EYE MOVEMENTS AND VISUAL PERCEPTION: A "TWO VISUAL SYSTEM" MODEL[1]

by

Richard L. Didday[2]
and
Michael A. Arbib[3]

COINS Technical Report 73C-9

December, 1973

DEPT. OF COMPUTER & INFORMATION SCIENCE
GRADUATE RESEARCH CENTER
UNIVERSITY OF MASSACHUSETTS
AMHERST, MASSACHUSETTS 01002

---

[1] Preparation of this paper was supported in part by the Public Health Service under Grant No. 5 R01 NS09755-03 COM from the National Institute of Neurological Diseases and Stroke. Arbib presented portions of this paper to the Information Sciences Program of the University of Hawaii, and the Psychology Department of the University of Western Australia. His thanks to David Pager and John Ross for their hospitality and stimulating discussion, and to Drs. Noton and Stark for helpful comments upon an earlier draft. Didday would like to express thanks to Prof. Walter Orvedahl for his helpful comments on an earlier draft of this paper.

[2] Department of Information Sciences
University of California at Santa Cruz
Santa Cruz, California 95060

[3] Department of Computer and Information Science
University of Massachusetts at Amherst
Amherst, Massachusetts 01002

1. Introduction

Eye movement is one of the few externally measurable activities of visual perception. There have been a number of studies of eye movements in the last 150 years; Yarbus [31] lists about a hundred relevant references. There seems to be a trend toward using saccadic eye movements (the rapid, coordinated rotation of the eyes from one "point" of fixation to the next) to infer the functioning of brain processes which are not directly measurable. For example, [27] and [20] use measurements of such movements to discover chess players' board analysis techniques. Noton and Stark [21], [22], [23] have used such measurements to test a hypothesis about a memory scheme and its relationship to patterns of eye movements. Our paper presents a model of the role of eye movements which builds on neurophysiological investigations of the "two visual systems" of [15]; and which yields an alternative interpretation of Noton and Stark's results.

Yarbus [31] recorded the eye movements of subjects examining complex visual scenes (paintings and photographs). The subjects examined the scenes for up to three minutes, so each record is the result of hundreds of saccadic movements. Thus, areas often fixated are densely recorded. Yarbus uses these dense regions as a measure of analytical effort and shows that fixations can be directed to different regions of the scene by posing such questions to the subject as: "estimate the material circumstances of the family in the picture," or "give the ages of the people" [31] p. 174.

In Yarbus' data there is a tendency for eye movement paths as well as fixation points to be dense, much as if some few sequences of eye movements

were preferred. Noton and Stark have confirmed the existence of preferred sequences and dubbed them scanpaths. Using the diffuse scleral reflection technique [26], they recorded data from subjects who were told to "look at" the line drawings they were shown. Figure 1 is taken from [23], where we see that the idealized scanpath of G crudely approximates the actual sequence of fixations shown in the traces A - F. Their data indicates that the scan path varies from picture to picture for the same subject and from subject to subject for the same picture.

Noton and Stark hypothesize that as part of perception, an internal representation of objects is constructed as an assemblage of internal representations of features which are linked together in a feature ring, i.e., a sequence of sensory and motor memory traces, alternately recording a feature of the object and the eye movement required to reach the next feature.

They state three hypotheses.

1. The memory of an object is a piecemeal affair: an assemblage of memory traces of features which must be matched serially with the features of the visual scene before recognition is achieved. (In this way, their model predicts scanpaths.)
2. The features of an object are the parts of it that yield the most information.
3. The memory of an object is the feature ring corresponding to it. (That is, eye movement commands are stored as well as feature representations.)

