experiment, the clcn condition matched the conditions used by Pelli and colleagues, and the clfn condition was in-between the two.

Figure 9 shows our letter-recognition results in terms of threshold contrasts, threshold signal energies, and efficiencies. A Tukey HSD test confirmed that recognition efficiency was not significantly different between the "flfn" condition and the "clfn" condition, but it was significantly lower for the "flfn" condition than for the "clcn" condition. (The same comparisons were also true in terms of contrast threshold and signal energy threshold.) Thus, the noise bandwidth (i.e. pixel resolution), but not the letter bandwidth, had an effect on efficiency.

Average efficiency was 16.3% for the low noise bandwidth ("clcn" condition), in good agreement with Pelli and colleagues. For the high noise bandwidth ("clfn" and "flfn" conditions) average efficiency was 12.5%. This is significantly higher than efficiency for recognizing small silhouettes (7.84%), which is our most comparable task as well as the one with the highest object-recognition efficiency. These findings reveal an efficiency advantage for letters over simple 3-D objects.

Summary of results

Table 2 summarizes the average efficiencies for object recognition, object detection, and letter recognition. In

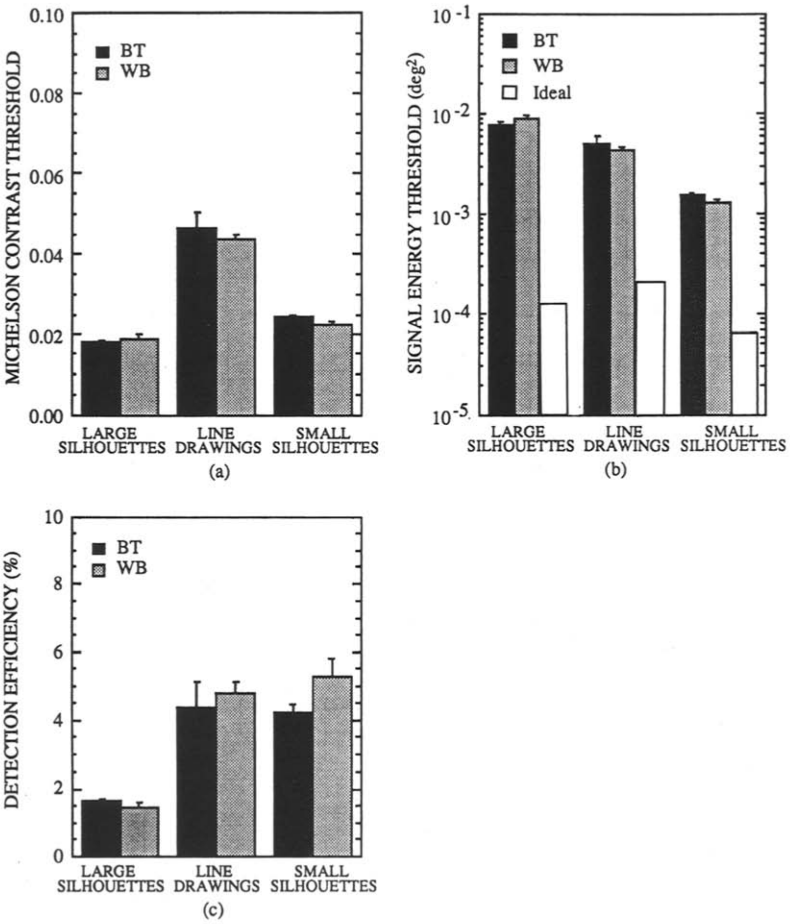

FIGURE 8. Human object-detection performance in noise [27.0 μ(deg^2)] at 79% correct: (a) Michelson contrast threshold, (b) signal energy threshold (log scale), (c) total detection efficiency. Error bars representing 1 SE are plotted for all human data (some may be too small to be visible).

general, object-recognition efficiency was low. Letter-recognition efficiency was higher than object-recognition efficiency for all rendering conditions, and object-detection efficiency was either higher or lower than object-recognition efficiency, depending on rendering condition.

DISCUSSION

Efficiency for object recognition was quite low, approximately 3–8%. These values are an order of magnitude lower than values of 30–60% reported for some other visual-information processing tasks. The low efficiency indicates that performance in object recognition is not limited by the information content of the stimulus, but by information-processing limitations intrinsic to the observer. The relatively small impact on efficiency of our different rendering cues may also suggest that the sources of inefficient information processing occur early, prior to specialized analysis of distinct cues.

Why is efficiency for object recognition low? In the remainder of this Discussion, we will draw on our results to consider seven factors that might limit object-recognition efficiency: internal noise, rendering condition, grouping across views, learning, spatial uncertainty, stimulus size and detection efficiency. Our data suggest that the last three impose a much larger limit on efficiency than the rest.

