Cognition, 18 (1984) 97–159

Visual routines*

SHIMON ULLMAN

Massachusetts Institute of Technology

Abstract

This paper examines the processing of visual information beyond the creation of the early representations. A fundamental requirement at this level is the capacity to establish visually abstract shape properties and spatial relations. This capacity plays a major role in object recognition, visually guided manipulation, and more abstract visual thinking.

For the human visual system, the perception of spatial properties and relations that are complex from a computational standpoint nevertheless often appears deceivingly immediate and effortless. The proficiency of the human system in analyzing spatial information far surpasses the capacities of current artificial systems. The study of the computations that underlie this competence may therefore lead to the development of new more efficient methods for the spatial analysis of visual information.

The perception of abstract shape properties and spatial relations raises fundamental difficulties with major implications for the overall processing of visual information. It will be argued that the computation of spatial relations divides the analysis of visual information into two main stages. The first is the bottom-up creation of certain representations of the visible environment. The second stage involves the application of processes called 'visual routines' to the representations constructed in the first stage. These routines can establish properties and relations that cannot be represented explicitly in the initial representations.

Visual routines are composed of sequences of elemental operations. Routines for different properties and relations share elemental operations. Using a fixed set of basic operations, the visual system can assemble different routines to extract an unbounded variety of shape properties and spatial relations.

---

*This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-80-C-0505 and in part by National Science Foundation Grant 79-23110MCS. Reprint requests should be sent to Shimon Ullman Department of Psychology and Artificial Intelligence Laboratory, M.I.T., Cambridge, MA 02139, U.S.A.

0010-0277/84/$19.40

At a more detailed level, a number of plausible basic operations are suggested, based primarily on their potential usefulness, and supported in part by empirical evidence. The operations discussed include shifting of the processing focus, indexing to an odd-man-out location, bounded activation, boundary tracing, and marking. The problem of assembling such elemental operations into meaningful visual routines is discussed briefly.

1. The perception of spatial relations

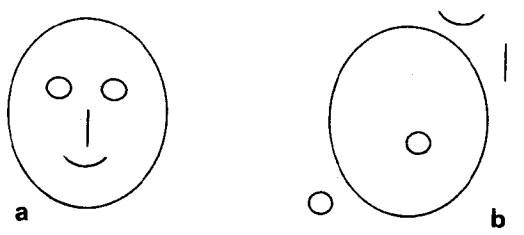
1.1. Introduction

Visual perception requires the capacity to extract shape properties and spatial relations among objects and objects' parts. This capacity is fundamental to visual recognition, since objects are often defined visually by abstract shape properties and spatial relations among their components.

A simple example is illustrated in Fig. 1a, which is readily perceived as representing a face. The shapes of the individual constituents, the eyes, nose, and mouth, in this drawing are highly schematized; it is primarily the spatial arrangement of the constituents that defines the face. In Fig. 1b, the same components are rearranged, and the figure is no longer interpreted as a face. Clearly, the recognition of objects depends not only on the presence of certain features, but also on their spatial arrangement.

The role of establishing properties and relations visually is not confined to the task of visual recognition. In the course of manipulating objects we often rely on our visual perception to obtain answers to such questions as "is *A* longer than *B*", "does *A* fit inside *B*", etc. Problems of this type can be solved without necessarily implicating object recognition. They do require, however,

Figure 1. *Schematic drawings of normally-arranged* (a) *and scrambled* (b) *faces. Figure 1a is readily recognized as representing a face although the individual features are meaningless. In 1b, the same constituents are rearranged, and the figure is no longer perceived as a face.*



the visual analysis of shape and spatial relations among parts.[1] Spatial relations in three-dimensional space therefore play an important role in visual perception.

In view of the fundamental importance of the task, it is not surprising that our visual system is indeed remarkably adept at establishing a variety of spatial relations among items in the visual input. This proficiency is evidenced by the fact that the perception of spatial properties and relations that are complex from a computational standpoint, nevertheless often appears immediate and effortless. It also appears that some of the capacity to establish spatial relations is manifested by the visual system from a very early age. For example, infants of 1–15 weeks of age are reported to respond preferentially to schematic face-like figures, and to prefer normally arranged face figures over 'scrambled' face patterns (Fantz, 1961).

The apparent immediateness and ease of perceiving spatial relations is deceiving. As we shall see, it conceals in fact a complex array of processes that have evolved to establish certain spatial relations with considerable efficiency. The processes underlying the perception of spatial relations are still unknown even in the case of simple elementary relations. Consider, for instance, the task of comparing the lengths of two line segments. Faced with this simple task, a draftsman may measure the length of the first line, record the result, measure the second line, and compare the resulting measurements. When the two lines are present simultaneously in the field of view, it is often possible to compare their lengths by 'merely looking'. This capacity raises the problem of how the 'draftsman in our head' operates, without the benefit of a ruler and a scratchpad. More generally, a theory of the perception of spatial relations should aim at unraveling the processes that take place within our visual system when we establish shape properties of objects and their spatial relations by 'merely looking' at them.

The perception of abstract shape properties and spatial relations raises fundamental difficulties with major implications for the overall processing of visual information. The purpose of this paper is to examine these problems and implications. Briefly, it will be argued that the computation of spatial relations divides the analysis of visual information into two main stages. The first is the bottom up creation of certain representations of the visible environment. Examples of such representations are the primal sketch (Marr, 1976) and the 2½-D sketch (Marr and Nishihara, 1978). The second stage involves the top–down application of visual routines to the representations constructed

[1]Shape properties (such as overall orientation, area, etc.) refer to a single item, while spatial relations (such as above, inside, longer–than, etc.) involve two or more items. For brevity, the term spatial relations used in the discussion would refer to both shape properties and spatial relations.

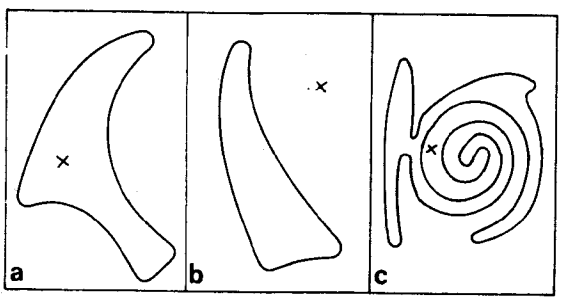
in the first stage. These routines can establish properties and relations that cannot be represented explicitly in the initial base representations. Underlying the visual routines there exists a fixed set of elemental operations that constitute the basic 'instruction set' for more complicated processes. The perception of a large variety of properties and relations is obtained by assembling appropriate routines based on this set of elemental operations.

The paper is divided into three parts. The first introduces the notion of visual routines. The second examines the role of visual routines within the overall scheme of processing visual information. The third (Sections 3 and 4) examines the elemental operations out of which visual routines are constructed.

1.2. An example: The perception of inside/outside relations

The perception of inside/outside relationships is performed by the human perceptual system with intriguing efficiency. To take a concrete example, suppose that the visual input consists of a single closed curve, and a small 'X' figure (see Fig. 2), and one is required to determine visually whether the X lies inside or outside the closed curve. The correct answers in Fig. 2a and 2b appear to be immediate and effortless, and the response would be fast and accurate.[2]

Figure 2. *Perceiving inside and outside. In 2a and 2b, the perception is immediate and effortless; in 2c, it is not.*



[2]For simple figures such as 2a, viewing time of less than 50 msec with moderate intensity, followed by effective masking is sufficient. This is well within the limit of what is considered immediate, effortless perception (e.g., Julesz, 1975). Reaction time of about 500 msec can be obtained in two-choice experiments with simple figures (Varanese, 1981). The response time may vary with the presentation conditions, but the main point is that in/out judgments are fast and reliable and require only a brief presentation.

One possible reason for our proficiency in establishing inside/outside relations is their potential value in visual recognition based on their stability with respect to the viewing position. That is, inside/outside relations tend to remain invariant over considerable variations in viewing position. When viewing a face, for instance, the eyes remain within the head boundary as long as they are visible, regardless of the viewing position (see also Sutherland (1968) on inside/outside relations in perception).

The immediate perception of the inside/outside relation is subject to some limitations (Fig. 2c). These limitations are not very restrictive, however, and the computations performed by the visual system in distinguishing 'inside' from 'outside' exhibit considerable flexibility: the curve can have a variety of shapes, and the positions of the X and the curve do not have to be known in advance.

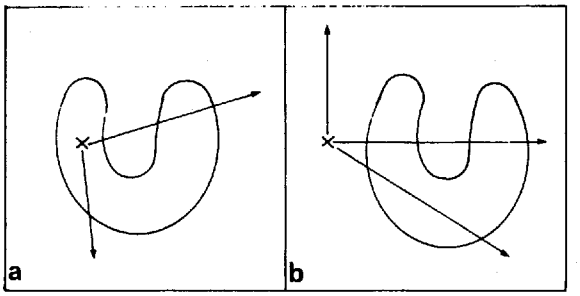
The processes underlying the perception of insde/outside relations are entirely unknown. In the following section I shall examine two methods for computing 'insideness' and compare them with human perception. The comparison will then serve to introduce the general discussion concerning the notion of visual routines and their role in visual perception.

1.2.1. Computing inside and outside

The ray-intersection method. Shape perception and recognition is often described in terms of a hierarchy of 'feature detectors' (Barlow, 1972; Milner, 1974). According to these hierarchical models, simple feature detecting units such as edge detectors are combined to produce higher order units such as, say, triangle detectors, leading eventually to the detection and recognition of objects. It does not seem possible, however, to construct an 'inside/outside detector' from a combination of elementary feature detectors. Approaches that are more procedural in nature have therefore been suggested instead. A simple procedure that can establish whether a given point lies inside or outside a closed curve is the method of ray-intersections. To use this method, a ray is drawn, emanating from the point in question, and extending to 'infinity'. For practical purposes, 'infinity' is a region that is guaranteed somehow to lie outside the curve. The number of intersections made by the ray with the curve is recorded. (The ray may also happen to be tangential to the curve without crossing it at one or more points. In this case, each tangent point is counted as two intersection points.) If the resulting intersection number is odd, the origin point of the ray lies inside the closed curve. If it is even (including zero), then it must be outside (see Fig. 3a, b).

This procedure has been implemented in computer programs (Evans, 1968; Winston, 1977, Ch. 2), and it may appear rather simple and straightforward. The success of the ray-intersection method is guaranteed, however, only if

Figure 3. *The ray intersection method for establishing inside/outside relations. When the point lies inside the closed curve, the number of intersections is odd* (a)*; when it lies outside, the number of intersections is even* (b).



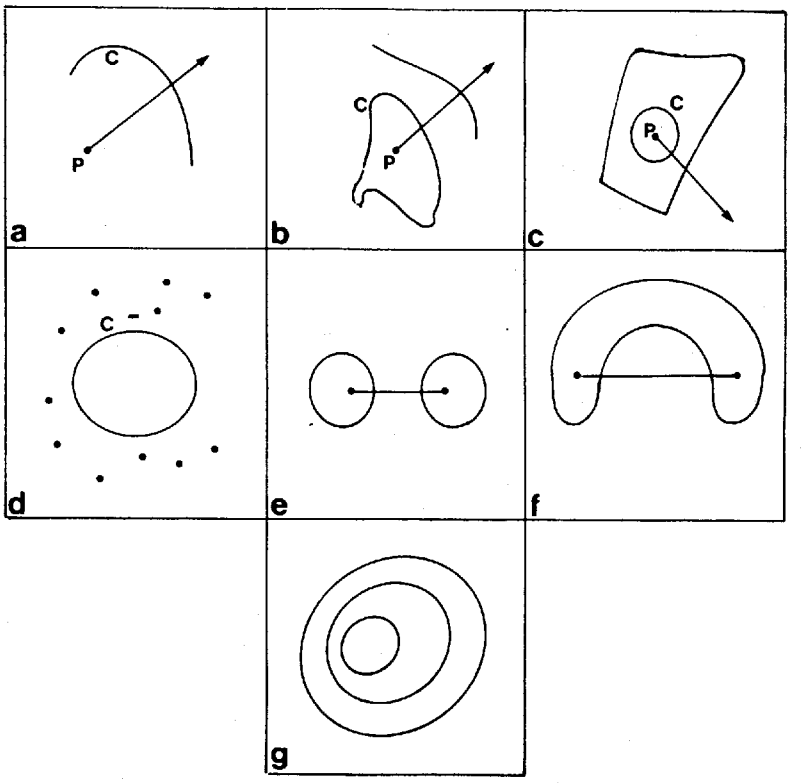
rather restrictive constraints are met. First, it must be assumed that the curve is closed, otherwise an odd number of intersections would not be indicative of an 'inside' relation (see Fig. 4a). Second, it must be assumed that the curve is isolated: in Figs. 4b and 4c, point p lies within the region bounded by the closed curve c, but the number of intersections is even.[3]

These limitations on the ray-intersection method are not shared by the human visual system: in all of the above examples the correct relation is easily established. In addition, some variations of the inside/outside problem pose almost insurmountable difficulties to the ray-intersection procedure, but not to human vision. Suppose that in Fig. 4d the problem is to determine whether any of the points lies inside the curve C. Using the ray-intersection procedure, rays must be constructed from all the points, adding significantly to the complexity of the solution. In Figs. 4e and 4f the problem is to determine whether the two points marked by dots lie inside the same curve. The number of intersections of the connecting line is not helpful in this case in establishing the desired relation. In Fig. 4g the task is to find an innermost point—a point that lies inside all of the three curves. The task is again straightforward, but it poses serious difficulties to the ray-intersection method.

It can be concluded from such considerations that the computations employed by our perceptual system are different from, and often superior to the ray-intersection method.