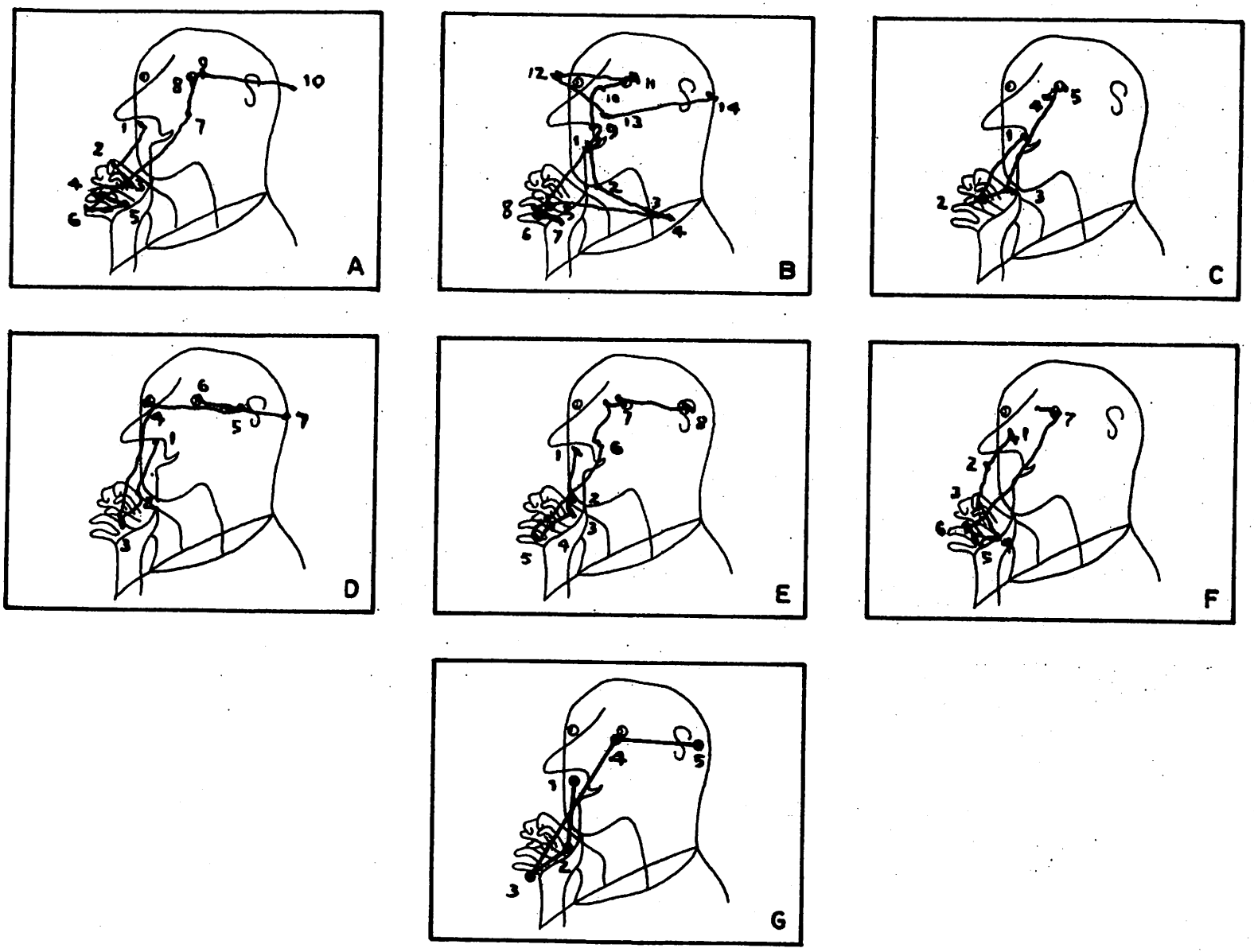


Figure 1

In the next section, we will present a model schema based on [3], [4], and [15] which also predicts scan paths but which is not a serial device and which does not store the attention shifts required to pass from feature to feature, i.e. does not involve feature rings. Let us first digress to use the second of Noton and Stark's assumptions above to ask a question.

What do we mean when we speak of the "information" in a visual scene? If we let a visual scene be a rectangular array of n small squares each one of which can (with equal probability) be black or white, we are forced to say that the information content of both scenes in Figure 2 is identical since each is one of the 2^n equally probable patterns. Clearly this approach (which ignores the "receiver") is not what is meant.

a. b.

Pictures of equal information content.

Figure 2

Information theory teaches that information is the reduction of uncertainty. However, the usual Shannonian theory is in purely probabilistic terms, in which it does not make sense to say that "the information gained by printing on a page was lost when the page was overprinted." To give "meaning" back to information, we should consider what sort of visual processing apparatus is observing the scene. For example, a scene presents different information content to a frog than the identical scene

presents to a cat. The cat is capable of perceiving "answers" to different visual "questions". However, "questions" may be set for the perceptual apparatus by "context", thus the information content of Yarbus' picture is different depending on which of the two questions mentioned above was asked of the observer.

For these reasons, we suggest that the phrase "features of an object are the parts of it that yield the most information" is simply a definition of the word "features", and gives little help in determining what features a particular visual system computes and uses. Attneave's famous sleeping cat [5] is often cited as evidence that humans use contours (edges) and points of high curvature of contours (angles) as features (regions of high information). All we can legitimately conclude, however, is that the transformation which Attneave applied to obtain his cat (i.e. blurring, thresholding, edge detection, selection of points of greatest edge curvature, and finally replacing edges by straight lines) does not destroy all the information which enables a human to answer the question, "What sort of animal is this and what is it doing?" There are, however, a great number of questions which a human looking at the original scene could answer, but can no longer answer from the transformed scene, such as "Does the cat have tufts of hair in its ears?" or "Is the cat's fur all one shade?" It is not sufficient to discuss the information <u>remaining</u> in a transformed scene in the hopes of deducing functions performed or features extracted in the visual system. It is also necessary to study the information <u>removed</u> by a transformation, since some types of transformation might be found to rule out specific classes of "questions", thus giving insights into the hierarchical interaction of various parts of the visual system.

Moreover, the notion that lines and angles are the features selected by the brain with which to build internal representations [23] is inconsistent with an assumption at the root of Noton and Stark's work--i.e. that a region of fixation is a feature. The fixation points shown in Figure 1 are not at all correlated with angles in the scene; in view of the fact that we can recognize single objects (such as a telephone) with a single fixation, we must expect fixation regions to contain complexes of low-level features. Experiments on chess board memorization [27] suggest that some fixations may take in at least four pieces and further, the relevant feature-regions seem to be high-level and both skill and problem dependent.

Let us turn now to one other point before presenting our model. Noton and Stark are interested in the question of whether object recognition is a serial or a parallel process. Although we could describe our model as a series of parallel processes, we tend to think that the serial/parallel dichotomy is not fruitful here. [Incidentally, [9] proposes a classification scheme for nervous system-like computational processes which assumes parallelism and discriminates on the basis of connectivities and information flow. This may turn out to be a more fruitful classification of nervous functions than serial-parallel.]

Recordings from the more peripheral areas of the visual system [13], [16], [19] indicate that visual information is being processed in parallel, with analysis of the entire visual field seeming to progress in parallel from layer to layer. If we view the brain as a collection of information processing units (the neurons), each of which is constantly using the values of its inputs to generate its output firing

level, then we might consider all the brain's activities as arising from a parallel process. Then again, if we follow the information flow as an image impinges upon the retina and its effect travels on through the visual system, we might note that until enough time has elapsed, a cell deep in the visual system will be unaffected by the latest input and so argue that the brain's operation is sequential. We are beginning to feel that terms like "serial", "sequential" and "parallel" are not fruitfully applied to brain processes since their meanings seem so overlapping and hazy. Perhaps describing the way(s) in which information flows through a system would be a more useful way to categorize what sort of processing is occurring. Let us, however, use one of the terms we have just disavowed to leave this point and return to our discussion of Noton and Stark's work.

At some level of abstraction we could perhaps say that a process performed by the brain is serial in organization, but we already know that the constraint imposed by the small area covered by foveal vision requires that the choice of next fixation point occurs serially irrespective of the mode of computation which controls it. Noton and Stark's model assumes more than this degree of seriality. They explicitly assume that serial eye movements become serial internal shifts of attention when an image is small enough to be seen in its entirety at one fixation point. While we cannot prove or disprove this, we feel that it becomes less likely in light of current notions [15], [24] that, to a first approximation, the processes of orientation ("where") are subserved by anatomically different regions of the mammalian brain than those of recognition ("what"). From this point of view, we would suspect that eye movements are dealt with by the orientation mechanisms and would have no a priori

reason to associate seriality to the recognition process. Even experiments which indicate that it takes longer to recognize more complex objects do not in and of themselves rule out a parallel process--in fact it is easy to define parallel processes with such behavior. This fact again indicates the questionable utility of the terms "serial" and "parallel".

2. A "Two Visual System" Model of Visual Perception

A previous paper [4] looked at memory and perception from what we called an "action-oriented" point of view. We concentrated on the idea that the "goal" of perception is not to build a little internal copy of the visual scene, but rather to assist making the decision of what action (if any) the organism should next initiate. We believe that any study which begins with the notion that perception is an easily identifiable, separable sub-part of the brain's activities runs the risk of conjuring up explanations and models which are flawed because they treat recognition as an end in itself. This does not deny that humans have evolved a "model-building" perceptual activity "atop" the basic "action-oriented" activity [3, pp. 16-17]. However, our contention remains that action-oriented perception is primary, and that no brain theory can succeed without such "first principle" considerations.