Internal noise

As described in the Introduction, inefficient human performance can result from two distinct generic sources: *equivalent noise* and *sampling efficiency*. Could the low efficiency we found for object recognition be explained by high levels of equivalent (internal) noise? The answer is no. Our *E–N* plots provided estimates of sampling efficiency that are not affected by equivalent noise. In addition, our measures of total efficiency in the highest noise condition were quite close to our estimates of sampling efficiency from the *E–N* plots. This means that our external noise levels were high enough to swamp the effects of the internal noise. We conclude that random sources of internal white noise cannot account for the low efficiencies we measured.*

*We qualify this conclusion by noting that non-white internal noise or, equivalently, a spatial filter applied to the images, might play a role in limiting efficiency. We address the issue of filtering in a separate paper (Braje, Tjan & Legge, 1995).

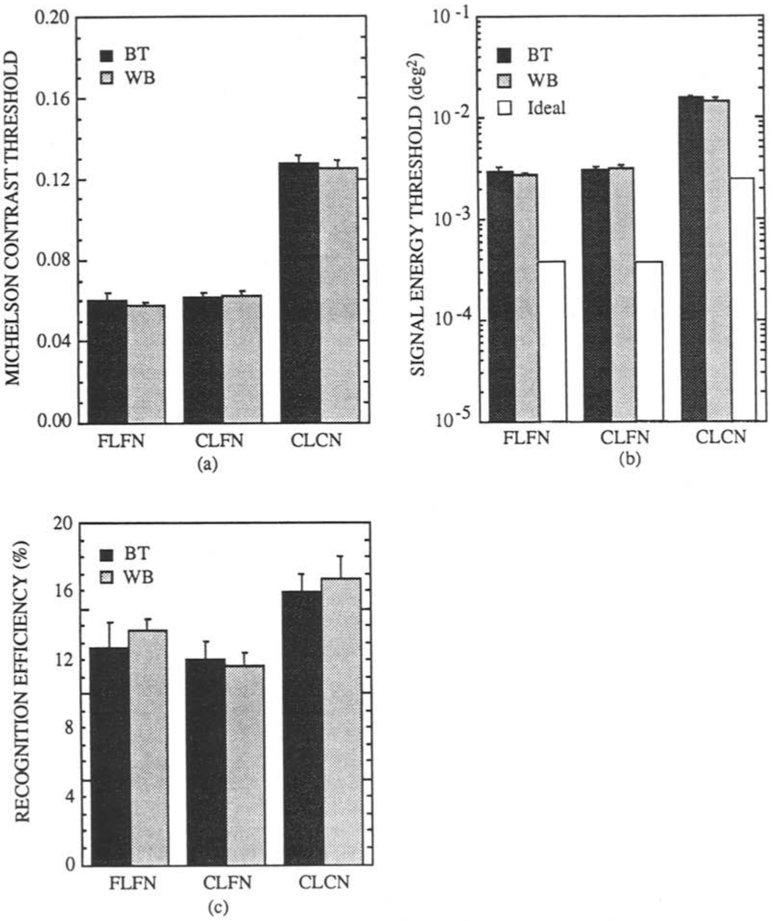

FIGURE 9. Human letter-recognition performance in noise at 79% correct with different letter/noise resolutions: (a) Michelson contrast threshold, (b) signal energy threshold (log scale), (c) total efficiency. Error bars representing 1 SE are plotted for all human data (some may be too small to be visible).

Effect of rendering

It is instructive to examine the performance of the ideal observer with the different rendering cues before discussing human performance. We use the ideal observer as an "information meter" because its signal-energy thresholds depend only on the information content of the stimuli: the lower the threshold, the higher the information content. In Fig. 10, the ideal observer's *E–N* curves for large and small silhouettes were very similar, showing that the information content is about the same. The thresholds for shaded images were lower than for silhouettes, consistent with our intuition. Surprisingly, however, the thresholds for line drawings were much lower than for shaded images. This indicates, paradoxically, that a line drawing (contour information only) contains more information than a shaded image (contour information plus luminance information). The paradox can be resolved by noticing that replacing the edges (lines) in a line drawing by shading information results in a reduction of contrast along these edges. If these edges are important for recognition, the addition of shading information may be more than offset by the loss of edge information.

For all four rendering conditions, human efficiency for recognition is low. Recognition efficiency for line drawings (2.70%) was the lowest, indicating that humans cannot make efficient use of the extra information presented in line drawings. In particular, efficiency for recognizing line drawings was lower than for recognizing silhouettes (4.51%), implying that humans do not use all the information carried by the internal contours.