[3]In Fig. 4c region p can also be interpreted as lying inside a hole cut in a planar figure. Under this interpretation the result of the ray–interaction method can be accepted as correct. For the original task, however, which is to determine whether p lies within the region bounded by c, the answer provided by the ray–intersection method is incorrect.

Figure 4. *Limitations of the ray-intersection method. a, An open curve. The number of intersections is odd, but* p *does not lie inside* C. *b—c, Additional curves may change the number of intersections, leading to errors. d—g, Variations of the inside/outside problem that render the ray-intersection method in ineffective. In* d *the task is to determine visually whether any of the dots lie inside* C, *in (—f), whether the two dots lie inside the same curve; in* g *the task is to find a point that lies inside all three curves.*



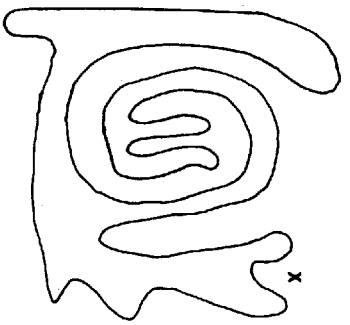
The 'coloring' method. An alternative procedure that avoids some of the limitations inherent in the ray-intersection method uses the operation of activating, or 'coloring' an area. Starting from a given point, the area around it in the internal representation is somehow activated. This activation spreads outward until a boundary is reached, but it is not allowed to cross the boundary. Depending on the starting point, either the inside or the outside of the curve, but not both, will be activated. This can provide a basis for separating inside from outside. An additional stage is still required, however, to complete the procedure, and this additional stage will depend on the specific problem at hand. One can test, for example, whether the region surrounding a 'point at infinity' has been activated. Since this point lies outside the curve in question,

it will thereby be established whether the activated area constitutes the curve's inside or the outside. In this manner a point can sometimes be determined to lie outside the curve without requiring a detailed analysis of the curve itself. In Fig. 5, most of the curve can be ignored, since activation that starts at the X will soon 'leak out' of the enclosing corridor and spread to 'infinity'. It will thus be determined that the X cannot lie inside the curve, without analyzing the curve and without attempting to separate its inside from the outside.[4]

Alternatively, one may start at an infinity point, using for instance the following procedure: (1) move towards the curve until a boundary is met; (2) mark this meeting point; (3) start to track the boundary, in a clockwise direction, activating the area on the right; (4) stop when the marked position is reached. If a termination of the curve is encountered before the marked position is reached, the curve is open and has no inside or outside. Otherwise, when the marked position is reached again and the activation spread stops, the inside of the curve will be activated. Both routines are possible, but, depending on the shape of the curve and the location of the X, one or the other may become more efficient.

The coloring method avoids some of the main difficulties with the ray-intersection method, but it also falls short of accounting for the performance of human perception in similar tasks. It seems, for example, that for human perception the computation time is to a large extent scale independent. That

Figure 5. *That the* x *does not lie inside the curve* C *can be established without a detailed analysis of the curve.*



[4]In practical applications 'infinity points' can be located if the curve is known in advance not to extend beyond a limited region. In human vision it is not clear what may constitute an 'infinity point', but it seems that we have little difficulty in finding such points. Even for a complex shape, that may not have a well-defined inside and outside, it is easy to determine visually a location that clearly lies outside the region occupied by the shape.

is, the size of the figures can be increased considerably with only a small effect on the computation time.[5] In contrast, in the activation scheme outlined above computation time should increase with the size of the figures.

The basic coloring scheme can be modified to increase its efficiency and endow it with scale independence, for example by performing the computation simultaneously at a number of resolution scales. Even the modified scheme will have difficulties, however, competing with the performance of the human perceptual system. Evidently, elaborate computations will be required to match the efficiency and flexibility exhibited by the human perceptual system in establishing inside/outside relationships.

The goal of the above discussion was not to examine the perception of inside/outside relations in detail, but to introduce the problems associated with the seemingly effortless and immediate perception of spatial relations. I next turn to a more general discussion of the difficulties associated with the perception of spatial relations and shape properties, and the implications of these difficulties to the processing of visual information.

1.3. Spatial analysis by visual routines

In this section, we shall examine the general requirements imposed by the visual analysis of shape properties and spatial relations. The difficulties involved in the analysis of spatial properties and relations are summarized below in terms of three requirements that must be faced by the 'visual processor' that performs such analysis. The three requirements are (i) the capacity to establish abstract properties and relations (abstractness), (ii) the capacity to establish a large variety of relations and properties, including newly defined ones (open-endedness), and (iii) the requirement to cope efficiently with the complexity involved in the computation of spatial relations (complexity).

1.3.1. Abstractness

The perception of inside/outside relations provides an example of the visual system's capacity to analyze abstract spatial relations. In this section the notion of abstract properties and relations and the difficulties raised by their perception will be briefly discussed.

Formally, a shape property P defines a set S of shapes that share this property. The property of closure, for example, divides the set of all curves

[5]The dependency of inside/outside judgments on the size of the figure is currently under empirical investigation. There seems to be a slight increase in reaction time as a function of the figure size.

into the set of closed curves that share this property, and the complementary set of open curves. (Similarly, a relation such as 'inside' defines a set of configurations that satisfy this relation.)

Clearly, in many cases the set of shapes S that satisfy a property P can be large and unwieldy. It therefore becomes impossible to test a shape for property P by comparing it against all the members of S stored in memory. The problem lies in fact not simply in the size of the set S, but in what may be called the size of the *support* of S. To illustrate this distinction, suppose that given a plane with one special point X marked on it we wish to identify the black figures containing X. This set of figures is large, but, given an isolated figure, it is simple to test whether it is a member of the set: only a single point, X, need be inspected. In this case the relevant part of the figure, or its support, consists of a single point. In contrast, the set of supports for the property of closure, or the inside/outside relation, is unmanageably large.

When the set of supports is small, the recognition of even a large set of objects can be accomplished by simple template matching. This means that a small number of patterns is stored, and matched against the figure in question.[6] When the set of supports is prohibitively large, a template matching decision scheme will become impossible. The classification task may nevertheless be feasible if the set contains certain regularities. This roughly means that the recognition of a property P can be broken down into a set of operations in such a manner that the overall computation required for establishing P is substantially less demanding than the storing of all the shapes in S. The set of all closed curves, for example, is not just a random collection of shapes, and there are obviously more efficient methods for establishing closure than simple template matching. For a completely random set of shapes containing no regularities, simplified recognition procedures will not be possible. The minimal program required for the recognition of the set would be in this case essentially as large as the set itself (cf. Kolmogorov, 1968).

The above discussion can now serve to define what is meant here by 'abstract' shape properties and spatial relations. This notion refers to properties and relations with a prohibitively large set of supports that can nevertheless be established efficiently by a computation that captures the regularities in the set. Our visual system can clearly establish abstract properties and

[6]For the present discussion, template–matching between plane figures can be defined as their cross–correlation. The definition can be extended to symbolic descriptions in the plane. In this case at each location in a plane a number of symbols can be activated, and a pattern is then a subset of activated symbols. Given a pattern P and a template T, their degree of match m is a function that is increasing in $P \cap T$ and decreasing in $P \cup T - P \cap T$ (when P is 'positioned over' T so as to maximize m).

relations. The implication is that it should employ sets of processes for establishing shape properties and spatial relations. The perception of abstract properties such as insideness or closure would then be explained in terms of the computations employed by the visual system to capture the regularities underlying different properties and relations. These computations would be described in terms of their constituent operations and how they are combined to establish different properties and relations.

We have seen in Section 1.2 examples of possible computations for the analysis of inside/outside relations. It is suggested that processes of this general type are performed by the human visual system in perceiving inside/outside relations. The operations employed by the visual system may prove, however, to be different from those considered in Section 1.2. To explain the perception of inside/outside relations it would be necessary, therefore, to unravel the constituent operations employed by the visual system, and how they are used in different judgments.

1.3.2. Open-endedness

As we have seen, the perception of an abstract relation is quite a remarkable feat even for a single relation, such as insideness. Additional complications arise from the requirement to recognize not only one, but a large number of different properties and relations. A reasonable approach to the problem would be to assume that the computations that establish different properties and relations share their underlying elemental operations. In this manner a large variety of abstract shape properties and spatial relations can be established by different processes assembled from a fixed set of elemental operations. The term 'visual routines' will be used to refer to the processes composed out of the set of elemental operations to establish shape properties and spatial relations.

A further implication of the open-endedness requirement is that a mechanism is required by which new combinations of basic operations can be assembled to meet new computational goals. One can impose goals for visual analysis, such as "determine whether the green and red elements lie on the same side of the vertical line". That the visual system can cope effectively with such goals suggests that it has the capacity to create new processes out of the basic set of elemental operations.

1.3.3. Complexity

The open-endedness requirement implied that different processes should share elemental operations. The same conclusion is also suggested by complexity considerations. The complexity of basic operations such as the bounded activation (discussed in more detal in Section 3.4) implies that differ-

ent routines that establish different properties and relations and use the bounded activation operation would have to share the same mechanism rather than have their own separate mechanisms.

A special case of the complexity consideration arises from the need to apply the same computation at different spatial locations. The ability to perform a given computation at different spatial positions can be obtained by having an independent processing module at each location. For example, the orientation of a line segment at a given location seems to be performed in the primary visual cortex largely independent of other locations. In contrast, the computations of more complex relations such as inside/outside independent of location cannot be explained by assuming a large number of independent 'inside/outside modules', one for each location. Routines that establish a given property or relation at different positions are likely to share some of their machinery, similar to the sharing of elemental operations by different routines.

Certain constraints will be imposed upon the computation of spatial relations by the sharing of elemental operations. For example, the sharing of operations by different routines will restrict the simultaneous perception of different spatial relations. The application of a given routine to different spatial locations will be similarly restricted. In applying visual routines the need will consequently arise for the sequencing of elemental operations, and for selecting the location at which a given operation is applied.

In summary, the three requirements discussed above suggest the following implications.

(1) Spatial properties and relations are established by the application of visual routines to a set of early visual representations.
(2) Visual routines are assembled from a fixed set of elemental operations.
(3) New routines can be assembled to meet newly specified processing goals.
(4) Different routines share elemental operations.
(5) A routine can be applied to different spatial locations. The processes that perform the same routine at different locations are not independent.
(6) In applying visual routines mechanisms are required for sequencing elemental operations and for selecting the locations at which they are applied.

1.4. Conclusions and open problems

The discussion so far suggests that the immediate perception of seemingly simple spatial relations requires in fact complex computations that are difficult to unravel, and difficult to imitate. These computations were termed above 'visual routines'. The general proposal is that using a fixed set of basic operations, the visual system can assemble routines that are applied to the visual representations to extract abstract shape properties and spatial relations.

The use of visual routines to establish shape properties and spatial relations raises fundamental problems at the levels of computational theory, algorithms, and the underlying mechanisms. A general problem on the computational level is which spatial properties and relations are important for object recognition and manipulation. On the algorithmic level, the problem is how these relations are computed. This is a challenging problem, since the processing of spatial relations and properties by the visual system is remarkably flexible and efficient. On the mechanism level, the problem is how visual routines are implemented in neural networks within the visual system.

In concluding this section, major problems raised by the notion of visual routines are listed below under four main categories.

(1) *The elemental operations*. In the examples discussed above the computation of inside/outside relations employed operations such as drawing a ray, counting intersections, boundary tracking, and area activation. The same basic operations can also be used in establishing other properties and relations. In this manner a variety of spatial relations can be computed using a fixed and powerful set of basic operations, together with means for combining them into different routines that are then applied to the base representation. The first problem that arises therefore is the identification of the elemental operations that constitute the basic 'instruction set' in the composition of visual routines.

(2) *Integration*. The second problem that arises is how the elemental operations are integrated into meaningful routines. This problem has two aspects. First, the general principles of the integration process, for example, whether different elemental operations can be applied simultaneously. Second, there is the question of how specific routines are composed in terms of the elemental operations. An account of our perception of a given shape property or relation such as elongation, above, next-to, inside/outside, taller-than etc. should include a description of the routines that are employed in the task in question, and the composition of each of these routines in terms of the elemental operations.

(3) *Control*. The questions in this category are how visual routines are

selected and controlled, for example, what triggers the execution of different routines during visual recognition and other visual tasks, and how the order of their execution is determined.

(4) *Compilation*. How new routines are generated to meet specific needs, and how they are stored and modified with practice.

The remainder of this paper is organized as follows. In Section 2 I shall discuss the role of visual routines within the overall processing of visual information. Section 3 will then examine the first of the problems listed above, the elemental operations problem. Section 4 will conclude with a few brief comments pertaining to the other problems.

2. Visual routines and their role in the processing of visual information

The purpose of this section is to examine how the application of visual routines fits within the overall processing of visual information. The main goal is to elaborate the relations between the initial creation of the early visual representations and the subsequent application of visual routines. The discussion is structured along the following lines.

The first half of this section examines the relation between visual routines and the creation of two types of visual representations: the bare representation (Section 2.1) that precedes the application of visual routines, and the incremental representations that are produced by them (Section 2.2). The second half examines two general problems raised by the nature of visual routines as described in the first half. These problems are: the initial selection of routines (Section 2.3) and the parallel processing of visual information (Section 2.4).

2.1. Visual routines and the base representations

In the scheme suggested above, the processing of visual information can be divided into two main stages. The first is the 'bottom-up' creation of some base representations by the early visual processes (Marr, 1980). The second stage is the application of visual routines. At this stage, procedures are applied to the base representations to define distinct entities within these representations, establish their shape properties, and extract spatial relations among them. In this section we shall examine more closely the distinction between these two stages.

