FINAL REPORT

Models of Speed Discrimination

Prepared for
Life Sciences Division
NASA Ames Research Center

Technical Contact: Dr. Misha Pavel
Department of Computer Science and Engineering
Oregon Graduate Institute
20000 N.W. Walker Road
P.O. Box 91000
Portland, OR 97291-1000

(503) 690-1155
pavel@cse.ogi.edu

Contents

1 Overview

The prime purpose of this project was to investigate various theoretical issues concerning the integration of information across visual space. To date, most of the research efforts in the study of the visual system seem to have been focused in two almost non-overlaping directions. One research focus has been the low level perception as studied by psychophysics. The other focus has been the study of high level vision exemplified by the study of object perception.

Most of the effort in psychophysics has been devoted to the search for the fundamental "features" of perception. The general idea is that the most peripheral processes of the visual system decompose the input into features that are then used for classification and recognition. The experimental and theoretical focus has been on finding and describing these analyzers that decompose images into useful components. Various models are then compared to the physiological measurements performed on neurons in the sensory systems.

In the study of higher level perception, the work has been focused on the representation of objects and on the connections between various physical effects and object perception. In this category we find the perception of 3D from a variety of physical measurements including motion, shading and other physical phenomena.

With few exceptions, there seem to be very limited development of theories describing how the visual system might combine the output of the analyzers to form the representation of visual objects. Therefore, the processes underlying the integration of information over space represent critical aspects of vision system. The understanding of these processes will have implications on our expectations for the underlying physiological mechanisms, as well as for our models of the internal representation for visual percepts.

In this project, we explored several mechanisms related to spatial summation, attention, and eye movements. The project comprised three components:

1. Modeling visual search for the detection of speed deviation
2. Perception of moving objects
3. Exploring the role of eye movements in various visual tasks

In the first component of this research project we examined several quantitative models of integration of speed information over space. In the second component we examined several aspects of integration of motion information underlying perception of two-dimensional and three-dimensional objects in motion. In the final component, we examined how the visual system may use eye movements to convert complex spatial tasks into "simpler" sequences of subtasks. In this final report we outline our two approaches, summarize the results, and indicate possible future directions. Many partial results of this work have been used to guide concurrent work by Dr. Stone and Dr. Verghese.

2 Speed Discrimination

One way to examine how the visual system combines motion information from different locations is to measure the ability to discriminate speed as a function of the number, and possibly the size, of moving patches. The general idea underlying this work is that signal (speed) discriminability is limited by the intrinsic noise in the sensory mechanisms. The ability of the visual system to combine information over space should, therefore, result in improvements in discriminability. Our starting assumption is that we can model the sensory mechanisms as efficient statistical detectors. The modeling approach was based on signal detection theory, under the assumption that all internal noise is dominated by that at the local sensory level. Uncertainty then affects only the decision process and not the stimulus representation. The empirically assessed improvements in discriminability with increasing opportunities to sample the stimuli are then compared to those predicted by the models. Any deviations must be interpreted as either less than ideal efficiency in the sensory mechanisms or lack of independence.

2.1 Stimulus Representation

Any quantitative account of the experimental results in speed discrimination requires a formal representation of the stimuli. Our objective was to account for the data in the speed discrimination experiments carried out at NASA Ames Research Center. The procedure in these experiments was based on a two-interval-forced-choice paradigm (2IFC), and the observers were asked to identify the interval with the faster stimulus velocity. A typical stimulus was sinusoidal grating multiplied by a Gaussian window. The number of these local motion stimuli was one of the independent variables. Since the direction of motion was the same at all locations for both intervals, the simplest approach was to represent the speed sensed at each location.

We assumed that the observer codes the speed of the grating in a small neighborhood by a single random variable $S_{x,i}$ where x represents the location of the stimulus and $i \in \{1, 2\}$ identifies the interval. We also assumed that these random variables from different locations and different temporal intervals are mutually independent. At the outset we also assumed that an observer's *attention* to a particular location does not affect the stimulus representation. Later in this report we reconsidered this assumption.

The distribution of these random variables F_v depends on the actual

stimulus speeds v at the corresponding stimulus locations. For simplicity we assumed that the actual stimulus speed determines only the expected value of this distribution. In that case, stimulus speed will shift the expected value of the corresponding distributions by a function that depends on the physical speed.

In that case, the distribution of the representation is given by

$$F_v(s) = F[s - d(v)],$$

where the function d represents the internal speed scale. It is possible that empirical results from future experiments would require a more general model in which the internal noise depends on the sensed speed. In that case the model would have to be generalized to

$$F_v(s) = F\left[\frac{s - d(v)}{n(v)}\right].$$

In order to make numerical predictions for the shape of the psychometric functions, we needed to assume a particular shape for the underlying distributions of the sensory representation. A typical assumption made by most researchers is that the underlying distribution of S is normal, i.e.,

$$Pr\{S < s\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{s} e^{-\frac{1}{2}\frac{t-d(v)}{\sigma}^2} dt. \tag{1}$$

Although a normal distribution is frequently used in psychophysical models, there are other distributions that might be more convenient for the speed discrimination experiments. For example, if the model is based on detection of maxima, it is more appropriate to assume that the underlying distributions are Weibull or double exponential, depending on the range of the random variable. The double exponential distribution can be expressed in a closed form as

$$Pr\{S < s\} = exp\left[e^{-s}\right]. \tag{2}$$

The form of a double exponential distribution is illustrated in Figure 1. The double exponential distribution is similar in shape to the Gaussian distribution. In practice it is difficult to distinguish empirically between the Gaussian and the double exponential distributions. We used the double exponential here because it has a convenient property that is useful for the calculation of the distribution of a maximum of a collection of identically distributed random variables.

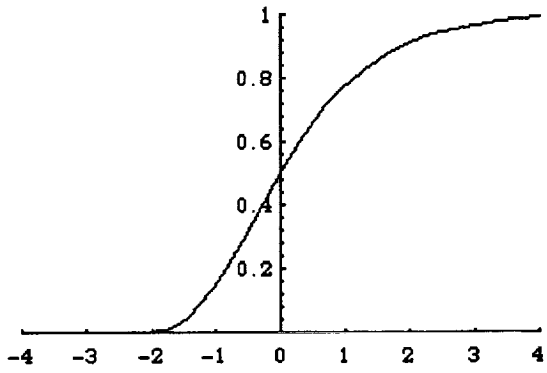

Figure 1: Double exponential distribution function shifted by $\log \log 2$ in order to be centered.

2.2 Optimal Observer

Given the stimulus representation by a collection of random variables, we assumed that the values of these random variables on each trial are used by the human observer to perform the speed discrimination task. The development of an optimal observer model requires assumptions about the amount of prior information that the observer can use to make the judgment. First, we consider an optimal observer with complete knowledge of the distributions and prior probabilities. An objective of an ideal observer is to minimize errors. Such an observer will make decision by computing the likelihood for each response alternative or hypothesis, and choose the one with the highest posterior odds. Thus, the optimal response r corresponds to the hypothesis H_r, where

$$r = argmax_\rho\{Pr\{Observation|H_\rho\}\}. \tag{3}$$

To compute predictions of the model, we must evaluate the likelihood of each of the two alternative hypotheses for different number of moving patches in each interval.

2.3 Single Stimulus in Each Interval

In a simple 2IFC task, an observer is confronted with two stimuli, presented in separate temporal intervals. His task was to determine the interval with the greater stimulus speed. The observer then entertains two hypotheses, H_1 and H_2, corresponding to the two possible locations of the faster stimulus.

We model this case by assuming that the sensory representation comprises two observations, $S_{x,1} = s_{x,1}$ and $S_{x,2} = s_{x,2}$, and the observer responds "1" if the likelihood

$$\lambda = \frac{Pr\{s_{x,1}, s_{x,2}|H_1\}}{Pr\{s_{x,1}, s_{x,2}|H_2\}} > 1. \quad (4)$$

If the probability of the faster stimulus in the first interval is p, then the performance of this model is given by the probability of an erroneous response

$$Pr\{Error\} = p\ Pr\{\lambda < 1|v_{x,1} > v_{x,2}\} + (1-p)Pr\{\lambda > 1|v_{x,1} < v_{x,2}\}. \quad (5)$$

If the faster stimulus is equally likely to be in both intervals, the probability of error is equal to the probability that the faster stimulus gives rise to a smaller sensory representation. Thus, the probability of error for the case when $v_{x,1} > v_{x,2}$ is given by the probability that $S_{x,1} < S_{x,2}$, or equivalently

$$Pr\{S_{x,1} - S_{x,2} < 0\}.$$

In order to predict a numerical value of the probability of error for an arbitrary values of v, i.e., a psychometric function, we must make assumptions about the underlying distributions. At this point we could assume either normal or double exponential distributions. In either case, the distribution of the difference can be easily calculated. For the normal distribution, the difference $(S_{x,1} - S_{x,2})$ is distributed normally, with the variance equal to the sum of the variances of the two random variables. In the case of a double exponential, the difference has a logistic distribution. In that case the resulting psychometric function is the upper curve shown in Figure 2.