Our models [4] cast the perceptual system as a hierarchically organized system which uses a current model-of-the-world (a short-term memory) constructed from current input information plus associations from long-term memory. A point of importance is that we view not only effector movements but also pre-processing sub-systems (arrays of feature detectors) as being controllable by the system. Thus, we hypothesize that the "features" being utilized by the system change depending on what perceptual task is being carried out. This does not seem to be an unreasonable assumption, although it does perhaps ask for an analysis (which we do not give) of exactly what is meant by the word "feature". That some feature detectors in the classical sense develop as a function of experience has been shown by [12] and [6].

The general form of our model is shown in Figure 3, which is related to Figure 4 of [4]. In the next section, we shall present a stripped-down version of the model which has been computer simulated, and use it to give a direct comparison with the feature ring model of Noton and Stark. Let us here give a verbal description of our model, following through Figure 3.

As is well known, the afferent signals from the eye (A) traverse two distinct pathways (among others), the geniculo-striate pathway leading to visual cortex (here caricatured by regions (B) and (C)), and the colli-cular pathway leading to the superior colliculus (H). We recognize that this dichotomy of destinations is not so strict in real visual systems and that there are more than two destinations for visual projections, but feel that it is premature to incorporate these details if we are to achieve a testable model.

The cortical visual information is processed to yield what we have termed "output feature clusters" [4]. These are internal encodings of high-level motor commands which, if released, act in concert to produce an appropriate response (one such response might be to move the eyes). Part of the effect of activating an "output feature cluster" is to enter into part of short-term memory a representation of the way the recognized visual feature will affect the peripheral visual mechanisms. We hypothe-size that there is a region (D) (called the "slide box" for reasons elaborated in [3, Section 4.1]) that holds the updated encoding of the "percepts" of a scene as a spatially-coded array of output feature clusters. Since the current input may suggest a number of possible ways to respond, there will at any time be a number of OFCs in (D) which cannot be acted

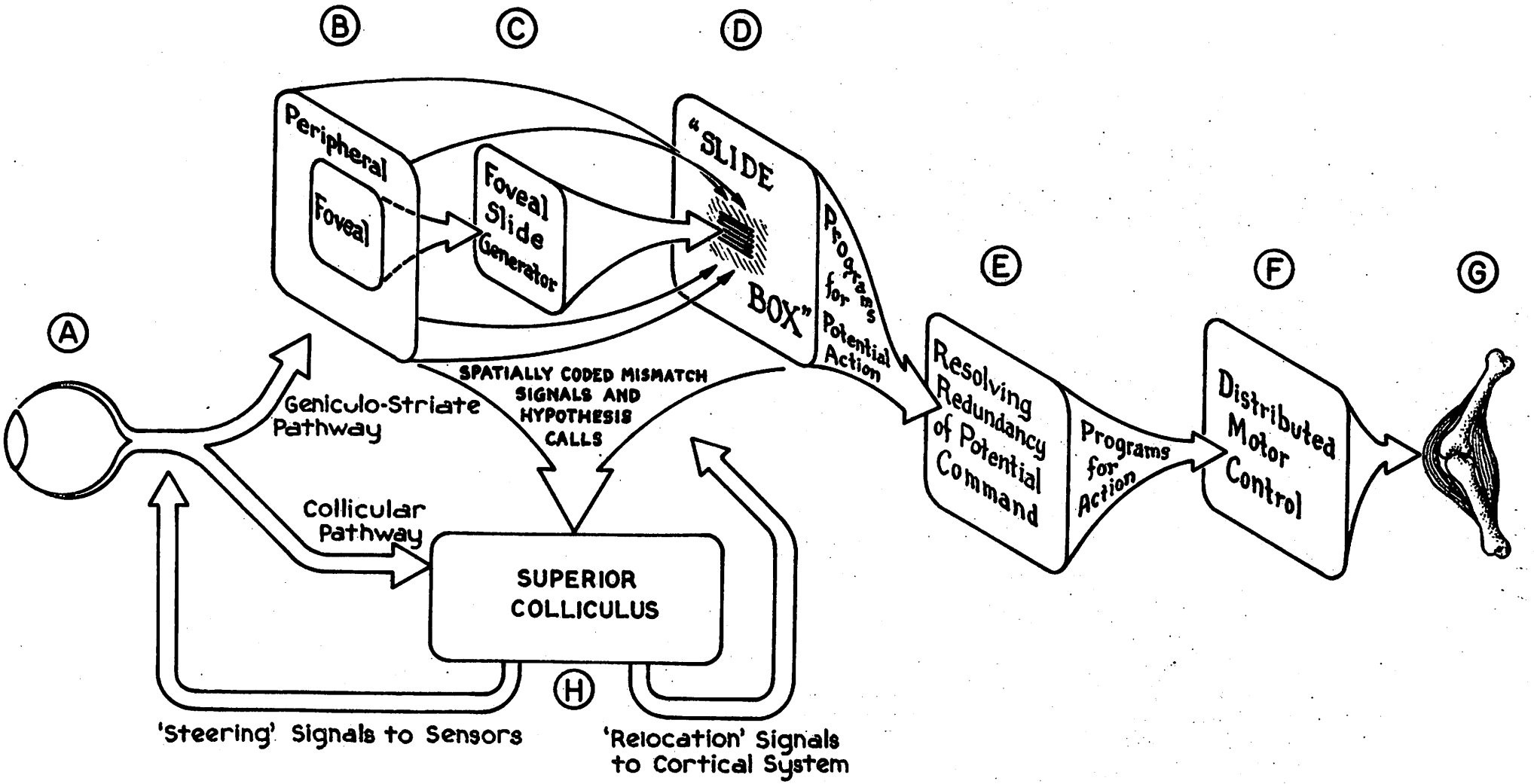

B
C
D
Peripheral
Foveal
Foveal Slide Generator
"SLIDE BOX"
Programs for Potential Action
E
F
G
A
SPATIALLY CODED MISMATCH SIGNALS AND HYPOTHESIS CALLS
Geniculo-Striate Pathway
Resolving Redundancy of Potential Command
Programs for Action
Distributed Motor Control
Collicular Pathway
SUPERIOR COLLICULUS
H
'Steering' Signals to Sensors
'Relocation' Signals to Cortical System

Figure 3

upon simultaneously. Box (E) then "resolves the redundancy of potential command" in this array, activating the distributed motor control (F) to recode the activated program in terms of the motoneuron activity which will control the skeletal musculature (G). We shall henceforth ignore those portions of the output system (E, F, G) which are uninvolved in the control of eye movements, and instead concentrate upon the interaction between the cortical system (B, C, D) and the midbrain system (H) in visual perception. (In [4], we considered some of the transformations that must occur in (E, F, G) to keep the internal model in register with the action frame of the organism.) The output from (H) labeled " 'steering' signals to sensors" is all that we will deal with from the output system, and we have, for convenience, drawn it separately from the rest of the output system.