2.1.1. The base representations

The first stage in the analysis of visual information can usefully be described as the creation of certain representations to be used by subsequent visual processes. Marr (1976) and Marr and Nishihara (1978) have suggested a division of these early representations into two types: the primal sketch, which is a representation of the incoming image, and the 2½-D sketch, which is a representation of the visible surfaces in three-dimensional space. The early visual representations share a number of fundamental characteristics: they are unarticulated, viewer-centered, uniform, and bottom-up driven. By 'unarticulated' I mean that they are essentially local descriptions that represent properties such as depth, orientation, color, and direction of motion at a point. The definition of larger more complicated units, and the extraction and description of spatial relationships among their parts, is not achieved at this level.

The base representations are spatially uniform in the sense that, with the exception of a scaling factor, the same properties are extracted and represented across the visual field (or throughout large parts of it). The descriptions of different points (e.g., the depth at a point) in the early representations are all with respect to the viewer, not with respect to one another. Finally, the construction of the base representations proceeds in a bottom-up fashion. This means that the base representations depend on the visual input alone.[7] If the same image is viewed twice, at two different times, the base representations associated with it will be identical.

2.1.2. Applying visual routines to the base representations

Beyond the construction of the base representations, the processing of visual information requires the definition of objects and parts in the scene, and the analysis of spatial properties and relations. The discussion in Section 1.3 concluded that for these tasks the uniform bottom-up computation is no longer possible, and suggested instead the application of visual routines. In contrast with the construction of the base representations, the properties and relations to be extracted are not determined by the input alone: for the same visual input different aspects will be made explicit at different times, depend-

[7]Although 'bottom–up' and 'top–down' processing are useful and frequently used terms, they lack a precise, well–accepted definition. As mentioned in the text, the definition I adopt is that bottom–up processing is determined entirely by the input. Top–down processing depends on additional factors, such as the goal of the computation (but not necessarily on object–specific knowledge).

Physiologically, various mechanisms that are likely to be involved in the creation of the base representation appear to be bottom–up: their responses can be predicted from the parameters of the stimulus alone. They also show strong similarity in their responses in the awake, anesthetized, and naturally sleeping animal (e.g., Livingstone and Hubel, 1981).

ing on the goals of the computation. Unlike the base representations, the computations by visual routines are not applied uniformly over the visual field (e.g., not all of the possible inside/outside relations in the scene are computed), but only to selected objects. The objects and parts to which these computations apply are also not determined uniquely by the input alone; that is, there does not seem to be a universal set of primitive elements and relations that can be used for all possible perceptual tasks. The definition of objects and distinct parts in the input, and the relations to be computed among them may change with the situation. I may recognize a particular cat, for instance, using the shape of the white patch on its forehead. This does not imply, however, that the shapes of all the white patches in every possible scene and all the spatial relations in which such patches participate are universally made explicit in some internal representation. More generally, the definition of what constitutes a distinct part, and the relations to be established often depends on the particular object to be recognized. It is therefore unlikely that a fixed set of operations applied uniformly over the base representations would be sufficient to capture all of the properties and relations that may be relevant for subsequent visual analysis.[8] A final distinction between the two stages is that the construction of the base representations is fixed and unchanging, while visual routines are open-ended and permit the extraction of newly defined properties and relations.

In conclusion, it is suggested that the analysis of visual information divides naturally into two distinct successive stages: the creation of the base representations, followed by the application of visual routines to these representations. The application of visual routines can define objects within the base representations and establish properties and spatial relations that cannot be established within the base representations.

It should be noted that many of the relations that are established at this stage are defined not in the image but in three-dimensional space. Since the base representations already contain three-dimensional information, the visual routines applied to them can also establish properties and relations in three-dimensional space.[9]

---

[8]The argument does not preclude the possibility that some grouping processes that help to define distinct parts and some local shape descriptions take place within the basic representations.

[9]Many spatial judgments we make depend primarily on three dimensional relations rather than on projected, two–dimensional ones (see e.g., Joynson and Kirk, 1960; Kappin and Fuqua, 1983). The implication is that various visual routines such as those used in comparing distances, operate upon a three–dimensional representation, rather than a representation that resembles the two–dimensional image.

2.2. The incremental representations

The creation of visual representations does not stop at the base representations. It is reasonable to expect that results established by visual routines are retained temporarily for further use. This means that in addition to the base representations to which routines are applied initially representations are also being created and modified in the course of executing visual routines. I shall refer to these additional structures as 'incremental representations', since their content is modified incrementally in the course of applying visual routines. Unlike the base representations, the incremental representations are not created in a uniform and unguided manner: the same input can give rise to different incremental representations, depending on the routines that have been applied.

The role of the incremental representations can be illustrated using the inside/outside judgments considered in Section 1. Suppose that following the response to an inside/outside display using a fairly complex figure, an additional point is lit up. The task is now to determine whether this second point lies inside or outside the closed figure. If the results of previous computations are already summarized in the incremental representation of the figure in question, the judgment in the second task would be expected to be considerably faster than the first, and the effects of the figure's complexity might be reduced.[10] Such facilitation effects would provide evidence for the creation of some internal structure in the course of reaching a decision in the first task that is subsequently used to reach a faster decision in the second task. For example, if area activation or 'coloring' is used to separate inside from outside, then following the first task the inside of the figure may be already 'colored'. If, in addition, this coloring is preserved in the incremental representation, then subsequent inside/outside judgments with respect to the same figure would require considerably less processing, and may depend less on the complexity of the figure.

This example also serves to illustrate the distinction between the base representations and the incremental representations. The 'coloring' of the curve in question will depend on the particular routines that happened to be employed. Given the same visual input but a different visual task, or the same task but applied to a different part of the input, the same curve will not be 'colored' and a similar saving in computation time will not be obtained. The general point illustrated by this example is that for a given visual stimulus but different computational goals the base representations remain the same,

[10]This example is due to Steve Kosslyn. It is currently under empirical investigations.

while the incremental representations would vary.

Various other perceptual phenomena can be interpreted in a similar manner in light of the distinction between the base and the incremental representations. I shall mention here only one recent example from a study by Rock and Gutman (1981). Their subjects were presented with pairs of overlapping red and green figures. When they were instructed to attend selectively to the green or red member of the pair, they were later able to recognize the 'attended' but not the 'unattended' figure. This result can be interpreted in terms of the distinction between the base and the incremental representations. The creation of the base representations is assumed to be a bottom-up process, unaffected by the goal of the computation. Consequently, the two figures would not be treated differently within these representations. Attempts to attend selectively to one sub-figure resulted in visual routines being applied preferentially to it. A detailed description of this sub-figure is consequently created in the incremental representations. This detailed description can then be used by subsequent routines subserving comparison and recognition tasks.

The creation and use of incremental representations imply that visual routines should not be thought of merely as predicates, or decision processes that supply 'yes' or 'no' answers. For example, an inside/outside routine does not merely signal 'yes' if an inside relation is established, and 'no' otherwise. In addition to the decision process, certain structures are being created during the execution of the routine. These structures are maintained in the incremental representation, and can be used in subsequent visual tasks. The study of a given routine is therefore not confined to the problem of how a certain decision is reached, but also includes the structures constructed by the routine in question in the incremental representations.

In summary, the use of visual routines introduces a distinction between two different types of visual representations: the base representations and incremental representations. The base representations provide the initial data structures on which the routines operate, and the incremental representations maintain results obtained by the application of visual routines.

The second half of Section 2 examines two general issues raised by the nature of visual routines as introduced so far. Visual routines were described above as sequences of elementary operations that are assembled to meet specific computational goals. A major problem that arises is the initial selection of routines to be applied. This problem is examined briefly in Section 2.3. Finally, sequential application of elementary operations seems to stand in contrast with the notion of parallel processing in visual perception. (Biederman *et al.*, 1973; Donderi and Zelnicker, 1969; Egeth *et al.*, 1972; Jonides and Gleitman, 1972; Neisser *et al.*, 1963). Section 2.4 examines the distinction

between sequential and parallel processing, its significance to the processing of visual information, and its relation to visual routines.

2.3. Universal routines and the initial access problem

The act of perception requires more than the passive existence of a set of representations. Beyond the creation of the base representations, the perceptual process depends upon the current computational goal. At the level of applying visual routines, the perceptual activity is required to provide answers to queries, generated either externally or internally, such as: "is this my cat?" or, at a lower level, "is *A* longer than *B*?" Such queries arise naturally in the course of using visual information in recognition, manipulation, and more abstract visual thinking. In response to these queries routines are executed to provide the answers. The process of applying the appropriate routines is apparently efficient and smooth, thereby contributing to the impression that we perceive the entire image at a glance, when in fact we process only limited aspects of it at any given time. We may not be aware of the restricted processing since whenever we wish to establish new facts about the scene, that is, whenever an internal query is posed, an answer is provided by the execution of an appropriate routine.

Such application of visual routines raises the problem of guiding the perceptual activity and selecting the appropriate routines at any given instant. In dealing with this problem, several theories of perception have used the notion of schemata (Bartlett, 1932; Biederman *et al.*, 1973; Neisser, 1967) or frames (Minsky, 1975) to emphasize the role of expectations in guiding perceptual activity. According to these theories, at any given instant we maintain detailed expectations regarding the objects in view. Our perceptual activity can be viewed according to such theories as hypothesizing a specific object and then using detailed prior knowledge about this object in an attempt to confirm or refute the current hypothesis.

The emphasis on detailed expectations does not seem to me to provide a satisfactory answer to the problem of guiding perceptual activity and selecting the appropriate routines. Consider for example the 'slide show' situation in which an observer is presented with a sequence of unrelated pictures flashed briefly on a screen. The sequence may contain arbitrary ordinary objects, say, a horse, a beachball, a printed letter, etc. Although the observer can have no expectations regarding the next picture in the sequence, he will experience little difficulty identifying the viewed objects. Furthermore, suppose that an observer does have some clear expectations, e.g., he opens a door expecting to find his familiar office, but finds an ocean beach instead. The contradiction to the expected scene will surely cause a surprise, but no

major perceptual difficulties. Although expectations can under some conditions facilitate perceptual processes significantly (e.g. Potter, 1975), their role is not indispensable. Perception can usually proceed in the absence of prior specific expectations and even when expectations are contradicted.

The selection of appropriate routines therefore raises a difficult problem. On the one hand, routines that establish properties and relations are situation-dependent. For example, the white patch on the cat's forehead is analyzed in the course of recognizing the cat, but white patches are not analyzed invariably in every scene. On the other hand, the recognition process should not depend entirely on prior knowledge or detailed expectations about the scene being viewed. How then are the appropriate routines selected?

It seems to me that this problem can be best approached by dividing the process of routine selection into two stages. The first stage is the application of what may be called *universal routines*. These are routines that can be usefully applied to any scene to provide some initial analysis. They may be able, for instance, to isolate some prominent parts in the scene and describe, perhaps crudely, some general aspects of their shape, motion, color, the spatial relations among them etc. These universal routines will provide sufficient information to allow initial indexing to a recognition memory, which then serves to guide the application of more specialized routines.

To make the motion of universal routines more concrete, I shall cite one example in which universal routines probably play a role. Studying the comparison of shapes presented sequentially, Rock *et al.* (1972) found that some parts of the presented shapes can be compared reliably while others cannot. When a shape was composed, for example, of a bounding contour and internal lines, in the absence of any specific instructions only the bounding contour was used reliably in the successive comparison task, even if the first figure was viewed for a long period (5 sec). This result would be surprising if only the base representations were used in the comparison task, since there is no reason to assume that in these representations the bounding contours of such line drawings enjoy a special status. It seems reasonable, however, that the bounding contour is special from the point of view of the universal routines, and is therefore analyzed first. If successive comparisons use the incremental representation as suggested above, then performance would be superior on those parts that have been already analyzed by visual routines. It is suggested, therefore, that in the absence of specific instructions, universal routines were applied first to the bounding contour. Furthermore, it appears that in the absence of specific goals, no detailed descriptions of the entire figure are generated even under long viewing periods. Only those aspects analyzed by the universal routines are summarized in the incremental representation. As

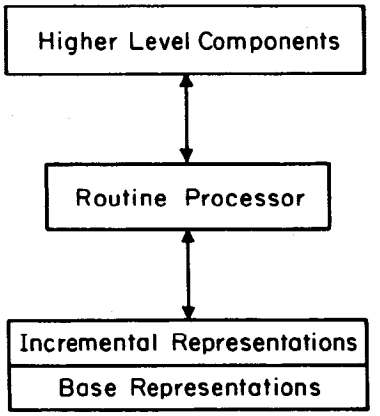
a result, a description of the outside boundary alone has been created in the incremental representation. This description could then be compared against the second figure. It is of interest to note that the description generated in this task appears to be not just a coarse structural description of the figure, but has template-like quality that enable fine judgments of shape similarity.

These results can be contrasted with the study mentioned earlier by Rock and Gutman (1981) using pairs of overlapping and green figures. When subjects were instructed to "attend" selectively to one of the subfigures, they were subsequently able to make reliable shape comparisons to this, but not the other, subfigure. Specific requirements can therefore bias the selection and application of visual routines. Universal routines are meant to fill the void when no specific requirements are set. They are intended to acquire sufficient information to then determine the application of more specific routines.