2.4 Multiple Stimuli

For the case with multiple stimuli, the observer views several simultaneous patches of moving gratings presented in each interval. His task is to identify the interval in which at least one of the stimuli was faster than the remaining ones. With more than a single stimulus presented in a given interval, there are several different possible conditions. These conditions differ in the number of "faster" stimuli presented in the "target" interval. In this section, we consider the response process of an observer in different stimulus conditions.

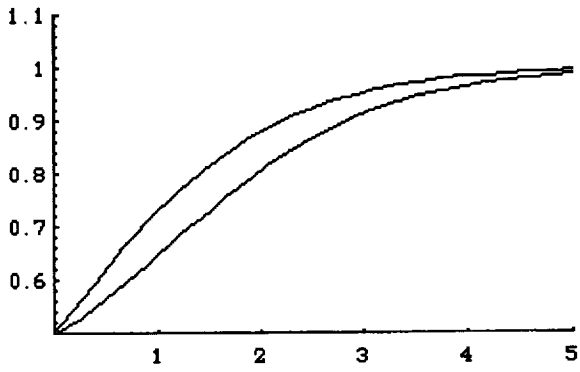

Figure 2: Resulting psychometric functions. The upper curve represents the results of 2IFC with one stimulus in each interval. The lower curve represents the psychometric function for 2IFC with two stimuli in each interval.

2.4.1 One-in-four Stimuli

First we consider the case where n stimuli are presented, but only one of them is potentially a target. First we consider a detector that would make the minimum possible number of errors given the stimulus representation described above.

Optimal Detector The actual speed of the stimuli is $v_{1,i}$ on the left, and $v_{1,i}$ on the right of the fixation point. The observer responds with "1" if

$$\lambda = \frac{Pr\{\vec{s}|H_1\}}{Pr\{\vec{s}|H_2\}} > 1, \tag{6}$$

where $\vec{s} = (s_{1,1}, s_{1,2}, s_{2,1}, s_{2,2})$ is the vector of all four observations. The response "1" is correct if the faster stimulus occurs in the left or in the right position. Thus, if the faster stimulus is equally likely in either interval and location then the probability of an observation given the hypothesis that the stimulus is in the first interval is:

$$Pr\{\vec{s}|H_1\} = f_d(s_{1,1})f_o(s_{1,2})f_o(s_{2,1})f_o(s_{2,2}) + f_o(s_{1,1})f_d(s_{1,2})f_o(s_{2,1})f_o(s_{2,2}), \tag{7}$$

where f_o and f_d are the distributions for the standard and the faster stimuli, respectively. The likelihood ratio for normally distributed sensory represen-

tation can be simplified by canceling identical terms to the following

$$\lambda = \frac{e^{-s_{1,1}d} + e^{-s_{1,2}d}}{e^{-s_{2,1}d} + e^{-s_{2,2}d}} > 1, \tag{8}$$

where d is the effect of the target's speed.

In a similar manner, for the double exponential distribution,

$$\lambda = \frac{exp\left((1-e^d)\left[e^{-s_{1,1}} + e^{-s_{1,2}}\right]\right)}{exp\left((1-e^d)\left[e^{-s_{2,1}} + e^{-s_{2,2}}\right]\right)} > 1. \tag{9}$$

In both cases, the likelihood ratio depends on the sum of exponential functions of the observations.

Maximum Detector Since the exponential functions are convex and rapidly increasing, a decision based on the sum these functions can be well approximated by the maximum value of the random variables observed in each interval. This approximation, frequently used in psychophysics was shown to predict performance indistinguishable from one generated by an optimal observer.

When the optimal observer is approximated by a maximum detector, the advantage of the double exponential becomes important. The maximum of a set of i.i.d. random variables with double exponential distributions is also distributed according to a double exponential. This fact enabled us to make predictions in closed form.

For the four-stimulus paradigm the observer is assumed to select the maximum observation in each interval and select that interval with the higher maximum. The resulting psychometric function shown in Figure 2 is derived from a logistic distribution,

$$Pr\{Correct\} = \frac{1}{1 + e^{-\log\frac{(1+e^d)}{2}}}. \tag{10}$$

These are predictions for an experiment in which the observers' task is to judge the interval containing the target with only one target present. The model can be tested by modifying the task and examining the resulting psychometric functions.

First, an experimenter may ask the observer to identify the faster stimulus. That is, in addition to indicating the interval, the observer must also

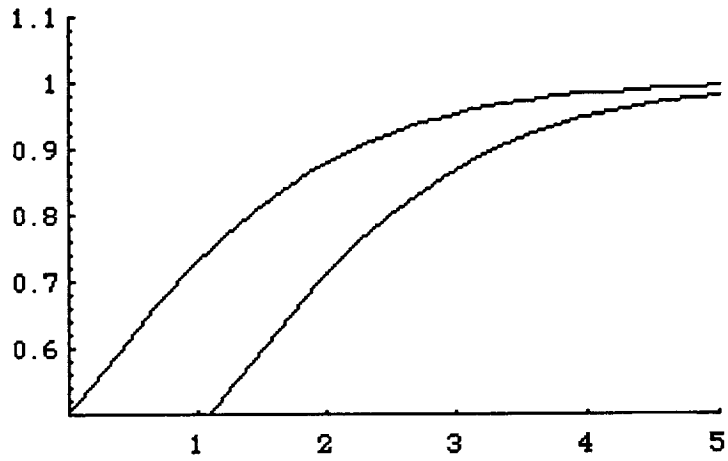

Figure 3: Psychometric function for identifying the fastest stimulus among four stimuli is the lower curve. The upper curve for a two-stimulus experiment is shown for reference.

identify the location of the faster stimulus. The results of this type of task are described by a logistic psychometric function arising from a single target and three distractors:

$$Pr\{Correct\} = \frac{1}{1 + e^{-d + \log 3}}, \tag{11}$$

shown in Figure 3

Another modification involves presenting the same speed for all stimuli in a given interval. Thus, during the interval with the faster stimulus, all patches would move with a faster speed. During the comparison interval, all stimulus patches would move with standard speed. The maximum detection model predicts that the number of patches should not affect the performance on this type of task. This is because the shift in the expected value for the maximum standard speed and the faster are identical. The optimal observer predicts slight improvements due to combining measurements in each interval.

An analog of this temporal alteration task in space is one in which the faster stimulus is presented in the same (or predictable) spatial location.

The optimal observer model examined in this section was assumed to have prior knowledge of the underlying sensory distributions, but did not use any prior knowledge of the stimulus presented on any particular trial.

Thus, large variations in the "standard" speed over trials would not affect its performance. It would be useful to determine whether human observers use prior knowledge about the approximate stimulus speed.

2.5 Effects of Attention

In our modeling effort so far, we have ignored any effects of attention. In the context of detection paradigms, attentional effects are usually interpreted to mean inhomogeneity of detection over space that is not explicable by the retinal factors. A typical example of attentional effects are experiments in which improvements in detection performance are obtained by providing observers with various amounts of prior knowledge about the stimulus parameters. In particular, the effects of spatial attention are demonstrated by improvements in performance due to a a reduction of observers' spatial uncertainty prior to stimulus presentation. These effects are often interpreted as being analogous to covert eye movements, and some researchers explain these attentional effects as resulting from an internal controllable beam of search light that increases contrast in the attended regions.

Because of its intangible or covert nature, attention has been difficult to model. One plausible assumption is that attention to a particular region of the visual field would reduce the internal noise, or alternatively, increase the signal to noise ratio for stimuli in that region. Such models are generally used to explain the effects of spatial uncertainty. In these models, attention is assumed to be under a complete control of the observer and set up prior to the presentation of the test scene.

This aspect of the experimental paradigm used to study attention is a subject of severe criticism. In particular, the experimental paradigm used in the study of attention rarely occurs in natural environments. We speculate that it is very unlikely for a visual scene to change instantaneously while the organism has a complete knowledge of the direction of his gaze. We would, therefore, argue that in natural environments, all attentional control is made relative to the visual scene and the stimulus. Observers confronted by the impoverished experimental paradigms are not likely to have complete control of their attention. Thus, even with a complete knowledge of the location of the stimuli, the observer may sample the scene less than optimally. This suboptimal behavior that could be modeled as additional noise, and may be responsible for the limitations of the models.

2.6 Recommendations

1. The psychometric functions for two-interval-forced-choice experiments would be better approximated by a logistic function or a mixture of two logistic functions rather than by Weibull distributions.

2. The general empirical approach that would aid in modeling would consist of finding the best fit psychometric function (underlying distribution) for single stimulus in each interval at a predictable location. Predictions for other conditions should be made using this psychometric function in conjunctions with various models of discrimination, such as maximum detector or summation.

3. Comparison of the effects of spatial and temporal uncertainty would provide the information for building models of spatial interactions.