TABLE 2. Total efficiency averaged across subjects for object recognition, object detection, and letter recognition

Object rendering	Object recognition	Object detection	Letter & noise resolutions	Letter recognition
Shaded objects	3.28 ± 0.12%			
Large silhouettes	4.51 ± 0.12%	1.53 ± 0.08%	CLCN	16.3 ± 0.78%
Line drawings	2.69 ± 0.08%	4.62 ± 0.37%	CLFN	11.8 ± 0.57%
Small silhouettes	7.84 ± 1.09%	4.74 ± 0.35%	FLFN	13.2 ± 0.78%

± intervals indicate ± 1 SE.

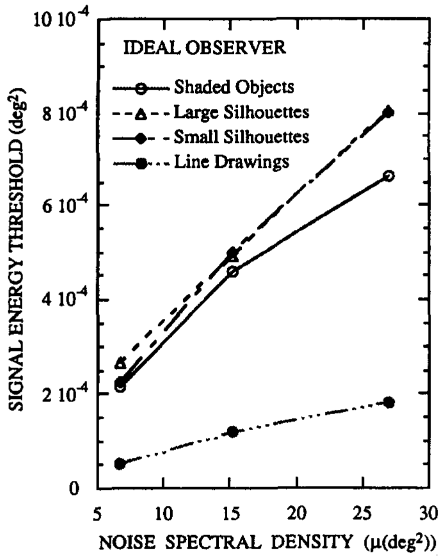

FIGURE 10. The ideal observer's signal energy threshold as a function of noise spectral density for recognizing objects in different rendering conditions.

Recognition efficiency for shaded images (3.28%) was also slightly lower than for silhouettes, implying that humans do not use all of the available shading information. On the whole, differences in efficiency due to rendering were small and cannot explain the low efficiencies we obtained.

Effect of size

The small (0.9 deg) silhouettes were scaled versions of the large (2.8 deg) silhouettes, containing fewer pixels and shorter bounding contours. As evident in Fig. 10, this size scaling had very little effect on the ideal observer's recognition thresholds, so the amount of stimulus information for object recognition is about equal for the two types of silhouettes. Nevertheless, efficiency for recognizing small silhouettes was about twice that for large silhouettes, 8.0% compared with 4.5%. This means that our subjects were less than ideal in integrating across space to extract information from the large silhouettes. In the following two paragraphs we rule out probability summation as an explanation for this inefficient behavior.

For visual detection, the ideal spatial summation relation (based on quantum catch) between threshold contrast and stimulus area is an inverse-square-root law. Except for very restricted domains, spatial summation for luminous disks (Barlow, 1958) and gratings (Legge, 1978; Robson & Graham, 1981) is suboptimal, showing a weaker dependence of threshold on area. Detection efficiency for gratings drops rapidly as the area (scaled for spatial frequency) increases (Kersten, 1984).

Probability summation is a model often invoked to account for the suboptimal summation effects in grating detection (Legge, 1978; Robson & Graham, 1981). However, it cannot account for the lower recognition efficiency obtained with large silhouettes. Recognition, unlike detection, cannot rely on isolated responses from localized independent detectors, which are assumed in probability summation. Recognition normally requires detection of distinguishing features and their spatial relations. While probability summation might play a role at the level of detecting visual features, the independence assumption precludes its application in determining their spatial relations. From our size data, we conclude that spatial integration in human object recognition is inefficient, but the effect cannot be ascribed to probability summation.

Effect of spatial uncertainty

The ideal observer takes advantage of information that humans may ignore because it is not of ecological significance. Our targets, for example, were always presented at fixed pixel locations on the display. Humans may fail to make use of the exact positional information available to the ideal observer. They may encode object information in a viewpoint invariant manner (e.g. Bülthoff & Edelman, 1992; Cooper, Biederman & Hummel, 1992; Hasselmo, Rolls, Baylis & Nalwa, 1992) in which at least some position-specific information is lost. This would result in reduced efficiency.*

Positional uncertainty can be introduced by displaying an object at random locations. The corresponding ideal observer is one who keeps several duplicate sets of image templates, one set for each possible location. The ideal observer's contrast threshold rises as the number of possible locations increases.

We used small silhouettes as targets to study the effects of positional uncertainty on object recognition. Before displaying a target, we randomly translated it by up to plus or minus 50 pixels (0.55 deg) in both the x and y-directions from its home position. The number of allowable translations, which we use as an index of the spatial uncertainty, ranged from 1 (no uncertainty) to 1000 (large uncertainty). Each of the translations was applied to all of the 32 original templates.

Figure 11 plots the threshold signal energy of the ideal observer as a function of spatial uncertainty (i.e. the number of translations). The ideal observer's threshold increased gradually as spatial uncertainty increased from 1 to 1000. When spatial uncertainty was 1000, the ideal observer's threshold energy was double that of the case when there was no spatial uncertainty.