For such a scheme to be of value in visual recognition, two interrelated requirements must be met. The first is that with universal routines alone it should be possible to gather sufficiently useful information to allow initial classification. The second requirement has to do with the organization of the memory used in visual recognition. It should contain intermediate constructs of categories that are accessible using the information gathered by the universal routines, and the access to such a category should provide the means for selecting specialized routines for refining the recognition process. The first requirement raises the question of whether universal routines, unaided by specific knowledge regarding the viewed objects, can reasonably be expected to supply sufficiently useful information about any viewed scene. The question is difficult to address in detail, since it is intimately related to problems regarding the structure of the memory used in visual recognition. It nonetheless seems plausible that universal routines may be sufficient to analyze the scene in enough detail to allow the application of specialized routines.

The potential usefulness of universal routines in the initial phases of the recognition process is supported in part by Marr and Nishihara's (1978) study of shape recognition. This work has demonstrated that at least for certain classes of shapes crude overall shape descriptions, which can be obtained by universal routines without prior knowledge regarding the viewed objects, can provide a powerful initial categorization. Similarly, the "perceptual 20 question game" of W. Richards (1982) suggests that a small fixed set of visual attributes (such as direction and type of motion, color, etc.) is often sufficient to form a good idea of what the object is (e.g., a walking person) although identifying a specific object (e.g., who the person is) may be considerably more difficult [cf. Milner, 1974]. These examples serve to illustrate the dis-

Figure 6. *The routine processor acts as an intermediary between the visual representations and higher level components of the system.*



tinction in visual recognition between universal and specific stages. In the first, universal routines can supply sufficient information for accessing a useful general category. In the second, specific routines associated with this category can be applied.

The relations between the different representations and routines can now be summarized as follows. The first stage in the analysis of the incoming visual input is the creation of the base representations. Next, visual routines are applied to the base representations. In the absence of specific expectations or prior knowledge universal routines are applied first, followed by the selective application of specific routines. Intermediate results obtained by visual routines are summarized in the incremental representation and can be used by subsequent routines.

2.3.1. Routines as intermediary between the base representations and higher-level components

The general role of visual routines in the overall processing of visual information as discussed so far is illustrated schematically in Fig. 6. The processes that assemble and execute visual routines (the 'routines processor' module in the figure) serve as an intermediary between the visual representations and higher level components of the system, such as recognition memory. Communication required between the higher level components and the visual representations for the analysis of shape and spatial relations are channeled via the routine processor.[11]

[11]Responses to certain visual stimuli that do not require abstract spatial analysis could bypass the routine processor. For example, a looming object may initiate an immediate avoidance response (Regan and Beverly, 1978). Such 'visual reflexes' do not require the application of visual routines. The visual system of lower animals such as insects or the frog, although remarkably sophisticated, probably lack routine mechanisms, and can perhaps be described as collections of 'visual reflexes'.

Visual routines operate in the middle ground that, unlike the bottom-up creation of the base representations, is a part of the top-down processing and yet is independent of object-specific knowledge. Their study therefore has the advantage of going beyond the base representations while avoiding many of the additional complications associated with higher level components of the system. The recognition of familiar objects, for example, often requires the use of knowledge specific to these objects. What we know about telephones or elephants can enter into the recognition process of these objects. In contrast, the extraction of spatial relations, while important for object recognition, is independent of object-specific knowledge. Such knowledge can determine the routine to be applied: the recognition of a particular object may require, for instance, the application of inside/outside routines. When a routine is applied, however, the processing is no longer dependent on object-specific knowledge.

It is suggested, therefore, that in studying the processing of visual information beyond the creation of the early representations, a useful distinction can be drawn between two problem areas. One can approach first the study of visual routines almost independently of the higher level components of the system. A full understanding of problems such as visually guided manipulation and object recognition would require, in addition, the study of higher level components, how they determine the application of visual routines, and how they are affected by the results of applying visual routines.

2.4. Routines and the parallel processing of visual information

A popular controversy in theories of visual perception is whether the processing of visual information proceeds in parallel or sequentially. Since visual routines are composed of sequences of elementary operations, they may seem to side strongly with the point of view of sequential processing in perception. In this section I shall examine two related questions that bear on this issue. First, whether the application of visual routines implies sequential processing. Second, what is the significance of the distinction between the parallel and sequential processing of visual information.

2.4.1. Three types of parallelism

The notion of processing visual information 'in parallel' does not have a unique, well-defined meaning. At least three types of parallelism can be distinguished in this processing: spatial, functional, and temporal. Spatial parallelism means that the same or similar operations are applied simultaneously to different spatial locations. The operations performed by the retina and the primary visual cortext, for example, fall under this category. Functional parallelism means that different computations are applied simultane-

ously to the same location. Current views of the visual cortex (e.g., Zeki, 1978a, b) suggest that different visual areas in the extra-striate cortex process different aspects of the input (such as color, motion, and stereoscopic disparity) at the same location simultaneously, thereby achieving functional parallelism.[12] Temporal parallelism is the simultaneous application of different processing stages to different inputs (this type of parallelism is also called 'pipelining'.[13]

Visual routines can in principle employ all three types of parallelism. Suppose that a given routine is composed of a sequence of operations O_1, O_2, ... O_n. Spatial parallelism can be obtained if a given operation O_i is applied simultaneously to various locations. Temporal parallelism can be obtained by applying different operations O_i simultaneously to successive inputs. Finally, functional parallelism can be obtained by the concurrent application of different routines.

The application of visual routines is thus compatible in principle with all three notions of parallelism. It seems, however, that in visual routines the use of spatial parallelism is more restricted than in the construction of the base representations.[14] At least some of the basic operations do not employ extensive spatial parallelism. The internal tracking of a discontinuity boundary in the base representation, for instance, is sequential in nature and does not apply to all locations simultaneously. Possible reasons for the limited spatial parallelism in visual routines are discussed in the next section.

2.4.2. Essential and non-essential sequential processing

When considering sequential *versus* spatially parallel processing, it is useful to distinguish between essential and non-essential sequentially. Suppose, for example, that O_1 and O_2 are two independent operations that can, in principle, be applied simultaneously. It is nevertheless still possible to apply them in sequence, but such sequentiality would be non-essential. The total computation required in this case will be the same regardless of whether the operations are performed in parallel or sequentially. Essential sequentiality, on the other hand, arises when the nature of the task makes parallel processing impossible or highly wasteful in terms of the overall computation required.

---

[12]Disagreements exist regarding this view, in particular, the role of area V4 in the rhesus monkey i processing color (Schein *et al.*, 1982). Although the notion o "one cortical area for each function" is too simplistic, the physiological data support in general the notion of functional parallelism.

[13]Suppose that a sequence of operations O_1, O_2 ... O_k is applied to each input in a temporal sequence I_1. I_2, I_3 .... First, O_1 is applied to I_1. Next, as O_2 is applied to I_1, O_1 can be applied to I_2. In general O_i, $1 < i < k$ can be applied simultaneously to I_{n-i}. Such a simultaneous application constitute temporal parallelism.

[14]The general notion of an extensively parallel stage followed by a more sequential one is in agreement with various findings and theories of visual perception (e.g., Estes, 1972; Neisser, 1967; Shiffrin *et al.*, 1976).

Problems pertaining to the use of spatial parallelism in the computation of spatial properties and relations were studied extensively by Minsky and Papert (1969) within the perceptrons model.[15] Minsky and Paper have established that certain relations, including the inside/outside relation, cannot be computed at all in parallel by any diamater-limited or order-limited perceptrons. This limitation does not seem to depend critically upon the perceptron-like decision scheme. It may be conjectured, therefore, that certain relations are inherently sequential in the sense that it is impossible or highly wasteful to employ extensive spatial parallelism in their computation. In this case sequentiality is essential, as it is imposed by the nature of the task, not by particular properties of the underlying mechanisms. Essential sequentiality is theoretically more interesting, and has more significant ramifications, than non-essential sequential ordering. In non-essential sequential processing the ordering has no particular importance, and no fundamentally new problems are introduced. Essential sequentiality, on the other hand, requires mechanisms for controlling the appropriate sequencing of the computation.

It has been suggested by various theories of visual attention that sequential ordering in perception is non-essential, arising primarily from a capacity limitation of the system (see, e.g., Holtzman and Gazzaniga, 1982; Kahneman, 1973; Rumelhart, 1970). In this view only a limited region of the visual scene (1 degree, Eriksen and Hoffman, 1972; see also Humphreys, 1981; Mackworth, 1965) is processed at any given time because the system is capacity-limited and would be overloaded by excessive information unless a spatial restruction is employed. The discussion above suggests, in contrast, that sequential ordering may in fact be essential, imposed by the inherently sequential nature of various visual tasks. This sequential ordering has substantial implications since it requires perceptual mechanisms for directing the processing and for concatenating and controlling sequences of basic operations.

Although the elemental operations are sequenced, some of them, such as the bounded activation, employ spatial parallelism and are not confined to a limited region. This spatial parallelism plays an important role in the inside/outside routines. To appreciate the difficulties in computing inside/outside relations without the benefit of spatial parallelism, consider solving a tactile

---

[15]In the perceptron scheme the computation is performed in parallel by a large number of units ϕ_i. Each unit examine a restricted part of the 'retina' R. In a diamater–limited perceptron, for instance, the region examined by each unit is restricted to lie within a circle whose diameter is small compared to the size of R. The computation performed by each unit is a predicate of its inputs (i.e., $\phi_i = 0$ or $\phi_i = 1$). For example, a unit may be a 'corner detector' at a particular location, signalling 1 in the presence of a corner and 0 otherwise. All the local units then feed a final decision stage, assumed to be a linear threshold device. That is, it tests whether the weighted sum of the inputs $\Sigma_i \omega_i \phi_i$ exceeds a predetermined threshold θ.

version of the same problem by moving a cane or a fingertip over a relief surface. Clearly, when the processing is always limited to a small region of space, the task becomes considerably more difficult. Spatial parallelism must therefore play an important role in visual routines.

In summary, visual routines are compatible in principle with spatial, temporal, and functional parallelism. The degree of spatial parallelism employed by the basic operations seems nevertheless limited. It is conjectured that this reflects primarily essential sequentiality, imposed by the nature of the computations.

3. The elemental operations

3.1. Methodological considerations

In this section, we examine the set of basic operations that may be used in the construction of visual routines. In trying to explore this set of internal operations, at least two types of approaches can be followed. The first is the use of empirical psychological and physiological evidence. The second is computational: one can examine, for instance, the types of basic operations that would be useful in principle for establishing a large variety of relevant properties and relations. In particular, it would be useful to examine complex tasks in which we exhibit a high degree of proficiency. For such tasks, processes that match in performance the human system are difficult to devise. Consequently, their examination is likely to provide useful constraints on the nature of the underlying computations.

In exploring such tasks, the examples I shall use will employ schematic drawings rather than natural scenes. The reason is that simplified artificial stimuli allow more flexibility in adapting the stimulus to the operation under investigation. It seems to me that insofar as we examine visual tasks for which our proficiency is difficult to account for, we are likely to be exploring useful basic operations even if the stimuli employed are artificially constructed. In fact, this ability to cope efficiently with artificially imposed visual tasks underscores two essential capacities in the computation of spatial relations. First, that the computation of spatial relations is flexible and open-ended: new relations can be defined and computed efficiently. Second, it demonstrates our capacity to accept non-visual specification of a task and immediately produce a visual routine to meet these specifications.

The empirical and computational studies can then be combined. For example, the complexity of various visual tasks can be compared. That is, the theoretical studies can be used to predict how different tasks should vary in

complexity, and the predicted complexity measure can be gauged against human performance. We have seen in Section 1.2 an example along this line, in the discussion of the inside/outside computation. Predictions regarding relative complexity, success, and failure, based upon the ray-intersection method prove largely incompatible with human performance, and consequently the employment of this method by the human perceptual system can be ruled out. In this case, the refutation is also supported by theoretical considerations exposing the inherent limitations of the ray-intersection method.

In this section, only some initial steps towards examining the basic operations problem will be taken. I shall examine a number of plausible candidates for basic operations, discuss the available evidence, and raise problems for further study. Only a few operations will be examined; they are not intended to form a comprehensive list. Since the available empirical evidence is scant, the emphasis will be on computational considerations of usefulness. Finally, some of the problems associated with the assembly of basic operations into visual routines will be briefly discussed.

3.2. Shifting the processing focus

A fundamental requirement for the execution of visual routines is the capacity to control the location at which certain operations take place. For example, the operation of area activation suggested in Section 1.2 will be of little use if the activation starts simultaneously everywhere. To be of use, it must start at a selected location, or along a selected contour. More generally, in applying visual routines it would be useful to have a 'directing mechanism' that will allow the application of the same operation at different spatial locations. It is natural, therefore, to start the discussion of the elemental operations by examining the processes that control the locations at which these operations are applied.

Directing the processing focus (that is, the location to which an operation is applied) may be achieved in part by moving the eyes (Noton and Stark, 1971). But this is clearly insufficient: many relations, including, for instance, the inside/outside relation examined in Section 1.2, can be established without eye movements. A capacity to shift the processing focus internally is therefore required.

Problems related to the possible shift of internal operations have been studied empirically, both psychophysically and physiologically. These diverse studies still do not provide a complete picture of the shift operations and their use in the analysis of visual information. They do provide, however, strong support for the notion that shifts of the processing focus play an important

role in visual information processing, starting from early processing stages. The main directions of studies that have been pursued are reviewed briefly in the next two sections.

3.2.1. Psychological evidence

A number of psychological studies have suggested that the focus of visual processing can be directed, either voluntarily or by manipulating the visual stimulus, to different spatial location in the visual input. They are listed below under three main categories.

The first line of evidence comes from reaction time studies suggesting that it takes some measurable time to shift the processing focus from one location to another. In a study by Eriksen and Schultz (1977), for instance, it was found that the time required to identify a letter increased linearly with the eccentricity of the target letter, the difference being on the order of 100 msec at 3° from the fovea center. Such a result may reflect the effect of shift time, but, as pointed out by Eriksen and Schultz, alternative explanations are possible.