4. Most models for the 2IFC procedure are based on the idea of a comparison of the representations between the first and second interval. The results predicted from these models should be insensitive to the variation of the "standard" speed. This notion that is critical to the modeling effort, should be tested by increasing the trial-to-trial variability of the "standard" speed.

5. Any deviations from the predicted integration effects could be expressed in terms of unobservable uncertainty. For example, if the addition of stimuli within each interval provides less than expected gains in the discriminability, the process of combination has statistical efficiency that is less than 100%. To account for the loss of efficiency we should investigate the possibility that human have only stochastic control over attention that is mostly driven by the visual scene.

3 Object Motion

In the project described in the first section we have examined the psychophysical aspects of integrating near-threshold motion information over space. In the projects described in this section, we assumed that the low level analyzers provide estimates of local velocities, and we addressed the issue of integration of information necessary to mediate the perception of two-dimensional and three-dimensional objects.

3.1 Aperture Problem

The motion of an image at a given point can be sensed only along the direction of the intensity gradient in the image. This technical constraint implies that the motion of a long line whose end-points are outside of the field of view can be measured only in the direction perpendicular to the direction along the line.

Prior experiments suggest that the perception of motion direction by the visual system depends on the motion of any visible endpoints of the line, whose motion is not ambiguous. The fact that this is true even if the endpoints result from an occluding aperture is best demonstrated using one or more lines moving diagonally within a rectangular aperture. Even though the motion of each line can be measured only in the perpendicular direction, the lines appear to move parallel to the longer side of the rectangle. This percept, called barber pole illusion, suggests that the visual system combines information along the extent of the line including the end points.

In general, the visual system seems to solve the problem of ambiguous motion of a line segment, called the *aperture problem*, by combining motion information across space. The integration-based approach is possible only if the integrated points on the moving object have intensity gradient pointing in different directions. Each integrated or fused point constrains the direction of the object motion along the direction of the local gradient, and the resulting *perceived* motion is then in the direction consistent with all these constraints.

This model for integration of information is based on the assumption that the visual system can determine which points belong to the same rigid object. Points, line segments, and other higher-dimensional features must follow a rigid motion in order for their combination to disambiguate the local velocity measurements. If a collection of local velocity measurements can be identified to belong to a single object, it is appropriate to combine the

motion information. Thus, the fundamental question is which point velocity measurements should be fused?

The selection of points whose motion is to be combined can be based on various aspects of the image, including similarity of color, spatial proximity or belonging to the same higher order *feature* in the image, or even having similarities of motion. Perhaps the least controversial of these aspects to be used for the fusion decision is for a point to be a part of the same image feature, such an edge or a contour.

Loosely speaking, two adjacent points form a contour if the maximum intensity gradient at each of these points is approximately perpendicular to the line connecting the points. The direction of the contour at a point is perpendicular to direction of the gradient. This definition of a contour does not exclude points at which the contour sharply changes its direction.

In many natural images, a contour that belongs to a single object may be interrupted by occlusions. Motion at the occlusion boundaries is affected by the boundaries, and the resulting motion of these points may conflict with the motion of the exposed object. In this situation, the visual system should reduce, or even ignore, the contribution of the boundary points to the overall motion determination. Experiments with stimuli that give rise to such conflicts can potentially provide information about the decision processes that underlie the integration.

We have previously shown that for simple motion,n such as translation, and for simple contours such as straight lines, the visual system is capable of integrating information from disjoint regions of the visual field. For example, a diamond translating back and forth along the horizontal axis was viewed through two disjoint apertures. The apertures were arranged so that the corners of the diamond were occluded, and each aperture alone would favor a different direction of motion. For a large range of velocities, observers could integrate the motion into a single percept of a translating diamond. The percept of a single object was unambiguous when the occluding edges were visible but persisted even when the occluding edges were only implied.

For a rotational motion, the visual system has more difficulty integrating motion information from disjoint regions. In fact when a rotating square is viewed through several apertures the percepts is nonrigid whenever the corners of the square are invisible. In the limit, this non-rigidity percept can be obtained by minimally occluding the apex of an rotating corner. Thus, it appears that the visual system is especially relying on the region of high curvature. These regions correspond to the elementary two-dimensional (2D)

features — corners. We refer to corners as 2D features because any 2D rigid transformation of a corner is completely specified by the final positions and orientations of these features. This work has been carried out with M. Shiffrar at Rutgers University. The results of this investigation were reported at the annual meeting of the Association for Research in Vision and Ophthalmology [1995].

3.2 Three-Dimensional Motion

Our discussion thus far was limited to 2D motion. The problem is considerably more complex when we consider motion in 3D. The increase in complexity is due to the fact that the image of a rigidly moving 3D object is, in general, not 2D rigid. Thus it is much more difficult to decide which points belong to a given moving object. The simplest possible stimulus to address this issue is a set of smoothly moving points sampled from the surface of a rigid 3D object. The situation where a set of moving points give rise to a 3D percept is called the kinetic depth effect (KDE).

One way to interpret the KDE is that the visual system assumes that the points define a rigid object and then use the 2D retinal image to determine the object's 3D coordinates. A complete 3D representation — Euclidean representation — exactly specifies all interpoint distances.

There is some prior empirical evidence suggesting that the visual system does not necessarily compute the complete 3D Euclidian representation of the object. In fact, the human visual system may be less sensitive to the distortions in depth than those in the frontal plane.

To investigate the complexity of the representation and the mode of information integration, we used stimuli consisting of moving points, similar to the standard KDE stimulus. In addition to rigid 3D motion we examined whether the observers can discriminate an affine type of motion from a rigid type. The main result is that the affine motion was much less discriminable from the rigid motion than the nonrigid motion with an equivalent amount of distortion. The amount of distortion was measured in terms of total energy that would be required to distort an elastic 3D object.

Our results suggest that the visual system uses more general approach than a complete 3D Euclidian representation. We speculate that the process of integration is accomplished by applying a test for a consistency with the internal representation while it is computing it. This work has been carried out in collaboration with D. Weinshall of Hebrew University and New York

University. The results of this investigation were reported at the annual meeting of the Association for Research in Vision and Ophthalmology [1997].

4 Eye Movements

In the first and second section we examined aspects of integration of information over space within a single view, i.e., without any eye movements. The signal detection model described can provide at best only a partial answer, because in most situations, people can scan images with their eyes and possibly combine information obtained from different gaze directions. In some tasks eye movements may not help very much. For example, detection of flicker my not benefit from eye movement. In this section we report initial analysis of the relationship between the complexity of a task and the utility of eye movements. Most of the text in this section was published as a chapter in a book on exploratory eye movements [1995].

A flexible, mobile sensor appears to be an essential component of most biological vision systems. In the human visual system, mobility is achieved by head and eye movements. The degree of importance of a manipulable sensor to a vision system is a critical question both for students of biological vision systems and for designers of artificial vision systems. In humans, eye movements appear to mediate a variety of functions ranging from image stabilization to visual search.

In this project, we examined the notion that eye movements mediate a tradeoff between various information processing demands on the visual system. In its most basic form, our hypothesis is that eye movements permit the visual system to convert parallel solutions of certain tasks that would require large amounts of hardware (or "wetware") into sequential algorithms that require considerably less complex signal processing mechanisms, although sacrificing processing speed. We will argue that the usefulness of sequential algorithms will increase with the difficulty or perceptual complexity of visual tasks.

In first part of this section we discuss several examples of perceptual tasks and consider the potential impact of eye movements on human performance. We note that one of the important performance characteristics affecting the impact of eye movements is the independence of performance from stimulus location, i.e., translation invariance. We will note that eye movements are generally more useful for more difficult tasks in which the human visual system is less translation invariant.

The key question in our analysis is how to determine the complexity of any particular visual task. To answer this question, we explore the potential of a formal analysis of algorithmic complexity to estimate the *perceptual*

complexity of visual tasks. Although algorithmic complexity theory is most relevant for the analysis of computer algorithms for unbounded problems, we suggest that it can also be used for finite problems confronting the visual system. We accomplish this by adding constraints on the computing mechanism. In particular, we examine the complexity of parallel computational networks whose depth is constrained to a small number of levels. Our analyses suggest that the perceptual complexity of a task generally correlates with the difficulty of the problem as measured by human performance on that task.

The final section of this section (Section 4.3) presents a detailed discussion of translation invariance. As we noted above, translation invariance is a key property of the visual system that may determine the need for eye movements.

4.1 Visual Task Performance

In this section we consider several examples of visual tasks and their perceptual complexity. We also discuss the role that eye movements might have in facilitating performance. In particular, we anticipate that the effects of eye movements on performance will depend on the ability of the visual system to perform various visual tasks equally well at different locations of the field of view. Finding performance to be independent of stimulus location would reflect translation invariance of the visual system with respect to those tasks. The impact of eye movements on task performance, in turn, is probably quite limited for those tasks that can be performed in a translation invariant manner.