Our model is based upon the premise that evolution (in the guise of comparative studies) can provide essential clues in our study of the human brain. Thus, rather than try to develop a model of human visual perception ab initio, we instead try to model it as a two-level system, in which the superior colliculus plays a role akin to that played by the tectum in the frog [7, 10, 14, 16, 19] (namely, orienting the system to local features of the environment) in addition to a new role played in interaction with the cortical mechanisms (B, C, D) which construct the organism's model of its current environment.

The "slide-box" (D) is to be thought of as containing an overall pattern of activation built up by the "insertion" of output feature clusters (OFCs) corresponding to different objects or well-defined sub-scenes in the environment. This "insertion" we now posit to be the task of the "foveal slide generator" (C). The crucial observation here is that

human vision is acute foveally but relatively diffuse peripherally. Thus we would predict that elicitation of all but the most familiar OFCs would require cuing by matching of an array of foveal features. Clearly then, the visual system must be able to steer the eye to bring such cuing arrays of features into the fovea so that system (C) may retrieve the OFCs for "insertion" into the "slide-box" (D).

It may help to think of the "slides" as "two-sided", with the "input-side" appropriate for matching with sensory features, while the "output-side" is so configured as to provide appropriate program signals to (E) and (F). At any time, the "slide-box" will be partially "filled", corresponding to the cumulative effect of past "insertions". Of this activity, only a fraction need correspond to current visual input--we have indicated this by the regions with horizontal hatching (foveal input) and diagonal hatching (peripheral input) in (D). We hypothesize that the "input-side" of this activity is compared in <u>parallel</u> with the current visual input to see how well the correspondence is made. We take it to be part of the specification of an OFC that the detail of some areas is crucial, while that of other areas is unimportant [if <u>that</u> OFC is indeed the correct one--this does not rule out that the foveal slide generator may seize upon these very features to elicit an OFC other than the one currently "active" in the "slide-box"].

Watanabe and Yoshida [29] report a tantalizing experiment, though--unfortunately--with only one subject. A TV screen was so connected to a device for detecting the subject's fixation point by using cornea-reflected light that all but that portion of the TV image in the 3° about the subject's line of sight was invisible to the subject at any time, no matter what his eye movements. In such a setup, the recognition

of simple patterns was delayed, and the subject was unable to recognize more complex patterns (Japanese letters composed of disconnected parts; caricatures of Nixon and de Gaulle) within 100 seconds. After viewing (but not recognizing) the complex letter, the subject was asked to sketch what he had seen--on observing the spatial relation between the "fragments" he had seen as made explicit in his drawing, the subject immediately recognized the letter. This suggests (the experiments are too sparse for us to draw a firm conclusion) that peripheral vision not only plays a role in directing saccades, but also provides a spatial framework or "organizing field" which maintains the spatial relationships between already inspected subpatterns. This accords with the subjective experience that if an object in peripheral vision of a well-perceived scene starts to move, one seems to know what is moving and where it is. Detailed studies with on-line computer generation of movies, coupled to line-of-sight monitors, seem called for to further resolve this intriguing problem.

Returning to Figure 3, we posit that mismatch of OFC and visual input in a given region of the spatial frame increases with discrepancy between the input and the "input-side" of the OFC, but decreases to the extent that that region is given low weight in the slide-specification. In any case, we posit that, as a result of this comparison process, a spatial array of mismatch signals is played down upon the superior colliculus, in spatial register with the "local feature map" played upon it by the collicular pathway. The visual cues reaching foveal slide generator (C) may correspond to several different slides, with further information being required to confirm or disconfirm these several

hypotheses. Often, the approximate spatial location of decisive information can be determined from the current activity in (C), and it is then posited that it is signals of this kind, weighted as to the strength they would have in confirming or disconfirming an hypothesis that make up the "hypothesis confirmation" map which provides the third spatial input array to superior colliculus in our model. There will also be contributions to this map from the "slide-box"--our current knowledge of the environment may suggest the presence of other objects in the environment even in the absence of visual stimuli cuing the object at that time.

Before suggesting how the three spatial arrays are combined in superior colliculus, we briefly mention two other mechanisms required for the proper functioning of the cortical system (B, C, D). Firstly, we require proper control of the pathways between (B), (C), and (D) to appropriately 'relocate' signals to the cortical system, despite varying orientations of the eyes. In [4], we have made a start in this direction, building upon the "reafference principle" of von Holst and Mittelstaedt [28]. Here we would merely add our agreement with MacKay [18] that one cannot expect such reafference to yield more than approximate compensation, and suggest that it is decomposition of "slide-box" activity in terms of OFCs corresponding to "objects" that then allows the necessary fine-tuning, since an OFC can then be transformed (or be transformed with respect to) in toto until sufficient register is obtained given an initial "rough" transformation. Occasionally, of course, an object will appear so unexpectedly different from a new perspective that a new "slide" must be found to represent it. Which brings us to our next point:

The foveal slide generator (C) may generate several OFCs before

actually "inserting" one in the "slide-box"; and what is essentially another resolution of redundancy of potential command mechanism must be provided to evaluate the slides suggested by (C) in comparison to the foveal input, the current "slide" in the horizontally hatched region of (D), and the overall context established by the current totality of activation in the "slide-box".

To complete our model-schema, it only remains to specify the function of the superior colliculus (H). In Didday's model of the frog tectum [7, 10], the tectal input was caricatured as a spatial array coding "food-ness"; and it was posited that a layer of computation resolved redundancy of potential command within this array, and that the resolved output played upon a distributed motor control to cause the frog to orient towards, or snap at, the prey object whose visual image achieved actual "command" (ironic term, poor worm) of the system. In our current model, we posit that the superior colliculus has a computing layer that resolves redundancy of potential command in its input array in the manner in which Didday's model of the tectum resolves redundancy; but that the resolved output plays upon a distributed motor control to cause a saccade of the eyes towards the region of the visual field in spatial register with the region of the input array which achieved actual "command" of the system. It remains to specify the nature of the input array to the redundancy of potential command resolver.