More direct evidence comes from a study by Posner *et al.* (1978). In this study a target was presented seven degrees to the left or right of fixation. It was shown that if the subjects correctly anticipated the location at which the target will appear using prior cueing (an arrow at fixation), then their reaction time to the target in both detection and identification tasks were consistently lower (without eye movements). For simple detection tasks, the gain in detection time for a target at 70 eccentricity was on the order of 30 msec.

A related study by Tsal (1983) employed peripheral rather than central cueing. In his study a target letter could appear at different eccentricities, preceded by a brief presentation of a dot at the same location. The results were consistent with the assumption that the dot initiated a shift towards the cued location. If a shift to the location of the letter is required for its identification, the cue should reduce the time between the letter presentation and its identification. If the cue precedes the target letter by k msec, then by the time the letter appears the shift operation is already k msec under way, and the response time should decrease by this amount. The facilitation should therefore increase linearly with the temporal delay between the cue and target until the delay equals the total shift time. Further increase of the delay should have no additional effect. This is exactly what the experimental results indicated. It was further found that the delay at which facilitation saturates (presumably the total shift time) increases with eccentricity, by about 8 msec on the average per 1° of visual angle.

A second line of evidence comes from experiments suggesting that visual sensitivity at different locations can be somewhat modified with a fixed eye

position. Experiments by Shulman *et al.* (1979) can be interpreted as indicating that a region of somewhat increased sensitivity can be shifted across the visual field. A related experiment by Remington (1978, described in Posner, 1980), showed an increase in sensitivity at a distance of 8° from the fixation point 50–100 msec after the location has been cued.

A third line of evidence that may bear on the internal shift operations comes from experiments exploring the selective readout from some form of short term visual memory (e.g., Shiffrin *et al.*, 1976; Sperling, 1960). These experiments suggest that some internal scanning can be directed to different locations a short time after the presentation of a visual stimulus.

The shift operation and selective visual attention. Many of the experiments mentioned above were aimed at exploring the concept of 'selective attention'. This concept has a variety of meanings and connotations (cf. Estes, 1972), many of which are not related directly to the proposed shift of processing focus in visual routines. The notion of selective visual attention often implies that the processing of visual information is restricted to small region of space, to avoid 'overloading' the system with excessive information. Certain processing stages have, according to this description, a limited total 'capacity' to invest in the processing, and this capacity can be concentrated in a spatially restricted region. Attempts to process additional information would detract from this capacity, causing interference effects and deterioration of performance. Processes that do not draw upon this general capacity are, by definition, pre-attentive. In contrast, the notion of processing shift discussed above stems from the need for spatially-structured processes, and it does not necessarily imply such notions as general capacity or protection from overload. For example, the 'coloring' operation used in Section 1.2 for separating inside from outside started from a selected point or contour. Even with no capacity limitations such coloring would not start simultaneously everywhere, since a simultaneous activation will defy the purpose of the coloring operation. The main problem in this case is in coordinating the process, rather than excessive capacity demands. As a result, the process is spatially structured, but not in a simple manner as in the 'spotlight model' of selective attention. In the course of applying a visual routine, both the locations and the operations performed at the selected locations are controlled and coordinated according to the requirement of the routine in question.

Many of the results mentioned above are nevertheless in agreement with the possible existence of a directable processing focus. They suggest that the redirection of the processing focus to a new location may be achieved in two ways. The experiments of Posner and Shulman *et al.* suggest that it can be 'programmed' to move along a straight path using central cueing. In other

experiments, such as Remmington's and Tsal's, the processing focus is shifted by being attracted to a peripheral cue.

3.2.2. Physiological evidence

Shift-related mechanisms have been explored physiologically in the monkey in a number of different visual areas: the superior colliculus, the posterior parietal lobe (area 7) the frontal eye fields, areas V1, V2, V4, MT, MST, and the inferior temporal lobe.

In the superficial layers of the superior colliculus of the monkey, many cells have been found to have an enhanced response to a stimulus when the monkey uses the stimulus as a target for a saccadic eye movement (Goldberg and Wurtz, 1972). This enhancement is not strictly sensory in the sense that it is not produced if the stimulus is not followed by a saccade. It also does not seem strictly associated with a motor response, since the temporal delay between the enhanced response and the saccade can vary considerably (Wurtz and Mohler, 1976a). The enhancement phenomenon was suggested as a neural correlate of "directing visual attention", since it modifies the visual input and enhances it at selective locations when the sensory input remains constant (Goldberg and Wurtz, *op. cit.*). The intimate relation of the enhancement to eye movements, and its absence when the saccade is replaced by other responses (Wurtz and Mohler, *op. cit.*, Wurtz *et al.*, 1982) suggest, however, that this mechanism is specifically related to saccadic eye movements rather than to operations associated with the shifting of an internal processing focus. Similar enhancement that depends on saccade initiation to a visual target has also been described in the frontal eye fields (Wurtz and Mohler, 1976b) and in prestriate cortex, probably area V4 (Fischer and Boch, 1981).

Another area that exhibits similar enhancement phenomena, but not exclusively to saccades, is area 7 of the posterior parietal lobe of the monkey. Using recordings from behaving monkeys, Mountcastle and his collaborators (Mountcastle, 1976, Mountcastle *et al.*, 1975) found three populations of cells in area 7 that respond selectively (i) when the monkey fixates an object of interest within its immediate surrounding (fixation neurons), (ii) when it tracks an object of interest (tracking neurons), and (iii) when it saccades to an object of interest (saccade neurons). (Tracking neurons were also described in area MST (Newsome and Wurtz, 1982).) Studies by Robinson *et al.* (1978) indicated that all of these neurons can also be driven by passive sensory stimulation, but their response is considerably enhanced when the stimulation is 'selected' by the monkey to initiate a response. On the basis of such findings it was suggested by Mountcastle (as well as by Posner, 1980; Robinson *et al.*, 1978; Wurtz *et al.*, 1982) that mechanisms in area 7 are

responsible for "directing visual attention" to selected stimuli. These mechanisms may be primarily related, however, to tasks requiring hand-eye coordination for manipulation in the reachable space (Mountcastle, 1976), and there is at present no direct evidence to link them with visual routines and the shift of processing focus discussed above.[16]

In area TE of the inferotemporal cortex units were found whose responses depend strongly upon the visual task performed by the animal. Fuster and Jervey (1981) described units that responded strongly to the stimulus' color, but only when color was the relevant parameter in a matching task. Richmond and Sato (1982) found units whose responses to a given stimulus were enhanced when the stimulus was used in a pattern discrimination task, but not in other tasks (e.g., when the stimulus was monitored to detect its dimming).

In a number of visual areas, including V1, V2, and MT, enhanced responses associated with performing specific visual tasks were not found (Newsome and Wurtz, 1982; Wurtz *et al.*, 1982). It remains possible, however, that task-specific modulation would be observed when employing different visual tasks. Finally, responses in the pulvinar (Gattas *et al.*, 1979) were shown to be strongly modulated by attentional and situational variables. It remains unclear, however, whether these modulations are localized (i.e., if they are restricted to a particular location in the visual field) and whether they are task-specific.

Physiological evidence of a different kind comes from visual evoked potential (VEP) studies. With fixed visual input and in the absence of eye movements, changes in VEP can be induced, for example, by instructing the subject to "attend" to different spatial locations (e.g., van Voorhis and Hillyard, 1977). This evidence may not be of direct relevance to visual routines, since it is not clear whether there is a relation between the voluntary 'direction of visual attention' used in these experiments and the shift of processing focus in visual routines. VEP studies may nonetheless provide at least some evidence regarding the possibility of internal shift operations.

In assessing the relevance of these physiological findings to the shifting of the processing focus it would be useful to distinguish three types of interactions between the physiological responses and the visual task performed by the experimental animal. The three types are task-dependent, task-location dependent, and location-dependent responses.

---

[16]A possible exception is some preliminary evidence by Robinson *et al.* (1978) suggesting that, unlike the superior colliculus, enhancement effects in the parietal cortex may be dissociated from movement. That is, a response of a cell may be facilitated when the animal is required to attend to a stimulus even when the stimulus is not used as a target for hand or eye movement.

A response is task-dependent if, for a given visual stimulus, it depends upon the visual task being performed. Some of the units described in area TE, for instance, are clearly task-dependent in this sense. In contrast, units in area V1 for example, appear to be task-independent. Task-dependent responses suggest that the units do not belong to the bottom-up generation of the early visual representations, and that they may participate in the application of visual routines. Task-dependence by itself does not necessarily imply, however, the existence of shift operations. Of more direct relevance to shift operations are responses that are both task- and location-dependent. A task-location dependent unit would respond preferentially to a stimulus when a given task is performed at a given location. Unlike task-dependent units, it would show a different response to the same stimulus when an identical task is applied to a different location. Unlike the spotlight metaphor of visual attention, it would show different responses when different tasks are performed at the same locations.

There is at least some evidence for the existence of such task-location dependent responses. The response of a saccade neuron in the superior colliculus, for example, is enhanced only when a saccade is initiated in the general direction of the unit's receptive field. A saccade towards a different location would not produce the same enhancement. The response is thus enhanced only when a specific location is selected for a specific task.

Unfortunately, many of the other task-dependent responses have not been tested for location specificity. It would be of interest to examine similar task-location dependence in tasks other than eye movement, and in the visual cortex rather than the superior colliculus. For example, the units described by Fuster and Jervey (1981) showed task-dependent response (responded strongly during a color matching task, but not during a form matching task). It would be interesting to know whether the enhanced response is also location specific. For example, if during a color matching task, when several stimuli are presented simultaneously, the response would be enhanced only at the location used for the matching task.

Finally, of particular interest would be units referred to above as location-dependent (but task-independent). Such a unit would respond preferentially to a stimulus when it is used not in a single task but in a variety of different visual tasks. Such units may be a part of a general 'shift controller' that selects a location for processing independent of the specific operation to be applied. Of the areas discussed above, the responses in area 7, the superior colliculus, and TE, do not seem appropriate for such a 'shift controller'. The pulvinar remains a possibility worthy of further exploration in view of its rich pattern of reciprocal and orderly connections with a variety of visual areas (Beneveneto and Davis, 1977; Rezak and Beneveneto, 1979).

3.3. Indexing

Computational considerations strongly suggest the use of internal shifts of the processing focus. This notion is supported by psychological evidence, and to some degree by physiological data.

The next issue to be considered is the selection problem: how specific locations are selected for further processing. There are various manners in which such a selection process could be realized. On a digital computer, for instance, the selection can take place by providing the coordinates of the next location to be processed. The content of the specified address can then be inspected and processed. This is probably not how locations are being selected for processing in the human visual system. What determines, then, the next location to be processed, and how is the processing focus moved from one location to the next?

In this section we shall consider one operation which seems to be used by the visual system in shifting the processing focus. This operation is called 'indexing'. It can be described as a shift of the processing focus to special 'odd-man-out' locations. These locations are detected in parallel across the base representations, and can serve as 'anchor points' for the application of visual routines.

As an example of indexing, suppose that a page of printed text is to be inspected for the occurrence of the letter 'A'. In a background of similar letters, the 'A' will not stand out, and considerable scanning will be required for its detection (Nickerson, 1966). If, however, all the letters remain stationary with the exception of one which is jiggled, or if all the letters are red with the exception of one green letter, the odd-man-out will be immediately identified.

The identification of the odd-man-out items proceeds in this case in several stages.[17] First the odd-man-out location is detected on the basis of its unique motion or color properties. Next, the processing focus is shifted to this odd-man-out location. This is the indexing stage. As a result of this stage, visual routines can be applied to the figure. By applying the appropriate routines, the figure is identified.

Indexing also played a role in the inside/outside example examined in Section 1.2. It was noted that one plausible strategy is to start the processing at the location marked by the X figure. This raises a problem, since the location of the X and of the closed curve were not known in advance. If the X can define an indexable location, that is, if it can serve to attract the

[17]The reasons for assuming several stages are both theoretical and empirical. On the empirical side, the experiments by Posner, Treisman, and Tsal provide support for this view.

processing focus, then the execution of the routine can start at that location. More generally, indexable locations can serve as starting points or 'anchors' for visual routines. In a novel scene, it would be possible to direct the processing focus immediately to a salient indexable item, and start the processing at that location. This will be particularly valuable in the execution of universal routines that are to be applied prior to any analysis of the viewed objects.

The indexing operation can be further subdivided into three successive stages. First, properties used for indexing, such as motion, orientation, and color, must be computed across the base representations. Second, an 'odd-man-out operation' is required to define locations that are sufficiently different from their surroundings. The third and final stage is the shift of the processing focus to the indexed location. These three stages are examined in turn in the next three subsections.

3.3.1. Indexable properties

Certain odd-man-out items can serve for immediate indexing, while others cannot. For example, orientation and direction of motion are indexable, while a single occurrence of the letter 'A' among similar letters does not define an indexable location. This is to be expected, since the recognition of letters requires the application of visual routines while indexing must precede their application. The first question that arises, therefore, is what the set of elemental properties is that can be computed everywhere across the base representations prior to the application of visual routines.

One method of exploring indexable properties empirically is by employing an odd-man-out test. If an item is singled out in the visual field by an indexable property, then its detection is expected to be immediate. The ability to index an item by its color, for instance, implies that a red item in a field of green items should be detected in roughly constant time, independent of the number of green distractors.