4.1.1 Detection

Perhaps the most straight-forward visual task is the detection of a luminous target (e.g., luminous disk) on a dark background, as shown in Figure 4a. In a detection experiment that consists of a sequence of trials, an observer is asked to fixate on a fixation point at the center of the display — indicated by the central cross in Figure 4a. On some trials, the target is displayed for a brief period of time at a location within the display area. On the remaining trials, the target is not displayed at all. On each , the observer is asked to indicate whether or not a target was present.

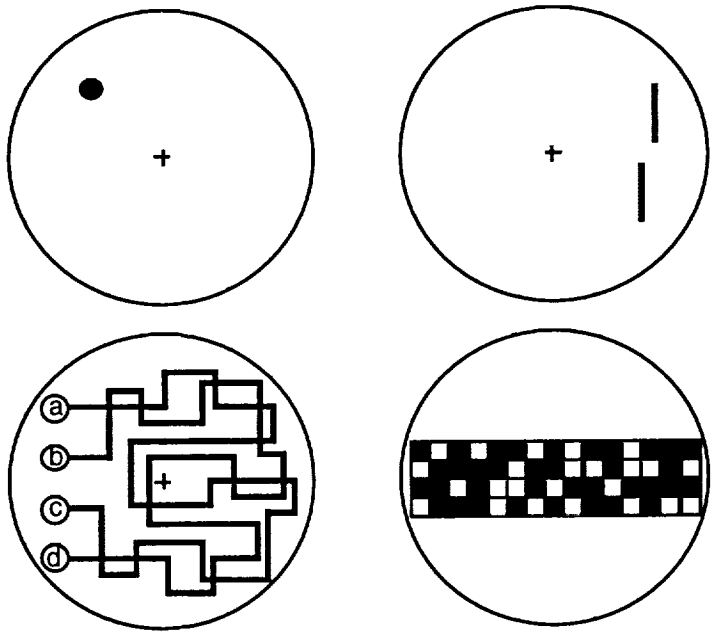

Figure 4: Examples of various visual tasks. The cross represents the fixation point. (a) Luminous object detected on a dark background. (b) Vernier acuity task. (c)Continuity puzzle. (d) Parity problem.

The ability of an observer to detect the target is generally found to depend on the contrast between the background and the target, target luminance, size of the target, and the distance of the target from the fixation point — eccentricity. For luminous, clearly visible targets, the task appears to be very easy and the observers' responses indicate nearly perfect performance. In that situation, the target eccentricity is not a critical variable; the visual system is fairly translation invariant, and we would not expect eye movements to improve performance.

A very similar task involves detecting a bright disk among dark disks (distractors) or a red disk among green ones. If the difference between the distractors and the target is such that the target is easily detected, then the performance is essentially independent of the target eccentricity and even the number of distractors (e.g., Triesman and Gelade, 1980) [1980]. For these tasks that appear to be fairly translation invariant, eye movements will provide little additional improvement in performance. Observers can perform these tasks without any noticeable effort or focused attention. Because of that, such tasks are sometimes referred to as preattentive (e.g., Bergen and Julesz, 1983) [1983].

In addition to the detection of a conspicuous target, the human visual system can also easily perform other tasks in the periphery such as detection of rapid fluctuations in luminance over time (i.e. flicker), motion at various eccentricities, and other. Eye movements will provide little help in performing these tasks.

4.1.2 Masking

Adding noise to mask the target information is a useful method employed by psychophysicists to measure the statistical efficiency of sensory mechanisms. This experimental method is frequently referred to as a "masking" paradigm. By manipulating the amount of noise added and measuring the decrease in performance, it is possible to measure certain internal limitations of the sensory mechanisms. For example, it is possible to evaluate the amount of internal noise that limits detection (e.g., Pelli 1983) [1983].

If the target luminance contrast is decreased or the visual field is contaminated by noise, observers' performance will decrease in most target detection paradigms. The decrease in performance will, in general, be more pronounced for targets located further from the fixation point. This dependence on location suggests that the visual system does not obey strict

translation invariance. Under these circumstances, eye movements toward the target are likely to improve performance.

In addition to the lack of translation invariance, there is some evidence that multiple looks can improve the detectability of a masked target (Rovamo and Virsu, 1979; Levi, Klein and Aitsebaomo, 1985) [1979, 1985]. Under these circumstances, eye movements are likely to facilitate improvements in performance.

It is worthwhile to note that in most detection situations, the amount of signal (contrast increment) that is just detectable is proportional to the amount of noise (e.g., standard deviation of the signal). A similar phenomenon holds for tasks in which observers are asked to detect increments or decrements in contrast. This type of scale invariance in psychophysics is called Weber's Law.

4.1.3 Localization

Another important visual task is the localization of objects in the visual field. Whereas absolute localization is relatively poor, the human visual system is capable of making accurate relative location judgments, such as length discrimination, e.g, Burbeck and Hadden [1993]. One way to summarize the empirical results is that the uncertainty in judgments obeys Weber's Law.

An interesting version of the relative location judgment is a task called vernier acuity, illustrated in Figure 4b. Subjects in the vernier acuity task are asked to identify whether the bottom bar is to the left or to the right of the top one. When the vernier stimulus is presented in the fovea, observers can make these judgments extremely accurately. For example, they can discriminate an offset in location down to 6 seconds of arc. As we shall discuss in depth later, the performance on the vernier task deteriorates quickly as the stimulus is moved from the central vision to the periphery.

The vernier acuity task requires the visual system to perform a more sophisticated task than simple detection. First, it requires the detection of both bars using luminance contrast. Second, it depends on the ability of the visual system to compare locations of two spatially separated objects.

4.1.4 Multidimensional Tasks

Spatial relations need not be limited to the relations between two points. One way to increase the requirements on spatial processing is to ask observers to

judge spatial contiguity. For example, consider the task of identifying which points are connected in Figure 4c. There are four starting points on the left of the display and they are pairwise connected. Observers are asked to identify the connected pairs. Although we are not aware of extensive experimental data, our limited observations suggest that eye movements are very useful if not essential in for this task.

Each straight line segment in the spatial contiguity task can be interpreted as a dimension of the task. There are ways of increasing the dimensionality of a task which do not depend on spatial location. For example, in the "parity" task depicted in Figure 4d. the observer is asked to judge whether the number of dark squares is odd or even. In this task, the position of the squares is irrelevant. The parity task is similar to a counting task. The performance on this task depends on the number of items and the area. For a limited number of items distributed over 2 degrees in fovea, eye movements are not very helpful Kowler and Steinman [1977]. As the area and the number of items increase, eye movements appear to be more useful.

In these two examples of multidimensional spatial tasks, the eye position might serve as a pointer. One possibility is that the motion of the pointed mediates conversion of a parallel task into a sequential one which is "easier" for humans.

4.1.5 Speed-Accuracy Tradeoff

Although the task difficulty is a central notion of this section, we have not yet defined the relationship between human performance and task difficulty. In fact, an experimenter's choice of a particular empirical performance measure can have critical implications for the assessment of task difficulty.

In psychophysical experiments, the observers' performance is typically characterized by two measures:

1. Accuracy — How accurate are the observers' responses.

2. Speed — How fast can a task be accomplished

In most experimental work, researchers have typically focused on one or the other measure. This emphasis might have undesirable consequences because, for most tasks, there is typically a significant tradeoff between speed and accuracy. In particular, faster responses are generally less accurate, and slower responses are typically more accurate. A comprehensive discussion

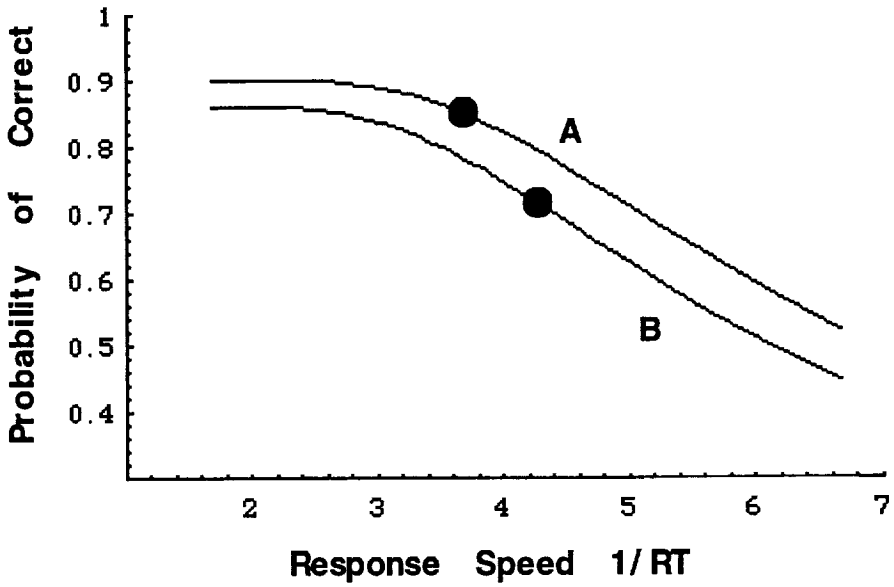

Figure 5: A tradeoff between speed and accuracy for two different tasks, A, and B.

of how these effects are critical for the interpretation of empirical results is beyond the scope of this section, but the interested reader can refer to an authoritative work of Sperling and Dosher [1986].