In each region of the collicular input array, we have three types of signal--a "low level feature" signal from the collicular pathway; and "mismatch" and "hypothesis confirmation" signals from the cortical system (B, C, D). These are combined to yield one "attention-worth" signal. The exact formula for combination need not concern us here--the crucial point is that it is monotonic in all three variables, and that the relative importance of these three variables can be changed depending on what we

shall loosely refer to as the "affective state" of the organism. [When "jittery", the collicular pathway may dominate--"we are easily distracted"--but when "daydreaming", the "hypothesis confirmation" signals may dominate visual perception to the virtual exclusion of "mismatch" and collicular pathway signals.]

The overall effect of this model is that different regions of the visual field will compete for the attention of the organism. The evaluation of "attention-worth" of the region will depend upon the intrinsic novelty (such as an unexpected flash of light) of the low-level features of the region, the degree of mismatch between high-level features of the region and the current internal model of that region, and the extent to which the organism posits that the region contains perceptually important information. We could also suggest that collicular input would be depressed in regions for which the slide-box contains OFCs which have received a high degree of confirmation. The regions then "compete", in a structure akin to that provided by [7]. If and when a region emerges as the "winner" from such a resolution of redundancy of potential command, the superior colliculus will both direct a saccade of the eyes towards this region, as well as send a "relocation" signal to the cortical system to ensure that foveal input will be routed towards the appropriate region of the slide-box.

model Qualifications logical model

Before proceeding to our simplified model, let us pause to make very clear the status in fact of a number of our specific hypotheses. Although we have tried to follow current neurophysiological thinking wherever feasible, this remains to great extent a conceptual or <u>logical</u> model. We have used anatomical terms to aid the exposition and have suggested where

certain subsystems might lie, but many of the specific details are open to doubt. For example, while the superior colliculus is strongly implicated in orientation behavior [15], and indeed in visual attention processes [30], and while a number of topographic spatially arrayed pathways from cortex to superior colliculus are known to exist [11], the specific nature of these pathways is not yet understood. Thus our guesses about specific sorts of information and their arrangements which are played down upon superior colliculus are to a large extent just that--guesses. Further, the locations and precise functioning of a number of sybsystems we have mentioned (feature detector layers, short-term memory) are largely unknown physiologically.

3. The Simplified Model

In order to verify that the model we have described actually works as we believe, we have computer simulated a simplified version of it and observed its behavior. We view a computer simulation used in this way much as a proof. A disadvantage which appears when this sort of proof is compared with a formal mathematical proof is that the computer simulation must deal in specifics--i.e. for each run, all parameters must be given specific values. Thus, we are not able to prove that our mechanism works as desired in all possible situations. Nonetheless, the problems of attempting a formal mathematical theorem which would capture our model's operation strongly argue, at least in the short run, the advantage of computer simulation. The other problems with computer simulations--that there may be programming errors which lead to a false idea of what the model predicts and that the computer program might in fact not simulate what the user says--are problems which are shared completely by mathematical proofs.

To obtain a model amenable to simple computer simulation, we have "collapsed" the model of Figure 3 to obtain that shown in Figure 4.

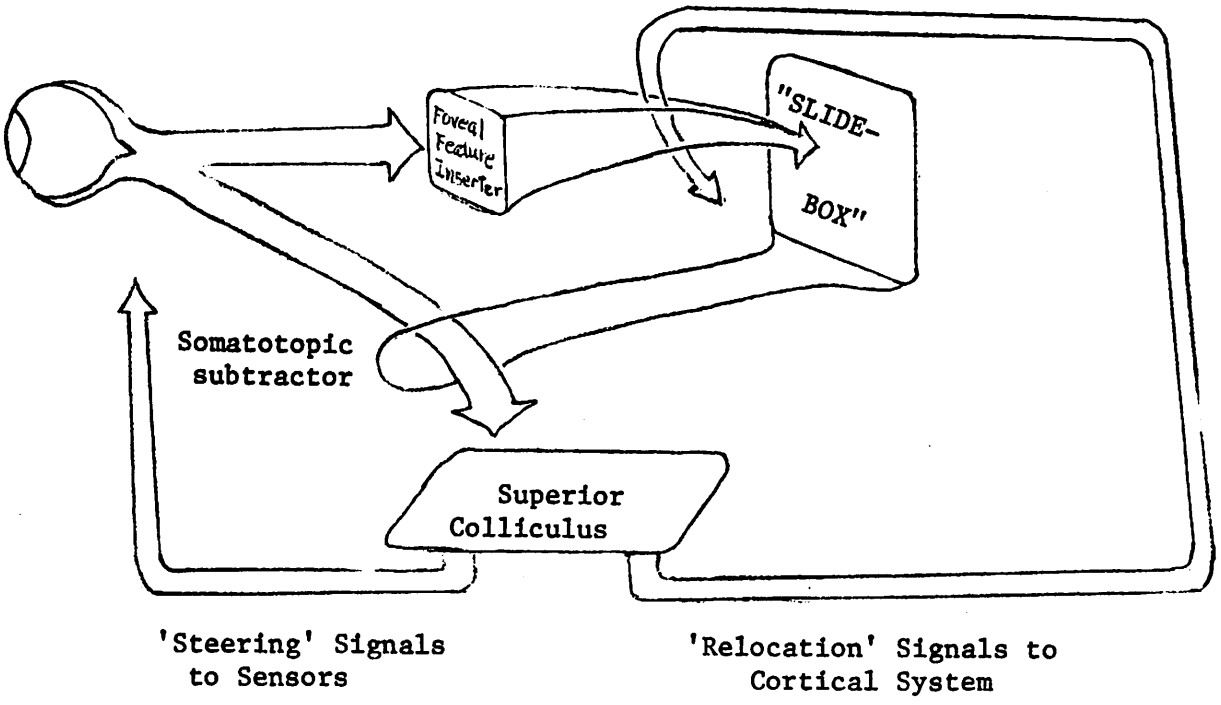


Figure 4

We collapse the cortical system (B, C, D) into a system whose job is simply to insert into the "slide-box" (which we think of as the short-term memory--though this is not the standard usage of the term) a representation of the feature-region to which the fovea is currently directed. This simplification means that, just as in Noton and Stark's feature ring system, long-term memory, OFC selection, and selection of the best percept (OFC) for a given foveal input scene all reduce to the equivalent of table look-up.