Using this and other techniques, A. Treisman and her collaborators (Treisman, 1977; Treisman and Gelade, 1980; see also Beck and Ambler, 1972, 1973; Pomerantz *et al.*, 1977) have shown that color and simple shape parameters can serve for immediate indexing. For example, the time to detect a target blue X in a field of brown T's and green X's does not change significantly as the number of distractors is increased (up to 30 in these experiments). The target is immediately indexable by its unique color. Similarly, a target green S letter is detectable in a field of brown T's and green X's in constant time. In this case it is probably indexable by certain shape parameters, although it cannot be determined from the experiments what the relevant parameters are. Possible candidates include (i) curvature, (ii) orientation, since the S contains some orientations that are missing in the X and

T, and (iii) the number of terminators, which is two for the S, but higher for the X and T. It would be of interest to explore the indexability of these and other properties in an attempt to discover the complete set of indexable properties.

The notion of a severely limited set of properties that can be processed 'pre-attentively' agrees well with Julesz' studies of texture perception (see Julesz (1981) for a review). In detailed studies, Julesz and his collaborators have found that only a limited set of features, which he termed 'textons', can mediate immediate texture discrimination. These textons include color, elongated blobs of specific sizes, orientations, and aspect ratios, and the terminations of these elongated blobs.

These psychological studies are also in general agreement with physiological evidence. Properties such as motion, orientation, and color, were found to be extracted in parallel by units that cover the visual field. On physiological grounds these properties are suitable, therefore, for immediate indexing.

The emerging picture is, in conclusion, that a small number of properties are computed in parallel over the base representations prior to the application of visual routines, and represented in ordered retinotopic maps. Several of these properties are known, but a complete list is yet to be established. The results are then used in a number of visual tasks including, probably, texture discrimination, motion correspondence, stereo, and indexing.

3.3.2. Defining an indexable location

Following the initial computation of the elementary properties, the next stage in the indexing operation requires comparisons among properties computed at different locations to define the odd-man-out indexable locations.

Psychological evidence suggests that only simple comparisons are used at this stage. Several studies by Treisman and her collaborators examined the problem of whether different properties measured at a given location can be combined prior to the indexing operation.[18] They have tested, for instance, whether a green T could be detected in a field of brown T's and Green X's. The target in this case matches half the distractors in color, and the other half in shape. It is the combination of shape and color that makes it distinct. Earlier experiments have established that such a target is indexable if it has a unique color or shape. The question now was whether the conjunction of two indexable properties is also immediately indexable. The empirical evidence indicates that items cannot be indexed by a conjunction of properties: the time to detect the target increases linearly in the conjunction task with the number of distractors. The results obtained by Treisman *et al.* were con-

[18]Triesman's own approach to the problem was somewhat different from the one discussed here.

sistent with a serial self-terminating search in which the items are examined sequentially until the target is reached.

The difference between single and double indexing supports the view that the computations performed in parallel by the distributed local units are severely limited. In particular, these units cannot combine two indexable properties to define a new indexable property. In a scheme where most of the computation is performed by a directable central processor, these results also place constraints on the communication between the local units and the central processor. The central processor is assumed to be computationally powerful, and consequently it can also be assumed that if the signals relayed to it from the local units contained sufficient information for double indexing, this information could have been put to use by the central processor. Since it is not, the information relayed to the central processor must be limited.

The results regarding single and double indexing can be explained by assuming that the local computation that precedes indexing is limited to simple local comparisons. For example, the color in a small neighborhood may be compared with the color in a surrounding area, employing, perhaps, lateral inhibition between similar detectors (Estes, 1972; Andriessen and Bouma, 1976; Pomerantz *et al.*, 1977). If the item differs significantly from its surround, the difference signal can be used in shifting the processing focus to that location. If an item is distinguishable from its surround by the conjunction of two properties such as color and orientation, then no difference signal will be generated by either the color or the orientation comparisons, and direct indexing will not be possible. Such a local comparison will also allow the indexing of a local, rather than a global, odd-man-out. Suppose, for example, that the visual field contains green and red elements in equal numbers, but one and only one of the green elements is completely surrounded by a large region of red elements. If the local elements signaled not their colors but the results of local color comparisons, then the odd-man-out alone would produce a difference signal and would therefore be indexable. To explore the computations performed at the distributed stage it would be of interest, therefore, to examine the indexability of local odd-men-out. Various properties can be tested, while manipulating the size and shape of the surrounding region.

3.3.3. Shifting the processing focus to an indexable location

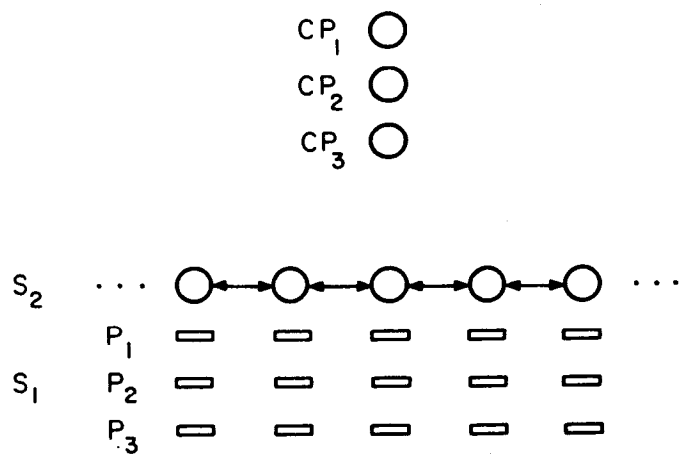
The discussion so far suggests the following indexing scheme. A number of elementary properties are computed in parallel across the visual field. For each property, local comparisons are performed everywhere. The resulting difference signals are combined somehow to produce a final odd-man-out signal at each location. The processing focus then shifts to the location of the strongest signal. This final shift operation will be examined next.

Several studies of selective visual attention likened the internal shift operation to the directing of a spotlight. A directable spotlight is used to 'illuminate' a restricted region of the visual field, and only the information within the region can be inspected. This is, of course, only a metaphor that still requires an agent to direct the spotlight and observe the illuminated region. The goal of this section is to give a more concrete notion of the shift in processing focus, and, using a simple example, to show what it means and how it may be implemented.

The example we shall examine is a version of the property-conjunction problem mentioned in the previous section. Suppose that small colored bars are scattered over the visual field. One of them is red, all the others are green. The task is to report the orientation of the red bar. We would like therefore to 'shift' the processing focus to the red bar and 'read out' its orientation.

A simplified scheme for handling this task is illustrated schematically in Fig. 7. This scheme incorporates the first two stages in the indexing operation discussed above. In the first stage (*S*1 in the figure) a number of different properties (denoted by P_1, P_2, P_3 in the figure) are being detected at each location. The existence of a horizontal green bar, for example, at a given location, will be reflected by the activity of the color- and orientation-detecting units at that location. In addition to these local units there is also a central common representation of the various properties, denoted by CP_1, CP_2, CP_3,

Figure 7. *A simplified scheme that can serve as a basis for the indexing operation. In the first stage* (S_1)*, a number of properties* (P_1, P_2, P_3 *in figure) are detected everywhere. In the subsequent stage* (S_2)*, local comparisons generate difference signals. The element generating the strongest signal is mapped onto the central common representations* (CP_1, CP_2, CP_3).



in the figure. For simplicity, we shall assume that all of the local detectors are connected to the corresponding unit in the central representation. There is, for instance, a common central unit to which all of the local units that signal vertical orientation are connected.

It is suggested that to perform the task defined above and determine the orientation of the red bar, this orientation must be represented in the central common representation. Subsequent processing stages have access to this common representation, but not to all of the local detectors. To answer the question, "what is the orientation of the red element", this orientation alone must therefore be mapped somehow into the common representation.

In section 3.3.2, it was suggested that the initial detection of the various local properties is followed by local comparisons that generate difference signals. These comparisons take place in stage *S*2 in Fig. 7, where the odd-man-out item will end up with the strongest signal. Following these two initial stages, it is not too difficult to conceive of mechanisms by which the most active unit in *S*2 would inhibit all the others, and as a result the properties of all but the odd-man-out location would be inhibited from reaching the central representation.[19] The central representations would then represent faithfully the properties of the odd-man-out item, the red bar in our example. At this stage the processing is focused on the red element and its properties are consequently represented explicitly in the central representation, accessible to subsequent processing stages. The initial question is thereby answered, without the use of a specialized vertical red line detector.

In this scheme, only the properties of the odd-man-out item can be detected immediately. Other items will have to await additional processing stages. The above scheme can be easily extended to generate successive 'shifts of the processing focus' from one element to another, in an order that depends on the strength of their signals in *S*2. These successive shifts mean that the properties of different elements will be mapped successively onto the common representations.

Possible mechanisms for performing indexing and processing focus shifts would not be considered here beyond the simple scheme discussed so far. But even this simplified scheme illustrates a number of points regarding shift and indexing. First, it provides an example for what it means to shift the processing focus to a given location. In this case, the shift entailed a selective

[19]Models for this stage are being tested by C. Koch at the M.I.T. A.I. Lab. One interesting result from this modeling is that a realization of the inhibition among units leads naturally to the processing focus being shifted continuously from item to item rather than 'leaping', disappearing at one location and reappearing at another. The models also account for the phenomenon that being an odd–man–out is not a simple all or none property (Engel, 1974). With increased dissimilarity, a target item can be detected immediately over a larger area.

readout to the central common representations. Second, it illustrates that shift of the processing focus can be achieved in a simple manner without physical shifts or an internal 'spotlight'. Third, it raises the point that the shift of the processing focus is not a single elementary operation but a family of operations, only some of which were discussed above. There is, for example, some evidence for the use of 'similarity enhancement': when the processing focus is centered on a given item, similar items nearby become more likely to be processed next. There is also some degree of 'central control' over the processing focus. Although the shift appears to be determined primarily by the visual input, there is also a possibility of directing the processing focus voluntarily, for example to the right or to the left of fixation (van Voorhis and Hillyard, 1977).

Finally, it suggests that psychophysical experiments of the type used by Julesz, Treisman and others, combined with physiological studies of the kind described in Section 3.2, can provide guidance for developing detailed testable models for the shift operations and their implementation in the visual system.

In summary, the execution of visual routines requires a capacity to control the locations at which elemental operations are applied. Psychological evidence, and to some degree physiological evidence, are in agreement with the general notion of an internal shift of the processing focus. This shift is obtained by a family of related processes. One of them is the indexing operation, which directs the processing focus towards certain odd-man-out locations. Indexing requires three successive stages. First, a set of properties that can be used for indexing, such as orientation, motion, and color, are computed in parallel across the base representation. Second, a location that differs significantly from its surroundings in one of these properties (but not their combinations) can be singled out as an indexed location. Finally, the processing focus is redirected towards the indexed location. This redirection can be achieved by simple schemes of interactions among the initial detecting units and central common representations that lead to a selective mapping from the initial detectors to the common representations.

3.4. Bounded activation (coloring)

The bounded activation, or 'coloring' operation, was suggested in Section 1.2. in examining the inside/outside relation. It consisted of the spread of activation over a surface in the base representation emanating from a given location or contour, and stopping at discontinuity boundaries.

The results of the coloring operation may be retained in the incremental representation for further use by additional routines. Coloring provides in

this manner one method for defining larger units in the unarticulated base representations: the 'colored' region becomes a unit to which routines can be applied selectively. A simple example of this possible role of the coloring operation was mentioned in Section 2.2: the initial 'coloring' could facilitate subsequent inside/outside judgments.

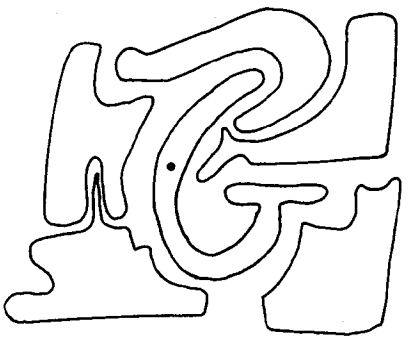
A more complicated example along the same line is illustrated in Fig. 8. The visual task here is to identify the sub-figure marked by the black dot. One may have the subjective feeling of being able to concentrate on this sub-figure, and 'pull it out' from its complicated background. This capacity to 'pull out' the figure of interest can also be tested objectively, for example, by testing how well the sub-figure can be identified. It is easily seen in Fig. 8 that the marked sub-figure has the shape of the letter G. The area surrounding the sub-figure in close proximity contains a myriad of irrelevant features, and therefore identification would be difficult, unless processing can be directed to this sub-figure.

The sub-figure of interest in Fig. 8 is the region inside which the black dot resides. This region could be defined and separated from its surroundings by using the area activation operation. Recognition routines could then concentrate on the activated region, ignoring the irrelevant contours. This examples uses an artificial stimulus, but the ability to identify a region and process it selectively seems equally useful for the recognition of objects in natural scenes.

3.4.1. Discontinuity boundaries for the coloring operation

The activation operation is supposed to spread until a discontinuity bound-

Figure 8. *The visual task here is to identify the subfigure containing the black dot. This figure (the letter 'C') can be recognized despite the presence of confounding features in close proximity to its contours, the capacity to 'pull out' the figure from the irrelevant background may involve the bounded activation operation.*



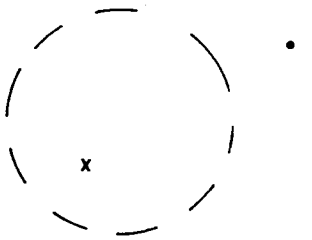
ary is reached. This raises the question of what constitutes a discontinuity boundary for the activation operation. In Fig. 8, lines in the two-dimensional drawing served for this task. If activation is applied to the base representations discussed in Section 2, it is expected that discontinuities in depth, surface orientation, and texture, will all serve a similar role. The use of boundaries to check the activation spread is not straightforward. It appears that in certain situations the boundaries do not have to be entirely continuous in order to block the coloring spread. In Fig. 9, a curve is defined by a fragmented line, but it is still immediately clear that the X lies inside and the black dot outside this curve.[20] If activation is to be used in this situation as well, then incomplete boundaries should have the capacity to block the activation spread. Finally, the activation is sometimes required to spread across certain boundaries. For example, in Fig. 10, which is similar to Fig. 8, the letter G is still recognizable, in spite of the internal bounding contours. To allow the coloring of the entire sub-figure in this case, the activation must spread across internal boundaries.