If human vision does not code exact spatial positions of objects, then increasing spatial uncertainty in the stimuli will not affect human thresholds. Human data (Tjan, Braje & Legge, 1994) showed that recognition thresholds were unaffected for translational uncertainties from 1 to 1000. Because the ideal threshold doubled in the same range, human recognition efficiency also doubled.

We conclude that humans cannot make use of exact positional information in object recognition. We

*Clearly, some 3-D positional information is encoded because people can grasp objects swiftly and effectively. Goodale and Milner (1992) have proposed that one visual pathway underlies motor activity and retains precise positional information, while another visual pathway underlies perceptual object recognition and retains much less positional information.

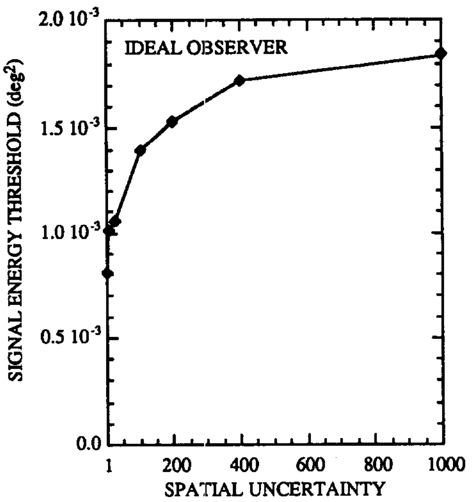

FIGURE 11. The ideal observer's signal energy threshold as a function of spatial uncertainty measured as the number of possible positions on the screen an image of a small silhouette object can appear.

speculate that the same may be true for absolute information regarding object pose (2-D and 3-D rotations).

Grouping across viewpoint

With comparable stimulus conditions, our results on letter recognition revealed an efficiency advantage of letters over simple 3-D objects (12.5 vs 7.84%). One plausible explanation is that variation of viewpoint plays a greater role in object recognition than in letter recognition. In our object-recognition experiment, there were eight views (requiring the ideal observer to have eight templates) for each of the four objects. In the letter-recognition experiment, there was only one "view" (requiring only one template) for each of the 26 letters (the total number of target images, 32 vs 26, was very similar). Assuming each view of each stimulus is presented with equal probability, the ideal decision rule for recognition is to select the label (one of the four objects, or one of the 26 letters) to maximize the sum of the likelihoods of the views associated with that label. If humans fail to sum likelihoods across viewpoints, but instead base their recognition decision solely on the "view" with maximum likelihood, their performance for the object-recognition task will be suboptimal. This is a version of the "maximum-of" rule described by Pelli (1985). On the other hand, the "maximum-of" strategy would be optimal in our letter-recognition experiment where there is only one "view" per letter. We estimated the reduction in efficiency for the "maximum-of" strategy by running a sub-ideal simulation, using the small silhouettes as targets. The sub-ideal observer operated like the ideal observer, except that it used the "maximum-of" decision rule. The threshold performance of the "maximum-of" sub-ideal observer yielded an efficiency of 95%.

We cannot be sure whether humans use the "maximum-of" strategy or not, but use of this strategy might account for the small relative difference between efficiency for object recognition and letter recognition. On the other hand, the mere 5% reduction in efficiency is too small to sufficiently explain why object-recognition efficiency is low in absolute terms.

Learning

The degree of familiarity with the stimuli may affect efficiency for recognition. Letters may be among the most familiar stimuli seen by subjects (often educated adults) who participate in psychophysical experiments, and this can be another reason for letter-recognition efficiency to be higher than object-recognition efficiency. In a recent study, Burns, Farell, Solomon and Pelli (1993) showed that contrast thresholds for letter identification were $\sqrt{2}$ (i.e. 0.15 log units) lower for adults than second graders.

In a pilot study, we also observed small learning effects in object recognition; the performance of naive subjects improved slightly with practice. The data in all of the figures of this paper, however, were collected with highly practiced subjects, who were also responsible for constructing the targets.

Object recognition and object detection

Must object-recognition efficiency always be lower than detection efficiency? The answer is no. Imagine a primitive visual system that can only sample a single pixel per trial from a computer display. Suppose a recognition task involves discrimination between target A, a 10 by 10 square of white pixels, and target B which is identical to A except for one black pixel in the upper left corner. There is only one informative pixel for this recognition task. If the visual system can sample strategically, it will look at the informative pixel and achieve 100% efficiency. In a detection task, all 100 pixels are useful, but the sampling capacity of the system is only one pixel. The system will therefore have a detection efficiency of only about 1%. From this example, it is clear that recognition efficiency can be higher than detection efficiency. This can happen only if there is some form of strategic (nonrandom) sampling that is well-adapted to the recognition task.