A second simplification is that we have expressed the computation of what we termed the "attention-worth"signal as simple somatotopic subtraction. Thus, the subsystem labeled "somatotopic subtractor" simply performs a pointwise subtraction between the spatially arrayed visual input to superior colliculus and the spatially arrayed short-term memory ("slide-box").

A third simplification concerns processing in the retina and feature detection layers (which are somewhat arbitrarily assumed to lie in the superficial layers of superior colliculus). Rather than representing a visual scene producing an intensity pattern impinging on a retina which sends excitation on, including through the feature detection layers, we have expressed the visual scene as an array of numerical values. The numerical values correspond to the level of excitation of the corresponding point in the visual projections after the feature detection layer. Thus, since we do not explicitly represent the feature detection layer, we simulate the effect of different feature detector settings by altering the numbers (but not their positions, of course) in the array storing the "visual scene". The only feature detection process that the program carries out is to slightly bias the input excitation levels in favor of ones which would lead to smaller eye movements.

Evidence indicates ([28], [18]) that voluntary eye movements are compensated for internally so that even though the visual scene has moved with respect to the retina, the perceived world holds still. We have implemented this by shifting the excitation which represents the system's short-term memory to account for effects of the eye movement carried out.

The simulation proceeds using techniques spelled out in [8] to specify the behavior of each simulated neuron and the position of the eye. The technique simulates the parallel operation of each unit by sweeping through and updating each cell's output firing level each time step. This is done using each cell's definition, which is an expression giving that cell's next output level in terms of prior values of its inputs.

Let us verbally follow the flow of visual information through the simulated system. The input is an array of numbers describing what features

are present at what points in the visual field. The position of the eye is specified by variables representing the vertical and horizontal positions of the fovea. The part of the visual scene which is before the fovea travels the simulated pathway to visual cortex where recognition of whatever feature appears there occurs, and an encoding of that feature is placed in short-term memory. Short-term memory is represented as a spatially arrayed layer of cells which holds all features represented so far, and which by shifting its excitation pattern to counteract eye movements maintains them in the same spatial relations in which they were found.

Let us back up for a moment and follow the excitation which leaves the eye by way of the "collicular" pathway of Figure 4. This pattern flows to the simulated colliculus through the "somatotopic subtractor". Thus, the only excitation which will reach the colliculus is in regions where short-term memory stores either no perceived feature or an incorrect feature.

Thus, the input to the simulated colliculus is a spatial array of excitation levels caused by as yet unperceived features (i.e. those which have not yet been analyzed foveally). The remnant of explicitly programmed feature detection occurs here, and biases the excitation levels so that those which would result in smaller eye movements are slightly favored. The task the "colliculus" faces is to choose the next region for foveal analysis by selecting an appropriate region (the region with maximal excitation) and by releasing that excitation to the spatially arrayed motor command units, causing an eye movement which will aim the fovea at the corresponding region of the visual field. In the simulation being

reported here, this process was implemented as in [7] and [10], but [9] analyzes several alternative methods of achieving this decision-making function. Essentially what happens is that a layer of cells called "sameness" cells (because of their similarity to the cells reported by [17]) conducts a competition process among sub-regions until one sub-region "wins" and is released. The criterion for release is set by a threshold in the sameness cell layer. If one cell's firing level crosses threshold, the corresponding region is released.

The position in the array of the excitation which was released determines exactly what eye movement is to be made. Eye movement is simulated simply by updating the variables storing the eye position. In addition, the representation of the visual scene in short-term memory is shifted so that the change in retinal input caused by the eye movement can be accounted for.

In operation, one eye movement follows another until all features in the visual scene have been "recognized" and entered into short-term memory. At that point, there are no discrepancies between the visual input and short-term memory, so no excitation enters the "colliculus" decision surface. We shall comment later that in a more realistic case we might expect processing to continue to refine the internal "understanding" of the visual scene. Here a feature seen foveally is perfectly recognized the first (and hence, last) time it is seen.

4. Results of Running the Computer Model

Noton and Stark's data illustrate first that in a given subject, scanpaths vary from picture to picture. They note that this rules out the possibility that scanpaths merely represent habitual ways of looking over a picture. Second, they show that for a given picture, scanpaths vary from subject to subject. They argue that this rules out the explanation that scanpaths ". . . result from peripheral feature detectors that control eye movements independent of the recognition process, since these detectors . . " would be low-level, genetically determined and hence similar in all subjects ([23] p. 40, italics ours). By adding the hypothesis that the weight given such feature detectors are in fact under control of the recognition process, we are able to demonstrate behavior similar to the feature ring. In fact, in the simplest form of our model, all that is necessary is that different subjects' visual systems value features differently.

Figure 6 illustrates the results of running our simulation model. We chose 7 different curves (Figure 5) to be the features and created three different scenes by rotating and translating the basic features. (The assumption of rotation and translation invariance of the feature detectors is irrelevant to our point and was done only to make Figure 6 more interesting for humans to look at.) The three scenes were presented to our system by giving the coordinates and level of excitation caused by each feature. Simulating different observers was accomplished by assigning different levels of excitation to each feature for the two observers.

Figure 6 shows that our system produces the same sort of results as Noton and Stark's feature rings, namely that the pattern of eye movements

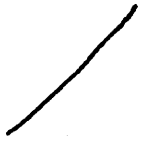
	feature	represented as one region of the visual input with excitation value: observer 1	observer 2
1		3.	5.
2		4.	9.
3		5.	7.
4		6.	4.
5		9.	6.
6		7.	2.
7		2.	3.

Figure 5

is different for different pictures seen by the same observer and different for the same picture seen by different observers.

Even this simplified model shows that a parallel system can, without storing any information about eye movements previously used, demonstrate behavior like that of the serial feature ring system which does explicitly store such information. Let us now discuss what details we must put back into our computer simulated model from our verbal model which will let it explain the experimental results as well.