One way to represent the effects of speed-accuracy tradeoff for two different tasks is shown in Figure 5. Each curve in Figure 5 — an operating characteristic curve — corresponds to human performance on one task. An operating characteristic curve is obtained by repeating the same experiment, but instructing observers to put different emphasis on accuracy or on reaction time. As the observers change their strategies from focusing on accuracy to decreasing their reaction times, they trace out a curve.

Note, that if an experimenter would perform only a single experiment for each task, he could obtain results indicated by the two black dots. An important implication of this example is that a definition of task difficulty on the basis of reaction time alone, would lead to a conclusion that task A is more difficult than task B. An examination of the operating curves leads to the opposite conclusion. In particular, for any fixed probability of correct responses, task B takes longer to complete than does task A. Thus, task B is actually uniformly more difficult that task A.

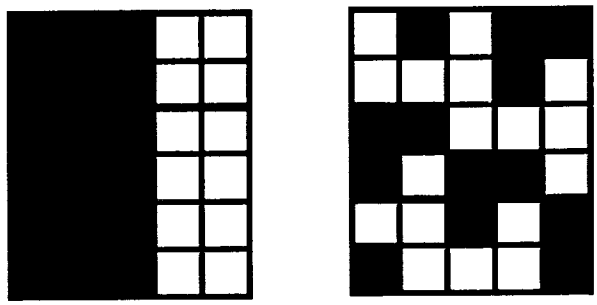

Figure 6: Example of a "simple" and a "complex" image.

4.2 Task Complexity

In this section we consider the notion that the different tasks described in the previous section can be characterized by an abstract measure of difficulty. The measure that we focus on is the notion of task complexity. We will then argue that many tasks that are too complex to benefit from using eye movements as a way of converting complex tasks into sequences of simpler ones.

4.2.1 Theory of Complexity

One appealing approach to describing task difficulty is the mathematical notion of complexity. A comprehensive introduction to complexity theory is well beyond the scope of this section, but we will present certain fundamental concepts that will be useful in our later discussions of the complexity of visual tasks. A clear presentation of some of these concepts can be found in Cover and Thomas [1991].

Loosely speaking, the complexity of an object is the shortest binary string (or program) that completely specifies or reproduces the object. We introduce the notions of complexity in the following examples. Consider, two binary $n \times n$ images shown in Figure 6.

The left image can be described by specifying the rectangle comprised of dark squares. This description would require on the order of $2 \log n$ bits, i.e., two integers describing the corners of the rectangle. In contrast, the image on the right would probably require nearly n bits to specify. According to this analysis, the left image is more complex than the right one. More formally, the complexity K of a string x is the minimum length program p

that generates x using computer $\mathcal{U}$.

$$K_{\mathcal{U}}(x) = \min_{p:\, \mathcal{U}(p)=x} l(p)$$

The analysis of the example in Figure 6 was based on an assumption that the second image is one of 2^{n^2} possible images. As it turns out the right image can be described as a repeated sequence $y = \{0101100\}$ that was written horizontally and wrapped around. The sequence y can be thought of a program that was used to generate the image. The length of the program required to generate the entire image is the length of y, i.e., $l(x) = 5 + m$ where m is the length of a small program specifying how y is used to generate the image. This type of efficient representation, however, can be used for a relatively small proportion of possible images. If all 2^{n^2} images are equally likely there is no savings in the length of representation.

This example illustrates that the length of the description depends on the computer, data representation, and possible data. If $\mathcal{U}$ and $\mathcal{A}$ are two computers sufficiently powerful (i.e., universal) computers, then the complexity of x can differ by a constant independent of the length of x,

$$K_{\mathcal{A}} \leq K_{\mathcal{U}}(x) + c.$$

The constant c represents the length of a set of instructions that programs computer $\mathcal{A}$ to behave as computer $\mathcal{U}$.

Within this framework, the complexity of string x of n elements can be written as

$$K(x) = c + l(x),$$

where $l(x)$ the length of the description of x, and c is the length of a program needed to convert one computer to another.

One of the aims of the theory of complexity proposed by Kolmogorov [1965], Solomon [1964], and Chaitin [1966] was to investigate the algorithmic complexity of very large objects, $n \to \infty$. In that case the finite constant c can be ignored. For our purpose, however, the size of the constant c is likely to be significant because it represents constraints arising from the specific mechanisms underlying the human visual system.

We will, therefore, examine some specific mechanisms and the effect of the constraints on the complexity of the visual problems discussed in Section 4.1.

Before we proceed to discuss the specific architectures, we would like to note that there is a close connection between complexity theory and information theory. It should be apparent from the examples in Figure 6 that

the description of an object depends on the number of other objects that must be distinguished. For example, the number of bits required to describe an integer in a computer depends on the largest possible integer N for the particular machine, and is equal to $l(x) = \log N$. This number of bits is required to distinguish N different integers.

A very closely related notion is the entropy of random variables in information theory. The entropy of a random variable X that takes on values from a set $\mathcal{X}$ is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log(x)$$

If each value of an integer X is equally likely, then the entropy of X is equal to $\frac{1}{N} log N$. In general, the entropy of a sequence of N random variables is approximately equal to the expected value of Kolmogorov complexity divided by N (Cover and Thomas [1991]), i.e.,

$$\frac{1}{N} E\{X_1, X_2, ..., X_N | N\} \approx H(X).$$

This fact can be useful in relating the complexity of visual tasks to the accuracy in human performance.

4.2.2 Capacity of Constrained Parallel Machines

As we noted above, for finite problems, the length of the computer-specific program c might be an important component of the overall complexity of a problem. Thus, for most of visual tasks, the visual system architecture and stimulus representation system will significantly affect the task complexity (difficulty). This effect of architecture on problem complexity is commonly used to infer aspects of the structure of the visual system by measuring the difficulty of different visual problems.

Our analysis here is based on an alternate approach. We start by assuming a two-layer parallel structure for the system architecture and examine its implications on the complexity of visual tasks. Because in many laboratory visual tasks, the input image and observers' responses are binary, we first assume an architecture of a Boolean machine based on Boolean algebra or logic rules (i.e., conjunctions, disjunctions and negations). Subsequently, we consider a more general extension of the Boolean machine based on neural networks.

Any Boolean function can be represented in Disjunctive Normal Form (DNF). DNF consists of disjunctions (*OR*) of binary variables combined by conjunctions (*AND*). An example of a DNF representation for three variable has the form:

$$y(x_1,\ x_2,\ x_3)\ =\ (x_1 \cap x_2 \cap x_3) \cup (\bar{x_1} \cap x_2 \cap x_3) \cup ...,$$

where x_i are Boolean variables, the bar represents negation, and $\cap$ and $\cup$ represent conjunction and disjunction, respectively. For example, an exclusive OR (*XOR*) can be written as

$$y(x_1,\ x_2)\ =\ (\bar{x_1} \cap x_2)\ \cup (x_1 \cap \bar{x_2}). \tag{12}$$

The building elements (basis functions) for boolean networks are single-valued boolean functions which can be in turn implemented by switching circuits consisting of *AND* and *OR* gates, and negations. DNF is then represented by two layers of gates — the first layer consists of *AND* gates and the second layer is a single *OR* gate. The complexity of these machines (i.e., the length of a "program") is defined by the number of gates and number of connections.

A DNF Boolean machine with an unlimited number of gates and connections represents a parallel machine that can compute all boolean functions. Although DNF is universal, there are several reasons why it is not the most desirable way to implement computations in practice. First, DNF is typically the most complex way of representing a function. There are many functions that could be computed by combinations of considerably fewer gates than prescribed by the complexity of DNF. Second, a DNF representation must in general be extended to more than two layers in order to accommodate constraints on fan-in, fan-out and connectivity.

Because switching circuits are fundamental components of digital computers, much effort has been devoted to evaluate complexity and finding ways to simplify implementations of Boolean functions, see for example Wegener [1987]. Many techniques have been developed to take advantage of particular properties of Boolean functions to be minimized. The results of these efforts suggest that major simplifications are achievable for functions that have certain properties, such as symmetry or monotonicity and those functions that are not completely specified, i.e., those that include many *Don't Cares*.

We can turn to the problem of estimating the complexity of the visual tasks illustrated in Figure 4 if we assume that all pixels are binary — black

or white. We also assume that the number of layers for processing is limited to two in order to minimize processing time.