For both large and small silhouettes, we found that detection efficiency was substantially lower than recognition efficiency (see Table 2). For silhouette detection, all of the white pixels are equally informative. The fact that recognition efficiency is higher implies that nonuniform subsampling takes place in the encoding of such stimuli. This is consistent with the popular view that object recognition is preceded by a stage of low-level feature extraction (Marr, 1982; Biederman, 1987; Malik, 1987; Ullman, 1989). Candidate features include edges, curves, corners, and junctions.

While recognition efficiency is not strictly limited by the efficiency for detecting an object, it may nonetheless be constrained by the efficiency for detecting low-level features of an object. Several low-level luminance features have been proposed to underlie human object recognition, including zero crossings (Marr & Hildreth, 1980), peaks and valleys (Mayhew & Frisby, 1981), and centroids (Watt & Morgan, 1983). To our knowledge, there are no studies of efficiency for detecting edges or other feature labels derived from these properties of luminance waveforms. Legge *et al.* (1987) reported sampling efficiency of 14% for contrast discrimination of small disks (13.6 min-arc in diameter) in static noise. If low-level features are often defined by localized contrast discontinuity (e.g. edge segments), then the result of Legge *et al.* (1987) will impose an efficiency limit on the detection and extraction of these features.

This limit on recognition efficiency for low-level features can be carried over to recognition of an object if the low-level features need to be detected and committed to a discrete label at an early stage, as suggested by most contemporary models of object recognition (Lowe, 1987; Biederman, 1987; Huttenlocher & Ullman, 1990). There are at least two empirical results that pose a problem for this view. First, Burgess *et al.* (1981) reported very high sampling efficiencies, around 70%, for detecting a target consisting of a few cycles of a 5 c/deg sine-wave grating. Second, Parish and Sperling (1991) have reported efficiencies as high as 42% for recognizing band-pass-filtered letters. These findings raise the possibility that human object recognition could not have committed a low-level (luminance discontinuity) feature to a label at an early stage. It is possible that human feature analysis relies on restricted bands of spatial frequency. Perhaps features are extracted from the output of specific spatial-frequency filters without being sufficiently identified to warrant a discrete label. If so, recognition efficiency should depend on the spatial frequency bandwidth of the image. We take up this issue in a separate paper (Braje *et al.*, 1995).

We conclude that visual subsampling, as revealed by low detection efficiency, does not completely explain low efficiencies for object recognition because the subsampling may not be uniform; however, it remains possible that recognition efficiency for the salient low-level features is limited by contrast-detection efficiency, and this limit may impose constraints on recognition efficiency for objects.

CONCLUSION

Efficiency tells us how well humans use visual information. Our purpose was to determine how efficiently humans recognize simple 3-D objects. We found that efficiency for object recognition was low (3–8%) compared to values of 30–60% reported for some other visual-information processing tasks, such as symmetry detection and recognition of bandpass-filtered letters. We considered seven factors that might account for low efficiency in object recognition.

The first four of these factors had only small effects on efficiency:

(1) *Internal Noise*: Even when the externally added noise level was high enough to swamp observers' internal noise, recognition efficiencies were low.

(2) *Rendering Condition*: The means for rendering objects on the screen—silhouettes, line drawings, or shaded images—had relatively small (though statistically significant) effects on recognition efficiency. Efficiency was highest in the most information-deprived rendering condition (silhouette).

(3) *Grouping Across Views*: The ideal observer optimizes performance by taking into account the similarity of a stimulus to all possible views of each object. Even if humans fail to use this grouping principle, and simply select an object based on the similarity of the stimulus and a single view, simulation results show that the reduction in efficiency would be small.

(4) *Learning*: There is evidence that exposure to stimuli over a very large number of trials can affect recognition efficiency, but the effects are small. In comparable measurements with letters and objects, we found recognition efficiencies to be slightly higher for letters, a difference which might be due to extra familiarity with letters.

Three other factors appear to have larger effects on efficiency:

(5) *Spatial Uncertainty*: When the objects were presented with 1000-fold positional uncertainty, human threshold energy was not affected but the ideal observer's threshold energy doubled. The result was a doubling of recognition efficiency. This implies that humans do not encode object information in a position-specific manner.

(6) *Stimulus Size*: Recognition efficiency doubled when the target diameter was reduced to about 1/3 of its previous size (2.8–0.9 deg). This size effect cannot be explained solely by the low-level probability-summation process used to explain size effects in detection.

(7) *Detection Efficiency*: Intuitively, it seems evident that there is a connection between efficiency for detection and recognition. This connection is not straightforward, however, because the stimulus features used in detection need not be identical to the stimulus features used in recognition. We found that detection efficiencies were similar to, but sometimes significantly lower than, recognition efficiencies. This finding implies that the visual system is selective in the extraction of information (features) for recognition. The efficiency with which these recognition-relevant features are detected or discriminated is likely to be a major factor limiting recognition efficiency.