In conclusion, the bounded activation, and in particular, its interactions with different contours, is a complicated process. It is possible that as far as the activation operation is concerned, boundaries are not defined universally, but may be defined somewhat differently in different routines.

3.4.2. A mechanism for bounded activation and its implications

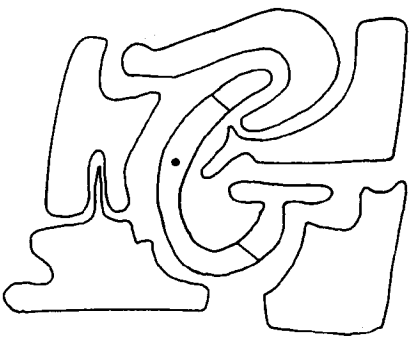
The 'coloring' spread can be realized by using only simple, local operations. The activation can spread in a network in which each element excites all of its neighbors.

Figure 9. *Fragmented boundaries. The curve is defined by a dashed line, but inside/outside judgments are still immediate.*



[20]Empirical results show that inside/outside judgments using dashed boundaries require somewhat longer times compared with continuous curves, suggesting that fragmented boundaries may require additional processing. The extra cost associated with fragmental boundaries is small. In a series of experiments performed by J. Varanese at Harvard University this cost averaged about 20 msec. The mean response time was about 540 msec (Varanese, 1983).

Figure 10. *Additional internal lines are introduced into the G-shaped subfigure. If bounded activation is used to 'color' this figure, it must spread across the internal contours.*



A second network containing a map of the discontinuity boundaries will be used to check the activation spread. An element in the activation network will be activated if any of its neighbors is turned on, provided that the corresponding location in the second, control network, does not contain a boundary. The turning on of a single element in the activation network will thus initiate an activation spread from the selected point outwards, that will fill the area bounded by the surrounding contours. (Each element may also have neighborhoods of different sizes, to allow a more efficient, multi-resolution implementation.)

In this scheme, an 'activity layer' serves for the execution of the basic operation, subject to the constraints in a second 'control layer'. The control layer may receive its content (the discontinuity boundaries) from a variety of sources, which thereby affect the execution of the operation.

An interesting question to consider is whether the visual system incorporates mechanisms of this general sort. If this were the case, the interconnected network of cells in cortical visual areas may contain distinct subnetworks for carrying out the different elementary operations. Some layers of cells within the retinotopically organized visual areas would then be best understood as serving for the execution of basic operations. Other layers receiving their inputs from different visual areas may serve in this scheme for the control of these operations.

If such networks for executing and controlling basic operations are incorporated in the visual system, they will have important implications for the interpretation of physiological data. In exploring such networks, physiological studies that attempt to characterize units in terms of their optimal stimuli would run into difficulties. The activity of units in such networks would be

better understood not in terms of high-order features extracted by the units, but in terms of the basic operations performed by the networks. Elucidating the basic operations would therefore provide clues for understanding the activity in such networks and their patterns of interconnections.

3.5. Boundary tracing and activation

Since contours and boundaries of different types are fundamental entities in visual perception, a basic operation that could serve a useful role in visual routines is the tracking of contours in the base representation. This section examines the tracing operation in two parts. The first shows examples of boundary tracing and activation and their use in visual routines. The second examines the requirements imposed by the goal of having a useful, flexible, tracing operation.

3.5.1. Examples of tracing and activation

A simple example that will benefit from the operation of contour tracing is the problem of determining whether a contour is open or closed. If the contour is isolated in the visual field, an answer can be obtained by detecting the presence or absence of contour terminators. This strategy would not apply, however, in the presence of additional contours. This is an example of the 'figure in a context' problem (Minsky and Papert, 1969): figural properties are often substantially more difficult to establish in the presence of additional context. In the case of open and closed curves, it becomes necessary to relate the terminations to the contour in question. The problem can be solved by tracing the contour and testing for the presence of termination points on that contour.

Another simple example which illustrates the role of boundary tracing is shown in Fig. 11. The question here is whether there are two X's lying on a common curve. The answer seems immediate and effortless, but how is it achieved? Unlike the detection of single indexable items, it cannot be mediated by a fixed array of two-X's-on-a-curve detectors. Instead, I suggest that this simple perception conceals, in fact, an elaborate chain of events. In response to the question, a routine has been compiled and executed. An appropriate routine can be constructed if the repertoire of basic operations included the indexing of the X's and the tracking of curves. The tracking provides in this task a useful identity, or 'sameness' operator: it serves to verify that the two X figures are marked on the same curve, and not on two disconnected curves.

This task has been investigated recently by Jolicoeur *et al.* (1984, Reference note 1) and the results strongly supported the use of an internal contour

tracing operation. Each display in this study contained two separate curves. In all trials there was an X at the fixation point, intersecting one of the curves. A second X could lie either on the same or on the second curve, and the observer's task was to decide as quickly as possible whether the two X's lay on the same or different curves. The physical distance separating the two X's was always 1.8° of visual angle. When the two X's lay on the same curve, their distance along the curve could be changed, however, in increments of 2.2° of visual angle (measured along the curve).

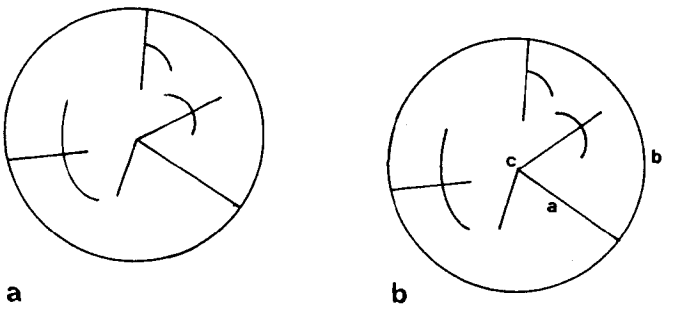
The main result from a number of related experiments was that the time to detect that the two X's lay on the same curve increased monotonically, and roughly linearly, with their separation along the curve. This result suggests the use of a tracing operation, at an average speed of about 24 msec per degree of visual angle. The short presentation time (250 msec) precluded the tracing of the curve using eye movements, hence the tracing operation must be performed internally.

Although the task in this experiment apparently employed a rather elaborate visual routine, it nevertheless appeared immediate and effortless. Response times were relatively short, about 750 msec for the fastest condition. When subjects were asked to describe how they performed the task, the main response was that the two X's were "simply seen" to lie on either the same curve or on different curves. No subject reported any scanning along a curve before making a decision.

The example above employed the tracking of a single contour. In other cases, it would be advantageous to activate a number of contours simultaneously. In Fig. 12a, for instance, the task is to establish visually whether there is a path connecting the center of the figure to the surrounding contour. The solution can be obtained effortlessly by looking at the figure, but again, it must involve in fact a complicated chain of processing. To cope with this

Figure 11. *The task here is to determine visually whether the two X's lie on the same curve. This simple task requires in fact complex processing that probably includes the use of a contour tracing operation.*

Figure 12. *The task in* a *is to determine visually whether there is a path connecting the center of the figure to the surrounding circle. In* b *the solution is labeled. The interpretation of such labels relys upon a set of common natural visual routines.*



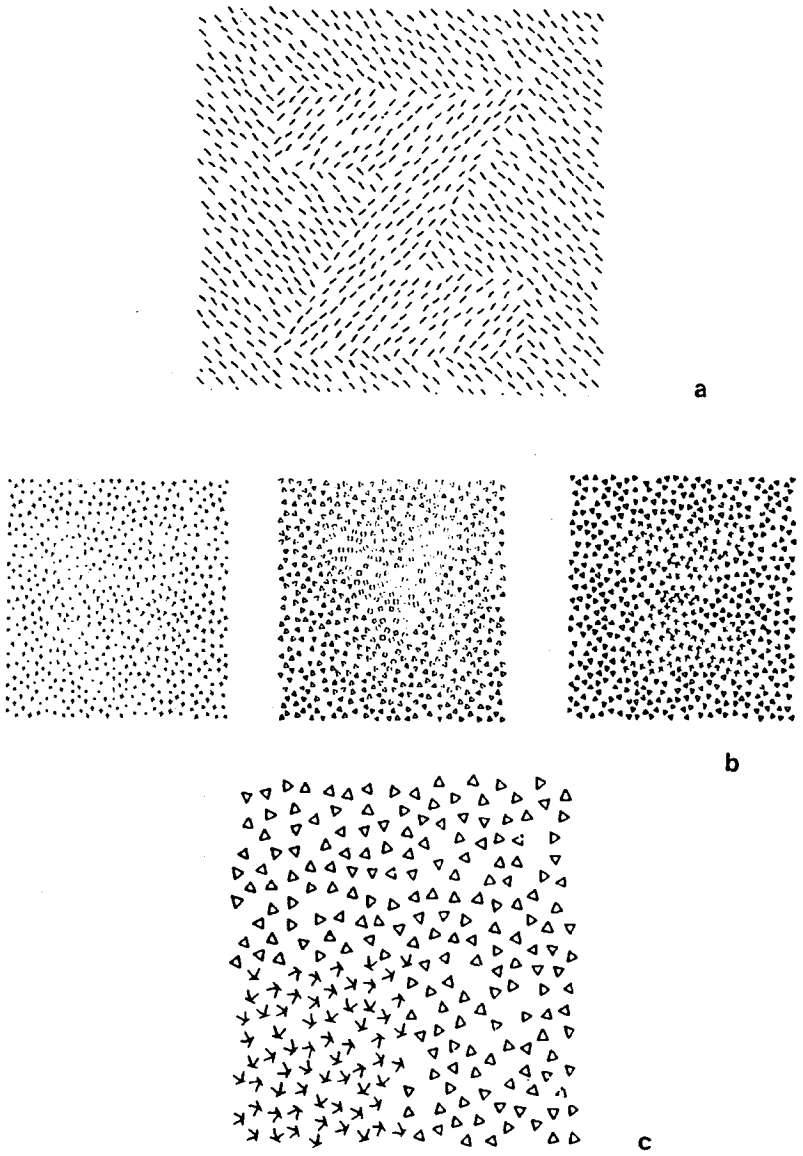
seemingly simple problem, visual routine must (i) identify the location referred to as "the center of the figure", (ii) identify the outside contour, and (iii) determine whether there is a path connecting the two. (It is also possible to proceed from the outside inwards.) By analogy with the area activation, the solution can be found by activating contours at the center point and examining the activation spread to the periphery. In Fig. 12b, the solution is labeled: the center is marked by the letter *c*, the surrounding boundary by *b*, and the connecting path by *a*. Labeling of this kind is common in describing graphs and figures. A point worth noting is that to be unambiguous, such notations must reply upon the use of common, natural visual routines. The label *b*, for example, is detached from the figure and does not identify explicitly a complete contour. The labeling notation implicitly assumes that there is a common procedure for identifying a distinct contour associated with the label.[21]

In searching for a connecting contour in Fig. 12, the contours could be activated in parallel, in a manner analogous to area coloring. It seems likely that at least in certain situations, the search for a connecting path is not just an unguided sequential tracking and exploration of all possible paths. A definite answer would require, however, an empirical investigation, for example, by manipulating the number of distracting culy-de-sac paths connected to the center and to the surrounding contour. In a sequential search, detection of the connecting path should be strongly affected by the addition of distracting paths. If, on the other hand, activation can spread along many paths simultaneously, detection will be little affected by the additional paths.

[21]It is also of interest to consider how we locate the center of figures. In Noton and Stark's (1971) study of eye movements, there are some indications of an ability to start the scanning of a figure approximately at its center.

Tracking boundaries in the base representations. The examples mentioned above used contours in schematic line drawings. If boundary tracking is indeed a basic operation in establishing properties and spatial relations, it is expected to be applicable not only to such lines, but also to the different types of contours and discontinuity boundaries in the base representations. Exper-

Figure 13. *Certain texture boundaries can delineate effectively shape for recognition (a), while others cannot (b). Micropatterns that are ineffective for delineating shape boundaries can nevertheless give rise to discriminable textures (c). (From Riley, 1981).*



iments with textures, for instance, have demonstrated that texture boundaries can be effective for defining shapes in visual recognition. Figure 13a (reproduced from Riley (1981)) illustrates an easily recognizable Z shape defined by texture boundaries. Not all types of discontinuity can be used for rapid recognition. In Fig. 13b, for example, recognition is difficult. The boundaries defined for instance by a transition between small k-like figures and triangles cannot be used in immediate recognition, although the texture generated by these micropatterns is easily discriminable (Fig. 13c)).

What makes some discontinuities considerably more efficient than others in facilitating recognition? Recognition requires the establishment of spatial properties and relations. It can therefore be expected that recognition is facilitated if the defining boundaries are already represented in the base representations, so that operations such as activation and tracking may be applied to them. Other discontinuities that are not represented in the base representations can be detected by applying appropriate visual routines, but recognition based on these contours will be considerably slower.[22]

3.5.2. Requirements on boundary tracing

The tracing of a contour is a simple operation when the contour is continuous, isolated, and well defined. When these conditions are not met, the tracing operation must cope with a number of challenging requirements. These requirements, and their implications for the tracing operation, are examined in this section.