The simple detection would require a single *OR* combination of all binary pixels — that computation could be accomplished with complexity on the order of $O(n)$. The vernier acuity task shown in Figure 4b requires, in general, a comparison of all pairs of pixels; thus, the complexity of the vernier task is on the order of $O(n^2)$. Both of these computations can be accomplished within two or three layers of gates. In contrast, the complexity for a parallel computation of connectivity and parity require exponential complexity for a two-layer circuit. Therefore, this type of computation would be limited to a relatively small field of view, and a sequential type of algorithms may be more desirable. Before we consider a sequential approach, we must ascertain that these conclusions were not the result of limitations due to a binary representation and computation. We will, therefore, examine a more general approach based on adaptive (neural) networks.

In order to extend the above results to continuous inputs, we consider a class of machines based on a network of units that compute a linear combination of the inputs. The basis functions computed by the units are monotone nonlinear transformations of weighted linear combinations of their inputs. We call these networks adaptive because the parameters of the network can be adjusted in response to the performance of the network on a given task.

If the nonlinear transformation is a step function and the inputs are binary the adaptive networks reduce to the Boolean machines described above. Thus, any Boolean function can be computed by a machine composed of two layers of these units.

The universality of the adaptive networks has been first put forth by Kolmogorov [1957], who demonstrated that with an appropriate choice of the nonlinear transformations, it is possible to approximate any continuous function with arbitrary accuracy using three layers of units [1].

For the purpose of analysis of the visual tasks, the adaptive networks must behave as classifiers. This is a natural function of the adaptive networks. In fact, for continuous inputs, the linear sum followed by a step-function (threshold) behaves as a classifier. In particular, these linear threshold units (LTU) compute linear discriminant functions. A two-dimensional example

[1]There are other bases (units) to model computation. For example, another useful basis involves radial basis functions. The overall complexity arguments will generalize to these bases.

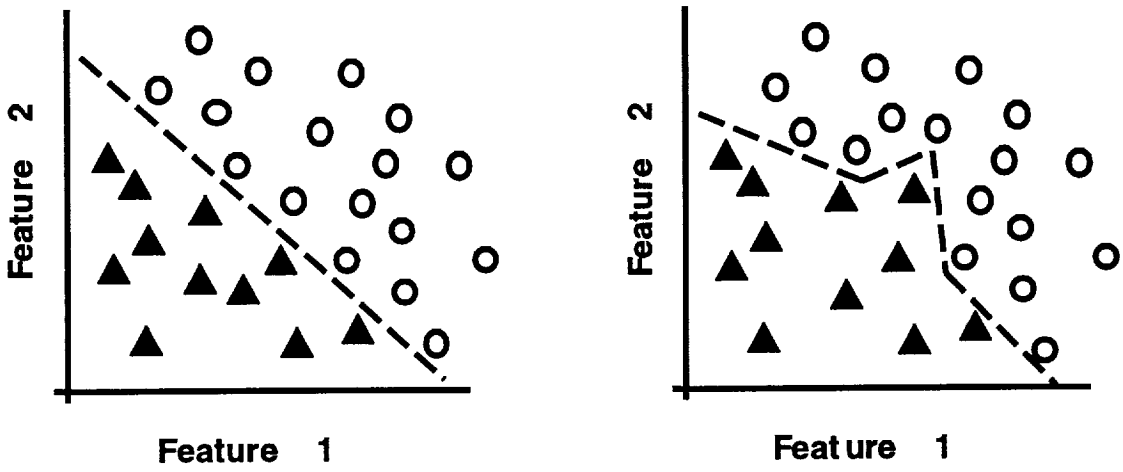

Figure 7: Classification in a two dimensional feature space. A linearly separable problem is shown in (a) and one that can't be solved by a linear combination of features is shown in (b).

is shown in Figure 7. Each input dimension represents the amount of the corresponding feature present at the input. Objects to be classified are represented as points in the feature space. The classification is represented by a surface or surfaces that separate the feature space into subsets corresponding to object-categories.

The separating surfaces are called decision surfaces and are generated by discriminant functions characterizing these surfaces. Figure 7 illustrates two simple two-dimensional feature spaces with two classes of objects to be distinguished. For example, a linear threshold unit can classify its input space into two regions separated by a hyperplane.

The complexity of a task is again specified by the number of operations which correspond, in the case of adaptive networks, to the number of units and number of connections. In the case of a classifier, there is an alternative way to specify complexity — in terms of the complexity of the discriminant surfaces required to perform the categorization task.

For example, the simplest class of problems that can be solved using a single linear threshold unit is called linearly separable. A useful analysis of various visual tasks, was published by Minsky [1969]. They classified tasks in terms of the highest order of a predicate [2] that is required to compute

[2]The order of predicate in logic is in terms of n-ary predicates

a correct response. In a simplified interpretation of their results, the order of a predicate represents the number of pixels that need to be considered simultaneously in order to perform the task.

The order of predicate is often directly related to the complexity of the machine. For example, according to Minsky and Papert's 1969 analysis, linearly separable problems correspond to first order predicates. That means that each pixel's contribution to a decision is independent of other pixels and the decision can be performed by summing of the contribution of each pixel. Thus, linearly separable tasks can be accomplished by a single layer LTU and their complexity is typically on the order of $O(n)$.

We can now turn back to the analysis of the visual tasks presented in Section 4.1. The detection of the presence of a luminous point on a dark background is a first order predicate, and can be accomplished by a single sum over the image. The sum will be independent of the location of the target location, and, therefore, this algorithm is translation invariant.

A different situation arises for the complexity of masking tasks. In a typical masking task, noise is added to each pixel and we cannot assume that the noise variance and the signal values are known exactly. An optimal way to perform a target detection is based on the comparison of each pixel to a number of its nearest neighbors. To perform this detection in a translation invariant manner increases the complexity of the detection task by a factor determined by the size of the neighborhood used for the comparison.

For visual tasks that require assessment of a length or distance, at least two pixels are required to compute it. Thus, vernier acuity, for example, is a second order predicate. Finally, to determine continuity of a contour or a parity of points requires simultaneous computation on all of the pixels or "features" in an image. For an adaptive network limited to two or three layers, the number of computations is exponential in the number of pixels. Thus, these results are consistent with those obtained for the Boolean machines. When limited to a few layers of computation units, there are problems like connectivity or parity that are exponentially difficult. We will now discuss the effect of allowing a large number of layers, or more simply, sequential algorithms.

4.2.3 Sequential Machines

With only two layers, visual tasks such as continuity or parity are exponentially complex. For example, a machine that determines parity and is

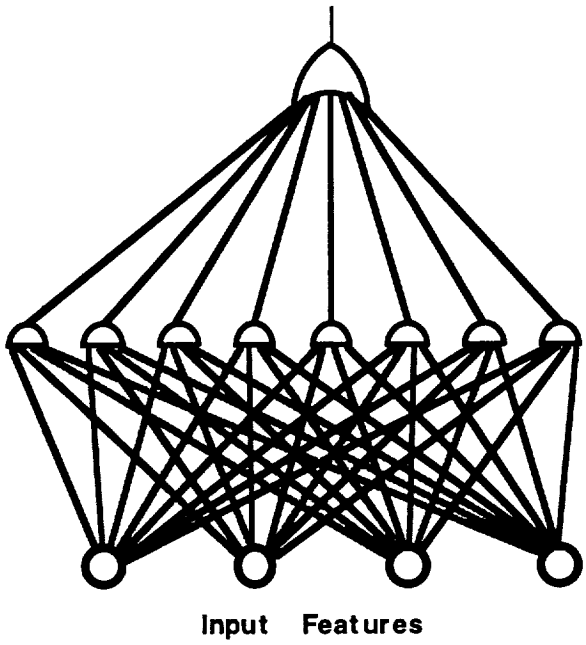

Figure 8: Disjunctive Normal Form of a combinatorial circuit to determine the parity of the input. The light connections represent negated inputs, the first layer are *AND* gates and the second layer are *OR* gates.

restricted to two layers (e.g., disjunctive normal form or DNF) would require exponentially many gates and connections, as illustrated in Figure 8.

Because the particular properties of the parity function, it is possible to find more efficient combinatorial circuits by permitting more layers of gates. For example, the parity function can be computed with the order of $O(5(n-1))$ gates arranged in $n-1$ layers. The resulting circuit is equivalent to a sequential application of a very simple function such that each level of gates corresponds to a time step for that sequential machine. The only complication of a sequential machine is that it must have a memory for the output of its prior output or state.