REFERENCES

Barlow, H. B. (1958). Temporal and spatial summation in human vision at different background intensities. *Journal of Physiology*, *141*, 337–350.

Barlow, H. B. (1977). Retinal and central factors in human vision limited by noise. In Barlow, H. B. & Fatt, P. (Eds), *Vertebrate photoreception* (pp. 337–358), New York: Academic Press.

Barlow, H. B. (1978). The efficiency of detecting changes of density in random dot patterns. *Vision Research*, *18*, 637–650.

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.

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychological Review*, *94(2)*, 115–147.

Braje, W. L., Tjan, B. S. & Legge, G. E. (1995). Human efficiency for recognizing and detecting low-pass filtered objects. *Vision Research, 35*, 2955–2966.

Brooks, R. A. (1983). Model-based 3-D interpretation of 2-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*, 140–150.

Bülthoff, H. H. & Mallot, H. A. (1988). Integration of depth modules: stereo and shading. *Journal of the Optical Society of America. Part A, Optics and Image Science, 5(10)*, 1749–1758.

Bülthoff, H. H. & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Science, U.S.A. 1989(1)*, 60–64.

Burgess, A. & Barlow, H. B. (1983). The precision of numerosity discrimination in arrays of random dots. *Vision Research, 23(8)*, 811–820.

Burgess, A. E., Wagner, R. F., Jennings, R. J. & Barlow, H. B. (1981). Efficiency of human visual signal discrimination. *Science, 214(4516)*, 93–94.

Burns, C. W. & Pelli, D. G. (1991). Recognition of letters and words in noise. *Investigative Ophthalmology and Visual Science, 32 (Suppl. 4)*, 1042.

Burns, C. W., Farell, B., Solomon, J. A. & Pelli, D. G. (1993). Bayesian perception. *Investigative Ophthalmology and Visual Science, 34, (Suppl. 4)*, 1417.

Cooper, L. & Schacter, D. (1992). Dissociations between structural and episodic representations of visual objects. *Current Directions in Psychological Science, 1*, 141–146.

Cooper, E. E., Biederman, I. & Hummel, J. E. (1992). Metric invariance in object recognition: a review and further evidence. *Canadian Journal of Psychology, 46(2)*, 191–214.

Duda, R. O. & Hart, P. E. (1973). *Pattern classification and scene analysis.* New York: Wiley.

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

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review, 96(2)*, 267–314.

Goodale, M. A. & Milner, D. (1992). Separate visual pathways for perception and action. *Trends in Neuroscience, 15*, 20–25.

Green, D. M. &U Swets, J. A. (1974). *Signal detection theory and psychophysics.* Huntington, New York: Krieger.

Hasselmo, M. E., Rolls, E. T., Baylis, G. C. & Nalwa, V. (1992). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research, 75(2)*, 417–429.

Hecht, S., Shlaer, S. & Pirenne, M. H. (1942). Energy, quanta, and vision. *Journal of General Physiology, 224*, 665–699.

Huttenlocher, D. P. & Ullman, S. (1990). Recognizing solid objects by alignment with an image. *International Journal of Computer Vision, 5(2)*, 195–212.

Kersten, D. (1984). Spatial summation in visual noise. *Vision Research, 24(12)*, 1977–1990.

Kersten, D. (1990). Statistical limits to image understanding. In Blakemore, C. (Ed.), *Vision: Coding and efficiency* (pp. 32–44). Cambridge: Cambridge University Press.

Legge, G. E. (1978). Space domain properties of a spatial frequency channel in human vision. *Vision Research, 18*, 959–969.

Legge, G. E., Kersten, D. & Burgess, A. E. (1987). Contrast discrimination in noise. *Journal of the Optical Society of America. Part A, Optics and Image Science, 4*, 391–404.

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

Liu, Z., Kersten, D. & Knill, D. C. (1995). Object classification for human and ideal observers. *Vision Research, 35*, 549–568.

Lowe, D. G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence, 31*, 355–395.

Malik, J. (1987). Interpreting line drawings of curved objects. *International Journal of Computer Vision, 1*, 73–103.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information.* New York: Freeman.

Marr, D. & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London B, 207*, 187–217.

Mayhew, J. E. W. & Frisby, J. P. (1981). Psychophysical and computational studies towards a theory of human stereopsis. *Artificial Intelligence, 17*, 349–385.

van Meeteren, A. & Barlow, H. B. (1981). The statistical efficiency for detecting sinusoidal modulation of average dot density in random figures. *Vision Research, 21*, 765–777.

Morgan, M. J. & Glennerster, A. (1991). Efficiency of locating centres of dot-clusters by human observers. *Vision Research, 31(12)*, 2075–2083.

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. (1981). The effects of visual noise. Doctoral dissertation, Physiology Department, Cambridge University.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America. Part A, Optics and Image Science, 2(9)*, 1508–1532.