Probably the most glaring difference between the models (i.e. Noton and Stark's and ours) and the data is that in real life, once a scanpath

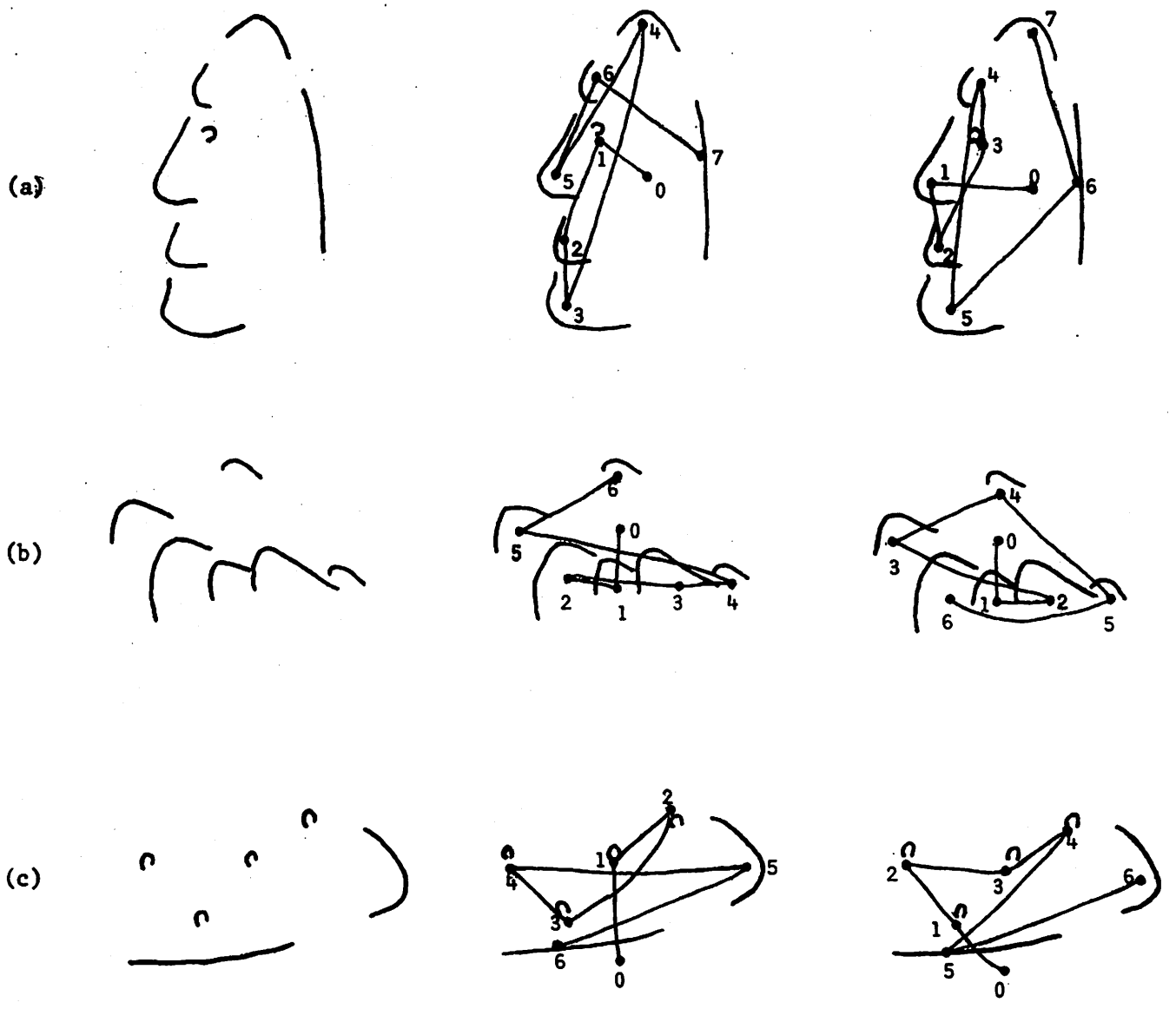


Figure 6

Scanpaths

(a) Easter Island Face

(b) Cloud Hazed Mountains

(c) Lower Side of a Dill Pickle

is completed, eye movements continue. The other is that scanpaths do not occur all the time--they represent a tendency, not certainty.

In studying the visual system, we are faced with a system of immense complexity and power. Although we feel safe in stating that perception is a hierarchical process, it is by no means clear how many levels there might be or along what lines they are organized. We would all agree that the following two situations illustrate a hierarchy, but where do we go from that?

1. A sudden, localized flash of light makes a subject jerk her eyes in the direction of the flash: we agree that the movement is a reflex.
2. We show a subject the painting Giovanni Arnolfini and His Bride by Jan Van Eyck, ask "Is the woman pregnant?" and observe her eye movements: we agree that the eye movements are very high-level and calculated.

In the situations we wish to explain, namely Noton and Stark's data, the subjects are operating in between these two levels. We look at the task of securing recognition by constructing an internal model in short-term memory as a task of hypothesis generation and testing. In a redundant visual world it is computationally more efficient to check for the presence of a feature which would be expected according to a current hypothesis than it is to go flitting about trying to recognize each foveal scene from scratch and then trying to deduce what object we have looked at.

In perceiving a scene, the subject looks from region to region of the scene until she has built up a satisfactory internal representation of the scene, with the next region of fixation depending on the currently attended

to region, and the subject's current visual hypotheses and their state of confirmation. These hypotheses are reflected in the settings of the feature detectors which in turn bias the choice of the next peripheral feature by the decision surface of Figure 3. Thus, repeated viewing would follow from a subject refining her hypotheses and non-scanpath eye movements would follow from momentary testing of a different internal perceptual hypothesis.

We find it interesting that Noton and Stark report that scanpaths occurred in the initial part of the recognition phase a large percentage of time--up to 65%. This is in concert with the notion that in recognition, the subject "knows" what she is looking for--feature detectors are purposively set to seek out features which will confirm the hypothesis that it is an already familiar object she is now seeing. Noton and Stark report [22] that in such cases subjects (when asked later) said they easily recognized the scenes as ones they had seen before.

One last elaboration is that our system builds up an internal model of a spatial array of objects, while Noton and Stark's feature ring system is specifically aimed at recognizing one object at a time, ignoring other objects as if they were noise. We suspect the notion of what an "object" is to be also dependent on what perceptual "question" is being worked on. In the face of Figure 1, at one time we might view the ear as an object, at another the head, at another the entire drawing and at another the entire picture (including the boundary).

5. Conclusions and Suggestions for Further Experiments

We have presented a model of visual perception and eye movements based on [4]. We have rigidly specified a simplified special case model in the form of a computer simulation and demonstrated that it predicts the same behavior as a system using feature rings. We have indicated that our full model is rich enough to explain (although at a vague level) Noton and Stark's experimental results. Let us close by considering some of the major distinctions between our model and that of Noton and Stark, and suggesting some experiments to better compare its merits with those of the model presented here.