In particular, the parity $y(n)$ after seeing n inputs can be expressed as a function of the parity after seeing only $n-1$ inputs and the current input x_n,

$$y(n) = \bar{y}(n-1) \cap x_n \cup y(n-1) \cap \bar{x_n}. \tag{13}$$

In this simple case the state (memory) is the output y. A corresponding two-

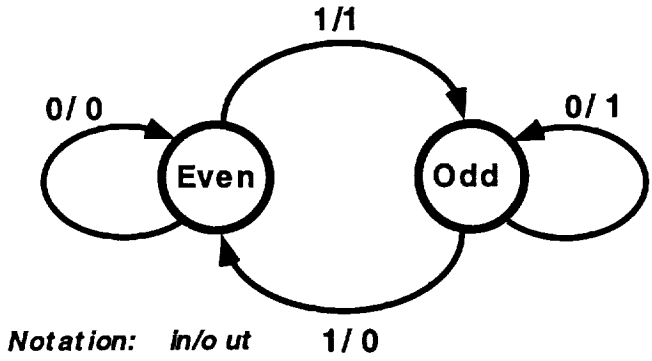

Figure 9: Sequential machine for the computation of parity for arbitrarily large inputs in the form of a simple logic circuit implementing the state transition diagram.

state sequential machine, illustrated in Figure 9, could determine the parity of an arbitrarily large input in time proportional to the size of the input.

The sequential machine in this example was much simpler than even the simplest corresponding parallel machine. But even more important is the fact that the sequential machine could perform the task, i.e. to determine parity, for input of any size without prior knowledge of the size.

In general for some of the computationally "difficult" tasks for parallel machines with limited number of layers, it is possible to construct sequential machines that can perform the task in polynomial time. We hypothesize that when the complexity of a visual task exceeds the capacity of the available parallel mechanisms, the visual system attempts to convert the parallel task into a sequential one. Although this conversion can be performed by eye movements, it is possible that this tradeoff is also mediated by covert attentional processes. In either case, the sequential approach which is typically more flexible and requires a less complex mechanism, takes generally longer than a parallel one.

4.2.4 Theoretical Speed-Accuracy Tradeoff

The discussion thus far has been based on the assumption that the results of the computations are always correct. Under this assumption, the complexity of a task is given by the minimum number of steps required to arrive at the correct answer. For sequential algorithms the number of required computa-

tions, and the thus the time required to complete a task, can be used as a direct measure of task complexity.

We noted in Section 4.1 that an assessment of human task difficulty must include both accuracy and reaction time because observers have a choice of strategies trading off speed and accuracy. The same type of tradeoff confronts an active sensor.

To interpret accuracy data within the complexity framework we must relate the probability of errors to the task complexity. One possible way to bridge the gap between accuracy and complexity is based on rate distortion theory (see for example Cover and Thomas [1991]) used in information coding and communication.

Rate distortion theory is applicable in situations where an errorless code is either impossible or impractical. For example, the representation of arbitrary real numbers in digital computers with a finite word length is generally contaminated by errors. Obviously the size of the error will decrease with increasing length of the binary representation. A designer must make a tradeoff between the complexity of the description (length of binary words) and the resulting error.

We must note that the quantitative measure of error depends on a somewhat arbitrarily selected distortion measure. The theoretical analysis is useful to the extent that the distortion measure such as Hamming distance or squared error is relevant to the task.

Loosely speaking,the relationship between the expected distortion D and the length of the optimally selected description is called the *rate distortion function* $R(D)$. If the code is optimized with respect to the selected distortion measure, it is possible to show that the rate distortion function is equal to the mutual information between the coded and the original representations (Cover and Thomas [1991]).

The exact shape of the rate distortion function depends on the distribution of the random variables, as well as on the selected measure of distortion. In situations in which the underlying distribution is Gaussian, and the distortion measure D is square error, the rate distortion function is given by

$$R(D) = \frac{1}{2} \log \frac{\sigma_s^2}{D},$$

where R is in bits, σ_s^2 is the variance of the underlying distribution, and $0 < D \leq \sigma_s^2$.

We entertain the hypothesis that the rate distortion function may account for a portion of the observed speed accuracy tradeoff. Suppose that the visual system refines its estimate of a normally distributed stimulus parameter over time by a binary search procedure. Then the response time would be proportional to the rate, i.e., $T = \alpha R$, where α is a positive constant. In that case, the variance in the stimulus representation after T seconds would be

$$\sigma^2 = \sigma_s^2 2^{-2\frac{T}{\alpha}} + \sigma_i^2,$$

where σ_i^2 is intrinsic visual system noise. The representation variance is the variability that would determine the stimulus discriminability and the response error rate.

Although it is not very likely the rate distortion function would account for the empirically observed speed-accuracy tradeoff, it may be a useful to analyze peripheral coding and processing of visual stimuli.

4.3 Translation Invariance

It is probably not surprising that continuity and parity tasks over a large visual field would benefit from eye movements. In the this section, however, we argue that even much less complex task, such as vernier acuity can benefit from a sequential approach that would compensate for the lack of translation invariance of the the visual system.

The capability of translation invariant spatial pattern recognition is of great importance to mechanical and biological visual systems. A translation invariant system is capable of recognizing patterns independently of their position in the visual field, as well as determining their position. The cost of translation invariant pattern recognition is reflected by an increase in complexity because any analysis must be performed at all locations in the visual field. For example, for the vernier acuity task described in Section 4.1 the complexity is on the order of $O(n^2)$. The complexity due to translation invariance might be quite high if the number of different patterns at any location increases. It would be useful to determine to what extent the human visual system obeys translation invariance.

Before we proceed any further, we must define translation invariance in terms of observable measures from behavioral experiments. A definition based on the ability to identify an object in the central and peripheral visual fields is not quite sufficient. Such a definition depends on the ensemble of

patterns used to test the invariance. For example, consider a letter identification task where observers are asked to identify a large capital letter "A" as different from the letter "B." The differences between the letters are so great, that the task is performed perfectly regardless of considerable distortions to the images. A more sensitive measurement (stricter definition) is required to determine whether the visual system is translation invariant.

A more strict definition of translation invariance can be based on the probability of discriminating between two similar objects. Let's denote the probability of discriminating between stimuli, defined by their local spatial parameters a and b located at eccentricity r, by $Pr\{a, b; r\}$. The local parameters might represent dimensions or relative positions of object features. For example the horizontal displacement of the lines could represent the stimulus in the vernier task. It is important that the value of the parameters a and b are chosen such that the discrimination probability is neither zero nor one. Such a choice of stimuli assures that they are not too different (always discriminable) nor too similar (indiscriminable). If a visual system is strictly translation invariant then

$$Pr\{a, b; r\} = Pr\{a, b; 0\}. \tag{14}$$

That is, the probability of correctly distinguishing between a and b should be independent of the location in the visual field.

Given this definition, it is obvious that the human visual system is not strictly translation invariant. In addition to the everyday experience that peripheral vision is not as acute as central vision, the lack of translation invariance is suggested by the nonuniform distribution of receptors in the retina. Most of the receptors are located in the central area of the retina called the fovea. The density of the receptors decreases rapidly with distance from the fovea.

Despite this severe violation of our strict definition of translation invariance, it is possible that the visual system is essentially translation invariant except for the non-uniform peripheral representation. This non-uniformity could be achieved by a spatial transformation, such as dilatation, whose parameters would depend on the eccentricity. In practice, such a transformation could be achieved by reducing the density of peripheral sampling and the corresponding representation of the retinal image. There appears to be some evidence of this type of transformation in physiological data.

If the lack of translation invariance can be fully accounted for by the peripheral representation, as some investigators have proposed, it should be

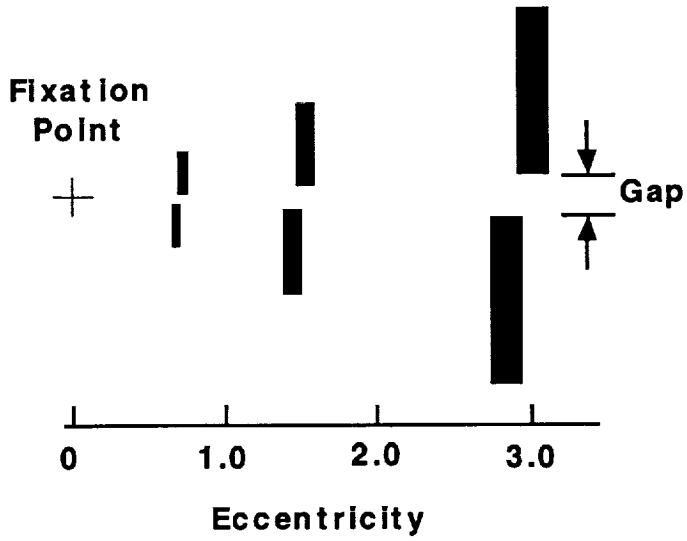

Figure 10: Vernier acuity task scaled for eccentricity to restore performance

possible to compensate for the loss of spatial sensitivity due to translation to the periphery, by a change in scale of the objects. Thus, two objects can be equally discriminable in the periphery as they are in the fovea provided that they appropriately enlarged. Mathematically, we restore translation invariance by enlarging each object by the same factor $m(r)$:

$$Pr\{m(r)a,\, m(r)b;\, r\} \;=\; Pr\{a,\, b;\, 0\}\,. \tag{15}$$

An example of restoration of translation invariance is shown in Figure 10. The scaling function $m(r)$ represents the derivative of the visual angle at eccentricity r with respect to the corresponding extent in the internal representation. Following the work of Anstis [1974] there are many studies concerning the performance of spatial discriminations in periphery (Levi [1985]). The results from various studies with different stimuli s are in general agreement that the scaling functions m_s are linear functions of the eccentricity, as shown in Figure 11. The difficulty of the task is expressed in terms of just noticeable differences (JND) which represent estimates of the physical offset, *delta*, that is required for 75% correct responses (the solid line). The scaling function for JNDs can be expressed as linear functions of eccentricity

$$m_s(r) \;=\; 1 \;+\; \frac{r}{r_s}, \tag{16}$$

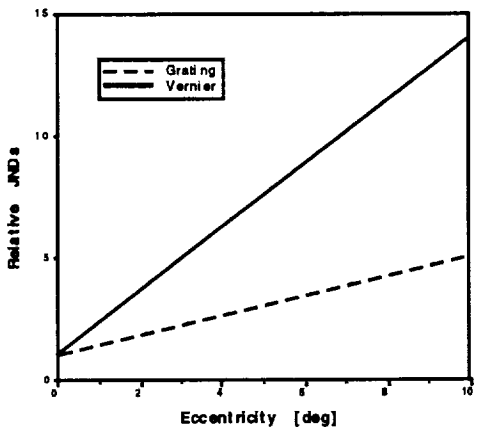

Figure 11: Plot of Just Noticeable Differences (JND) for a vernier acuity task as a function of the eccentricity of the vertical bars (solid line), and detection of grating (dashed line). JND is defined as the physical offset required for 75% correct discrimination on the vernier acuity task, and the contrast for 75% detection of sinusoidal gratings. Both functions were normalized to be equal to unity at the center of the visual field (fovea).

where r_s is a positive constant. The linearity of the function m, however, is neither necessary nor sufficient for the existence of translation invariance. That is, the scaling function m could be any positive function of eccentricity.

Recently, more careful investigations of the validity of translation invariance have been undertaken. To investigate whether stimulus scaling can compensate for a translation to the periphery requires that all spatial dimensions of the stimulus be scaled equally. Cunningham [1986] have examined the effect of the gap in the vernier acuity target, as shown in Figure 4b. The resulting performance as a function of gap size for stimuli presented at different eccentricities is shown in Figure 12. The fact that the performance at different eccentricities, as functions of gap size, are parallel can be used to prove that the visual system is translation invariant after appropriate scaling of all dimensions of the stimuli by the function m.

Whereas these results are encouraging in terms of restoring translation invariance, there is considerable evidence that the scaling function m depends on the specific task. We hypothesize that the scaling function m depends on the task complexity. In particular, the simpler, first order predicate spatial tasks seem to require lower scaling factors m than do the more complex

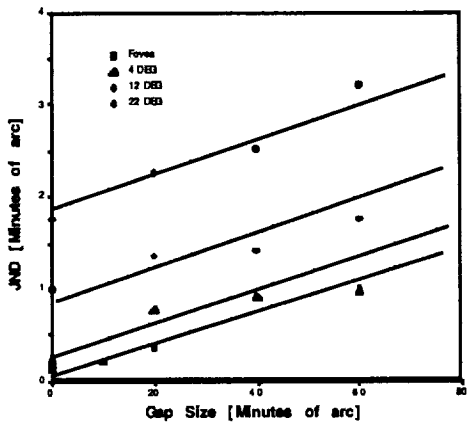

Figure 12: Plot of Just Noticeable Differences (JND) from a vernier acuity task as a function of the gap between the vertical bars. The parameter is the eccentricity of the stimulus (on nasal retina).

tasks. For example, performance on detection of the presence of a sinusoidal grating in the periphery deteriorates with eccentricity at a slower rate than does the performance of vernier acuity. In general, the performance on tasks requiring second and higher order of predicates deteriorates more rapidly with eccentricity.

To summarize, the human visual system can recognize patterns in the periphery in a similar manner as in central vision, but with lower spatial acuity. This is because a uniform increase in the size of objects is equivalent to a corresponding reduction in spatial frequencies. Thus, the visual system can be thought of as performing the preliminary, low acuity analysis in the periphery, and a high acuity analysis in central vision. The peripheral analysis provides the low acuity analysis and approximate information on the location of the object's image. This preliminary peripheral analysis is followed by a precise pattern analysis in central vision after the image of the object is centered using eye movements. Thus, a parallel pattern recognition task, requiring a large visual field and a high precision, is converted to a serial process consisting of locating, centering, and then analyzing individual patterns.

4.4 Summary of Eye Movements Effectiveness

Our objective of this part of the project was to examine how eye movements, mediate a tradeoff between performance and cost requirements on the visual vision system. The cost is in the complexity of physiological processing, the time, and the accuracy required to perform various tasks.

In the conclusion we would like to reiterate the three main points of our discussion:

1. We argued that computational complexity of a particular tasks can be used to assess the difficulty of such a task for the human visual system. Although the full power of the Kolmogorov-Chaitin complexity theory may not be applicable directly, the general approach involving the complexity of individual objects was shown to be potentially useful for anticipation of human observers performance.

2. Eye movements can be used to mediate a tradeoff between the complexity of a fixed, but fast pattern recognition machine and that of a sequential and slow, but flexible pattern recognition system. For complex visual tasks, a parallel solution would require too much parallel hardware that would have to be continuously adapted to each task. The visual system appears to economize on the number of parallel computations by taking advantage of the fact that sequential algorithms are actually preferable for some tasks. For those tasks, eye movements (or a moving camera) can provide a simple and effective means for converting computationally complex parallel tasks into serial ones.

3. Although the human visual system is, strictly speaking, not translation invariant, it can be transformed to one by a particular scaling (dilatation) transformation. The resulting representation provides a convenient tradeoff between requirements for high accuracy (high complexity) capabilities, and the size of the visual field that can be monitored in parallel.

We hope that our discussion will motivate more rigorous analyses of the performance of the human visual system. The results of such analyses would improve our understanding of the human visual system and at the same time provide engineers with new directions for designing artificial vision systems.

References

[1974] Anstis, S. M. (1974). A chart demonstrating variations in acuity with retinal position. *Vision Research, 14*, 589–592.

[1983] Bergen, J. R. & Julesz, B. (1983). Rapid discrimination of visual patterns. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13(5)*, 857–863.

[1993] Burbeck, C. A. & Hadden, S. (1993). Scaled position integration areas: accounting for weber's law for separation. *Journal of the Optical Society of America A, 10*, 5–15.

[1966] Chaitin, G. J. (1966). On the length of programs for computing binary sequences. *Journal of the Association for Computing Machinery, 13*, 547–569.

[1991] Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory.* New York: Wiley.

[1986] Cunningham, H. A. & Pavel, M. (1986). Judgements of position in near and far peripheral fields. *Invest. Ophthalmol. Visual Sci. Suppl., 27*, 95.

[1957] Kolmogorov, A. N. (1957). On the representation of continuous functions of many ariables by superposition of continuous functions of one variable nd addition. *Doklady Akademii Nauk USSR, 114*, 953–956.

[1965] Kolmogorov, A. N. (1965). Three approaches to quantitative definition of information. *Problems in Information Transmission, 1*, 4–7.

[1977] Kowler, E. & Steinman, R. M. (1977). The role of small saccades in counting. *Vision Research, 17*, 141–146.

[1985] Levi, D. M., Klein, S. A., & Aitsebaomo, P. K. (1985). Vermier acuity, crowding and cortical magnification. *Vision Research, 25*, 963–977.

[1995] M. Shiffrar, J. L. & Pavel, M. (1995). What is a corner? *Invest. Ophthalmol. Visual Sci. Suppl., 35*, 1277.

[1969] Minsky, M. & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry.* Cambridge, MA: MIT Press.

[1995] Pavel, M. & Cunningham, H. (1995). Eye movements and the complexity of visual processing. In M. Landy, L. T. Maloney, & M. Pavel (Eds.), *Active Vision.* New York: Springer.

[1997] Pavel, M. & Weinshall, D. (1997). Non-rigid perception of motion in 3d. *Invest. Ophthalmol. Visual Sci. Suppl., 37,* 95.

[1983] Pelli, D. G. (1983). The spatiotemporal spectrum of the equivalent noise of human vision. *Invest. Ophthalmol. Visual Sci. Suppl., 24,* 46.

[1979] Rovamo, J. & Virsu, V. (1979). An estimation and application of human magnification factor. *Experimental Brain Research, 37,* 495–510.

[1964] Solomonoff, R. J. (1964). A formal theory of inductive inference. *Information and Control, 4,* 224–254.

[1986] Sperling, G. & Dosher, B. (1986). Strategy and optimization in human information processing. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of Psychology.* New York: Wiley.

[1980] Triesman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97–136.

[1987] Wegener, I. (1987). *The Complexity of Boolean Functions.* New York: Wiley.