Pelli, D. G. (1990). The quantum efficiency of vision. In Blakemore, C. (Ed.), *Vision: coding and efficiency* (pp. 3–24). Cambridge: University Press.

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

Pentland, A. (1986). Perceptual organization and the representation of natural form. *Artificial Intelligence, 28*, 293–331.

Pentland A. (1989). Shape information from shading: a theory about human perception. *Spatial Vision, 4(2–3)*, 165–182.

Poggio, T., Gamble, E. B. & Little J. J. (1988). Parallel integration of vision modules. *Science, 242(4877)*, 437–440.

Robson, J. G. & Graham, N. (1981). Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research, 21(3)*, 409–418.

Solomon, J. A. & Pelli, D. G. (1994). The visual filter mediating letter identification. *Nature, 369(2)*, 395–397.

Sperling, G. & Landy, M. S. (1989). Kinetic depth effect and identification of shape. *Journal of Experimental Psychology Human Perception Performance, 15(4)*, 826–840.

Tanner, W. P. & Birdsall, T. G. (1958). Definitions of d′ and η as psychophysical measures. *Journal of the Acoustical Society of America, 30(10)*, 922–928.

Tarr, M. J. & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21(2)*, 233–282.

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

Todd, J. T. & Akerstrom, R. A. (1987). Perception of three-dimensional form from patterns of optical texture. *Journal of Experimental Psychology Human Perception Performance, 13(2)*, 242–255.

Todd, J. T. & Bressan, P. (1990). The perception of 3-dimensional affine structure from minimal apparent motion sequence. *Perception & Psychophysics, 48(5)*, 419–430.

Ullman, S. (1989). Aligning pictorial descriptions: an approach to object recognition. *Cognition, 32*, 193–254.

Voorhees, H. & Poggio, T. (1988). Computing texture boundaries from images. *Nature, 333(6171)*, 364–367.

Watt, R. J. & Morgan, M. J. (1983). The recognition and representation of edge blur: evidence for spatial primitives in human vision. *Vision Research, 23(12)*, 1465–1477.

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.

Young, A. W., Hellawell, D. J. & Welch, J. (1992). Neglect and visual recognition. *Brain, 115(1)*, 51–71.

Acknowledgements—This study was supported by NIH Grant EY02857. Partial results of this research were previously reported at the Optical Society of America meeting (1991) and the Association for Research in Vision and Ophthalmology meeting (1993).

APPENDIX A

Definitions

Michelson contrast

Targets were rendered as bright objects on a uniform, dim background. Let L_0 be the background luminance, and $L_{\max}$ be the luminance of the target pixel with maximum luminance in the absence of noise. The *Michelson contrast* of the target is defined to be

$$C_{\text{Michelson}} = \frac{(L_{\max} - L_0)}{(L_{\max} + L_0)}. \tag{A1}$$

RMS contrast

Let L_i be the luminance of the ith pixel of an image. The *contrast of the ith pixel* is defined as

$$C_i = \frac{L_i - L_0}{L_0}. \tag{A2}$$

If there are M pixels in an image, then the *RMS (Root-Mean-Square) contrast* of the image is defined to be

$$C_{\text{RMS}} = \sqrt{\frac{1}{M}\sum_{i=1}^{M} C_i^2}. \tag{A3}$$

The *signal energy* of an image is defined to be

$$E = (C_{\text{RMS}})^2\,(M)\,(\textit{Pixel\_Area}) \tag{A4}$$

where a pixel is assumed to be square, and *Pixel_Area* is measured in degrees of visual angle squared. Thus, the units for E are deg^2. The physical interpretation of E is the total signal energy $[(C_{\text{RMS}})\,(M)]$ per unit 2-sided bandwidth. We define E as such to be consistent with our choice of using noise spectral density (see next) to describe the noise strength.

Spectral density

When independent samples of static Gaussian luminance noise of zero mean and standard deviation σ are added to each pixel of an image, the *noise spectral density* (i.e. noise energy per unit bandwidth) is defined to be the variance of the noise, normalized by the square of background luminance, divided by the 2-sided bandwidth of the noise. That is,

$$N = \left(\frac{\sigma}{L_0}\right)^2 \Big/ \omega \tag{A5}$$

where the 2-sided noise bandwidth ω is given by

$$\omega = \frac{1}{\Delta x\ \Delta y} \tag{A6}$$

in which Δx, Δy are the respective x and y dimensions of a noise pixel in degrees of visual angle. The product $\Delta x\ \Delta y$ equals the *Pixel_Area* of equation (A4).

The term σ/L_0 in (A5) can be thought of as the noise measured in contrast terms (in the sense of A2). We shall use this notion of *contrast noise*, which is linearly proportional to luminance noise, in Appendix B to simplify our arguments.