Probably the most crucial distinction is over the need to explicitly store eye movement commands with stimulus features. The basic idea of a feature ring which stores the scanpath required to recognize an object is not an uneconomical way of storing the memory of an object. However, to explain the experimental data, Noton was forced to elaborate the scheme to make provision for memory traces recording other eye movements between features not adjacent [or recorded] in the ring. The original ring is then taken to represent the preferred and habitual order of processing rather than the inevitable order, thus "allowing" the occasional substitution of an abnormal order for the scanpath. Noton suggests that each scanpath is selected arbitrarily, but then through habit becomes fixed and characteristic for a given person viewing a given pattern.

However, once one stores the whole feature network, rather than just the scanpath, the combinatorics involved make it seem more economical to adopt our "slide-box" strategy of storing the internal representation as some homomorph of a two-dimensional array in which eye-movement commands

are only implicit. The feature ring idea seems more like current digital computer data structures than a model of perception.

The other main difference in the two models lies in the question of how pervasive is the seriality forced by the nature of the human eye. Our model assumes that the process which implements the shift of visual attention from one point to another is parallel (in so far as the terms "parallel" and "serial" are appropriate to the brain). Noton and Stark's model assumes that this structure is itself serial in nature.

When a scene is sufficiently reduced in size, two points that were previously separated sufficiently to require a saccade between their fixations are now close enough to be fully inspected in a single fixation. Noton and Stark hypothesize that small shifts of attention, e.g. 1/2°, be carried out internally to move the analysis to a new area of the visual field in the neighborhood of the center of fixation--but that the matching of an image to its internal representation should involve the same feature network, with only the nature of the shifts of attention (external or internal) varying with change of scale. However, Noton and Stark's experiments do not show whether "internal fixation points" are placed along a scanpath, nor do they rule out the possibility that it is the question "Where do you think you are looking" that forces a localization that might otherwise be absent.

We would thus expect the Noton-Stark model to yield the following precise prediction--that the sequence of fixations yielding the scanpath for a scene at low magnification should be a "lumping" of the scanpath for the scene at high magnification, with the fixation time for each "lump" of the "lumped" scanpath approximating the sum of the fixation times for all the points of the original scanpath which were lumped into it. The approximate nature results from the reasonable assumption that Noton and

Stark would not expect the time taken for an internal shift of attention to equal that required for an overt eye movement, but would expect the "processing time" for a feature to be independent of the mode of attention shift.

Another type of experiment which would more clearly draw a distinction between the two models would consist of trying to fool subjects by changing features which do not lie on their scanpath. Since scanpaths occur so frequently during the "recognition" phase of their experiments, it would seem that according to their model, a subject could miss noticing the changed (non-scanpath) features. According to our (verbal) model, such altered regions would quickly become the target for fixations since they would contradict the internal hypothesis that the current scene was a familiar one.

It is our hypothesis (specified as our model) that scanpaths arise from the process whereby the brain secures an acceptable internal representation rather than being <u>part</u> of the internal representation. It is no more the case that the memory traces recording sensory features are assembled into the complete internal representation by other traces explicitly recording the shifts of attention required to pass from feature to feature than it is the case that the movements we make between pencil strokes in printing the letter "A" are part of the "memory trace" of graphite deposits that remain when the drawing is completed.

References

[1] M. A. Arbib, Theories of Abstract Automata, Englewood Cliffs, N.J.: Prentice-Hall (1969).

[2] M. A. Arbib, Math. Biosciences, 11, 95-107 (1971).

[3] M. A. Arbib, The Metaphorical Brain, Wiley-Interscience, New York (1972).

[4] M. A. Arbib and R. L. Didday, J. Cybernetics, 1, 3-18 (1971).

[5] F. Attneave, Psychol. Rev., 61, 183-193 (1954).

[6] C. Blakemore and G. F. Cooper, Nature, 228, 447-448 (1970).

[7] R. L. Didday, Tech. Report 6112-1, Information Systems Laboratory, Stanford Univ. (1970).

[8] R. L. Didday, Intl. J. Man-Machine Studies, 3, 99-126 (1971).

[9] R. L. Didday, Intl. J. Man-Machine Studies, 4(4) 439-457 (1972).

[10] R. L. Didday, in preparation.

[11] B. Gordon, Sci. Amer., 227, 6, 72-82 (1972).

[12] H.V.B. Hirsch and D. N. Spinelli, Science, 168, 869-871 (1970).

[13] D. H. Hubel and T. N. Wiesel, J. Physiol. 160, 106-154 (1962).

[14] D. Ingle, Brain, Behav. Evol., 3, 57-71 (1970).

[15] D. Ingle, G. E. Schneider, C. B. Trevarthen and R. Held, Psychologische Forschung, 31 (1967).

[16] J. Y. Lettvin, et al., Proc. IRE, 47, 1940-1951 (1959).

[17] J. Y. Lettvin, et al., Sensory Communication, (ed. W. A. Rosenblith), 757-776 (1959).

[18] D. M. MacKay, Invest. Ophthalmology, 11, 6, 518-524 (1972).

[19] H. R. Maturana, et al., J. Gen. Physiol. 43 Suppl., 129-175 (1960).

[20] A. Newell, in Theoretical Approaches to Non-Numerical Problem-Solving (R. Banerji, M. D. Mesarović, Eds.) Springer-Verlag, 363-400 (1970).

[21] D. Noton, IEEE Trans. System Science and Cybernetics SSC-6, 349-357 (1970).

[22] D. Noton and L. Stark, Science, 171, 308-311 (1971).

[23] D. Noton and L. Stark, Sci. Amer., 224, 34-43 (1971).

[24] G. E. Schneider, Science, 163, 895-902 (1969).

[25] P. H. Schiller, Invest. Opthalmology, 11, 6, 451-460 (1972).

[26] L. Stark, Neurological Control Systems, Washington: Plenum Press (1968).

[27] O. K. Tichomirov and E. D. Poznyanskaya, Soviet Psychology, 5, 2 (winter 1966-1967).

[28] E. von Holst and H. Mittelstaedt, Naturwiss, 37, 464-476 (1950).

[29] A. Watanabe and T. Yoshida, Role of Central and Peripheral Vision in Pattern Perception, NHK Technical Monograph 20 (1973), Tokyo: NHK Broadcasting Science Research Laboratories.

[30] R. H. Wurtz and M. E. Goldberg, Invest. Ophthalmology, 11, 6, 441-450 (1972).

[31] A. L. Yarbus, Eye Movements and Vision, Plenum: New York (1967).