APPENDIX B

Interpretations of E–N Plot, Sampling Efficiency and Total Efficiency

In this appendix we present informal arguments of three claims made in the paper.

Claim 1: An Ideal Observer's E–N plot for object recognition in static Gaussian contrast noise is a straight line passing through the origin

Any image of M pixels can be represented by a point in an M-dimensional feature space. Each axis represents the contrast of one pixel. A stimulus in our experiments can be thought of as being formed by adding independent contrast noise samples drawn from a Gaussian distribution with zero mean and standard deviation σ/L_0 to the contrast value of each pixel of an object template (signal). All of the possible stimuli produced by adding noise to a particular template form an isotropic multivariate Gaussian cloud with zero covariances in the feature space centered at the point representing the template. For a task involving many objects, the feature space is populated with many clouds, one for each view of each object. The threshold task involves reducing the contrasts of all the objects (signals), leaving the noise characteristics unchanged, until percent correct recognition drops to a criterion value. Signal contrast is changed by multiplying the contrasts of the templates' pixels by a constant before adding the noise. This is equivalent to moving the centers of the clouds along radial lines from the origin. A change in the ensemble's contrast results in a geometrically similar pattern of template center points in the feature space, but with all distances scaled (reduced in the case of a contrast reduction). The spread of the noise clouds around the centers is, however, unchanged by the scaling associated with the signal contrast change.

The distances between the clouds relative to the size of the clouds determine the proportion correct of the observer's decisions if an optimal decision rule is used at all times. In other words, the recognition accuracy is determined by the distances between the centers of these clouds normalized by the standard deviation of the noise (i.e. the size of the clouds). Therefore, if the standard deviation of the noise is increased by a factor of k, the signal contrast will have to be increased by the same proportion to retain the same performance. Thus at any given threshold, the signal contrast is linearly related to the standard deviation of the noise. Since the signal energy E is linearly proportional to the sum of the squares of the pixel contrasts of the signal, and since the noise spectral density N is linearly proportional to the square of the noise standard deviation, it follows that the E is linearly proportional to N. Thus, we have proven Claim 1.

Claim 2: If a sub-ideal observer randomly sub-samples m out of the M image pixels, but is otherwise identical to the ideal observer, then the sampling efficiency of the sub-ideal observer is m/M

According to the definition of sampling efficiency, we need to show that the slope of the E–N plot of the sub-ideal observer is M/m times that of the ideal observer. In our argument for Claim 1, we pointed out that the proportion correct is determined by the distances between the centers of the clouds normalized by the standard deviation of the noise. The distance between two centers A and B is given as

$$\| A - B \| = \sqrt{\sum_{i=1}^{M} (A_i - B_i)^2} \tag{B1}$$

where A_i and B_i are the contrast values at the ith pixel of templates A and B respectively.

When the sub-ideal observer randomly sub-samples m of M image pixels, the distance between the same two centers becomes

$$\| A - B \| = \sqrt{\sum_{i=1}^{m} (A_i - B_i)^2}. \tag{B2}$$

On average, this distance is reduced to $\sqrt{m/M}$ of the original distance because there are only m terms in the summation instead of the previous M. This shortening of the distance can be compensated for by scaling up the signal contrast by $\sqrt{M/m}$. Hence we have proven Claim 2. (Note that for this to be true, m should be larger than the dimensionality of the space spanned by the centers of the clouds so that the general configuration of the decision space is not altered.)

Claim 3: Total efficiency approximates sampling efficiency at high external noise levels

We just showed that if a sub-ideal observer sub-samples with a proportion $1/p$ of the input image (let $p = M/m$), then the slope of its E–N plot is p times that of the ideal observer, or that the sampling efficiency of the sub-ideal observer is $1/p$. In addition to having difficulty retaining all informative samples, an observer (such as a human observer) may also have difficulty in accurately encoding the samples. This inaccuracy can be modeled by a noise source internal to the sub-ideal. If we assume this noise (known as *equivalent noise*) to be Gaussian and additive to the external noise (see Pelli, 1981), then the E–N plot of the sub-ideal observer will be shifted to the left by the noise spectral density n of the internal noise source. Therefore, if q is the slope of the ideal's E–N plot, and thus $E\_ideal = qN$, then the relationship between the signal energy and the stimulus (external) noise spectral density of this generalized sub-ideal is

$$E_{\text{sub-ideal}} = pq\ (N + n). \tag{B3}$$

Recall that total efficiency is defined as

$$\textit{Total efficiency} = \frac{E_{\text{ideal}}}{E_{\text{sub-ideal}}} = \frac{qN}{pq\ (N + n)} = \frac{N}{p\ (N + n)}. \tag{B4}$$

When N is sufficiently large, total efficiency approaches $1/p$, which is the sampling efficiency. Thus, we have proven Claim 3.