(a) Tracing incomplete boundaries. The incompleteness of boundaries and contours is a well-known difficulty in image processing systems. Edges and contours produced by such systems often suffer from gaps due to such problems as noise and insufficient contrast. This difficulty is probably not confined to man–made systems alone; boundaries detected by the early processes in the human visual system are also unlikely to be perfect. The boundary tracing operation should not be limited, therefore, to continuous boundaries only. As noted above with respect to inside/outside routines for human perception, fragmented contours can indeed often replace continuous ones.

[22]M. Riley (1981) has found a close agreement between texture boundaries that can be used in immediate recognition and boundaries that can be used in long–range apparent motion (cf. Ullman, 1979). Boundaries participating in motion correspondence must be made explicit within the base representations, so that they can be matched over discrete frames. The implication is that the boundaries involved in immediate recognition also preexist in the base representations.

(b) Tracking across intersections and branches. In tracing a boundary crossings and branching points can be encountered. It will then become necessary to decide which branch is the natural continuation of the curve. Similarity of color, contrast, motion, etc. may affect this decision. For similar contours, collinearity, or minimal change in direction (and perhaps curvature) seem to be the main criteria for preferring one branch over another.

Tracking a contour through an intersection can often be useful in obtaining a stable description of the contour for recognition purposes. Consider, for example, the two different instances of the numeral '2' in Fig. 14a. There are considerable differences between these two shapes. For example, one contains a hole, while the other does not. Suppose, however, that the contours are traced, and decomposed at places of maxima in curvature. This will lead to the decomposition shown in Fig. 14b. In the resulting descriptions, the

Figure 14. *The tracking of a contour through an intersection is used here in generating a stable description of the contour. a, Two instances of the numeral '2'. b, In spite of the marked difference in their shape, their eventual decomposition and description are highly similar.*

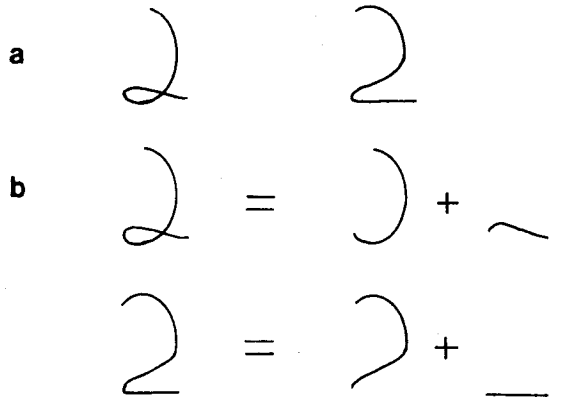
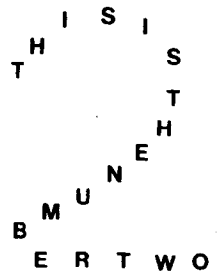


Figure 15. *Tracing a skeleton. The overall figure can be traced and recognized without recognizing first all of the individual components.*



decomposition into strokes, and the shapes of the underlying strokes, are highly similar.

(c) Tracking at different resolutions. Tracking can proceed along the main skeleton of a contour without tracing its individual components. An example is illustrated in Fig. 15, where a figure is constructed from a collection of individual tokens. The overall figure can be traced and recognized without tracing and identifying its components.

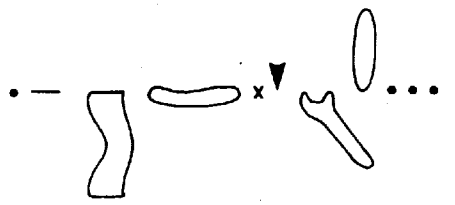
Examples similar to Fig. 15 have been used to argue that 'global' or 'holistic' perception precedes the extraction of local features. According to the visual routines scheme, the constituent line elements are in fact extracted by the earliest visual process and represented in the base representations. The constituents are not recognized, since their recognition requires the application of visual routines. The 'forest before the trees' phenomenon (Johnston and McLelland, 1973; Navon, 1977; Pomerantz *et al.*, 1977) is the result of applying appropriate routines that can trace and analyze aggregates without analyzing their individual components, thereby leading to the recognition of the overall figure prior to the recognition of its constituents.

The ability to trace collections of tokens and extract properties of their arrangement raises a question regarding the role of grouping processes in early vision. Our ability to perceive the collinear arrangement of different tokens, as illustrated in Fig. 16, has been used to argue for the existence of sophisticated grouping processes within the early visual representations that detect such arrangements and make them explicit (Marr, 1976). In this view, these grouping processes participate in the construction of the base representations, and consequently collinear arrangements of tokens are detected and represented throughout the base representation prior to the application of visual routines. An alternative possibility is that such arrangements are identified in fact as a result of applying the appropriate routine. This is not to deny the existence of certain grouping processes within the base representations. There is, in fact, strong evidence in support of the existence of such processes.[23] The more complicated and abstract grouping phenomena such as in Fig. 16 may, nevertheless, be the result of applying the appropriate routines, rather than being explicitly represented in the base representations.

Finally, from the point of view of the underlying mechanism, one obvious possibility is that the operation of tracing an overall skeleton is the result of applying tracing routines to a low resolution copy of the image, mediated by low frequency channels within the visual system. This is not the only possibil-

[23]For evidence supporting the existence of grouping processes within the early creation of the base representations using dot-interference patterns see Glass (1969), Glass and Perez (1973), Marroquin (1976), Stevens (1978). See also a discussion of grouping in early visual processing in Barlow (1981).

Figure 16. *The collinearity of tokens (items and endpoints) can easily be perceived. This perception may be related to a routine that traces collinear arrangements, rather than to sophisticated grouping processes within the base representations.*



ity, however, and in attempting to investigate this operation further, alternative methods for tracing the overall skeleton of figures should also be considered.

In summary, the tracing and activation of boundaries are useful operations in the analysis of shape and the establishment of spatial relations. This is a complicated operation since flexible, reliable, tracing should be able to cope with breaks, crossings, and branching, and with different resolution requirements.

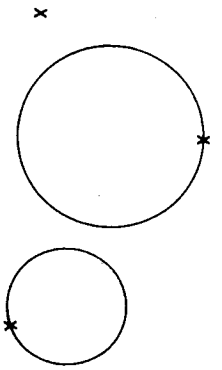
3.6. Marking

In the course of applying a visual routine, the processing shifts across the base representations from one location to another. To control and coordinate the routine, it would be useful to have the capability to keep at least a partial track of the locations already processed.

A simple operation of this type is the marking of a single location for future reference. This operation can be used, for instance, in establishing the closure of a contour. As noted in the preceding section, closure cannot be tested in general by the presence or absence of terminators, but can be established using a combination of tracing and marking. The starting point of the tracing operation is marked, and if the marked location is reached again the tracing is completed, and the contour is known to be closed.

Figure 17 shows a similar problem, which is a version of a problem examined in the previous section. The task here is to determine visually whether there are two X's on the same curve. Once again, the correct answer is perceived immediately. To establish that only a single X lies on the closed curve c, one can use the above strategy of marking the X and tracking the

Figure 17. *The task here is to determine visually whether there are two X's on a common curve. The task could be accomplished by employing marking and tracing operations.*



curve. It is suggested that the perceptual system has marking and tracing in its repertoire of basic operations, and that the simple perception of the X on the curve involved the application of visual routines that employ such operations.

Other tasks may benefit from the marking of more than a single location. A simple example is visual counting, that is, the problem of determining as fast as possible the number of distinct items in view (Atkinson *et al.*, 1969; Kowler and Steinman, 1979).

For a small number of items visual counting is fast and reliable. When the number of items is four or less, the perception of their number is so immediate, that it gave rise to conjecture regarding special *Gestalt* mechanisms that can somehow respond directly to the number of items in view, provided that this number does not exceed four (Atkinson *et al.*, 1969).

In the following section, we shall see that although such mechanisms are possible in principle, they are unlikely to be incorporated in the human visual system. It will be suggested instead that even the perception of a small number of items involves in fact the execution of visual routines in which marking plays an important role.

3.6.1. Comparing schemes for visual counting

Perception-like counting networks. In their book *Perceptrons*, Minsky and Papert (1969, Ch. 1) describe parallel networks that can count the number of elements in their input (see also Milner, 1974). Counting is based on computing the predicates "the input has exactly M points" and "the input has between M and N points" for different values of M and N. For any given

value of M, it is thereby possible to construct a special network that will respond only when the number of items in view is exactly M. Unlike visual routines which are composed of elementary operations, such a network can adequately be described as an elementary mechanism responding directly to the presence of M items in view. Unlike the shifting and marking operations, the computation is performed by these networks uniformly and in parallel over the entire field.

Counting by visual routines. Counting can also be performed by simple visual routines that employ elementary operations such as shifting and marking. For example, the indexing operation described in Section 3.3 can be used to perform the counting task provided that it is extended somewhat to include marking operations. Section 3.3 illustrated how a simple shifting scheme can be used to move the processing focus to an indexable item. In the counting problem, there is more than a single indexable item to be considered. To use the same scheme for counting, the processing focus is required to travel among all of the indexable items, without visiting an item more than once.

A straightforward extension that will allow the shifting scheme in Section 3.3 to travel among different items is to allow it to mark the elements already visited. Simple marking can be obtained in this case by 'switching off' the element at the current location of the processing focus. The shifting scheme described above is always attracted to the location producing the strongest signal. If this signal is turned off, the shift would automatically continue to the new strongest signal. The processing focus can now continue its tour, until all the items have been visited, and their number counted.

A simple example of this counting routine is the 'single point detection' task. In this problem, it is assumed that one or more points can be lit up in the visual field. The task is to say 'yes' if a single point is lit up, and 'no' otherwise. Following the counting procedure outlined above, the first point will soon be reached and masked. If there are no remaining signals, the point was unique and the correct answer is 'yes'; otherwise, it is 'no'.

In the above scheme, counting is achieved by shifting the processing focus among the items of interest without scanning the entire image systematically. Alternatively, shifting and marking can also be used for visual counting by scanning the entire scene in a fixed predetermined pattern. As the number of items increases, programmed scanning may become the more efficient strategy. The two alternative schemes will behave differently for different numbers of items. The fixed scanning scheme is largely independent of the number of items, whereas in the traveling scheme, the computation time will depend on the number of items, as well as on their spatial configuration.

There are two main differences between counting by visual routines of one

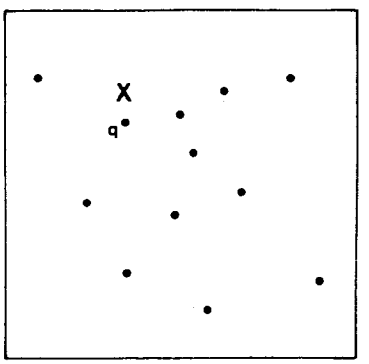
type or another on the one hand, and by specialized counting networks on the other. First, unlike the perception-like networks, the process of determining the number of items by visual routines can be decomposed into a sequence of elementary operations. This decomposition holds true for the perception of a small number of items and even for the single item detection. Second, in contrast with a counting network that is specially constructed for the task of detecting a prescribed number of items, the same elementary operations employed in the counting routine also participate in other visual routines.

This difference makes counting by visual routines more attractive than the counting networks. It does not seem plausible to assume that visual counting is essential enough to justify specialized networks dedicated to this task alone. In other words, visual counting is simply unlikely to be an elementary operation. It is more plausible in my view that visual counting can be performed efficiently as a result of our general capacity to generate and execute visual routines, and the availability of the appropriate elementary operations that can be harnessed for the task.

3.6.2. Reference frames in marking

The marking of a location for later reference requires a coordinate system, or a frame of reference, with respect to which the location is defined. One general question regarding marking is, therefore, what is the referencing scheme in which locations are defined and remembered for subsequent use by visual routines. One possibility is to maintain an internal 'egocentric' spatial map that can then be used in directing the processing focus. The use of marking would then be analogous to reaching in the dark: the location of one or more objects can be remembered, so that they can be reached (approximately) in the dark without external reference cues. It is also possible to use an internal map in combination with external referencing. For example, the

Figure 18. *The use of an external reference. The position of point* q *can be defined and retained relative to the predominant X nearby.*



position of point q in Fig. 18 can be defined and remembered using the prominent X figure nearby. In such a scheme it becomes possible to maintain a crude map with which prominent features can be located, and a more detailed local map in which the position of the marked item is defined with respect to the prominent feature.

The referencing problem can be approached empirically, for example by making a point in figures such as Fig. 18 disappear, then reappear (possibly in a slightly displaced location), and testing the accuracy at which the two locations can be compared. (Care must be taken to avoid apparent motion.) One can test the effect of potential reference markers on the accuracy, and test marking accuracy across eye movements.

3.6.3. Marking and the integration of information in a scene

To be useful in the natural analysis of visual scenes, the marking map should be preserved across eye motions. This means that if a certain location in space is marked prior to an eye movement, the marking should point to the same spatial location following the eye movement. Such a marking operation, combined with the incremental representation, can play a valuable role in integrating the information across eye movements and from different regions in the course of viewing a complete scene.[24]

Suppose, for example, that a scene contains several objects, such as a man at one location, and a dog at another, and that following the visual analysis of the man figure we shift our gaze and processing focus to the dog. The visual analysis of the man figure has been summarized in the incremental representation, and this information is still available at least in part as the gaze is shifted to the dog. In addition to this information we keep a spatial map, a set of spatial pointers, which tell us that the dog is at one direction, and the man at another. Although we no longer see the man clearly, we have a clear notion of what exists where. The 'what' is supplied by the incremental representations, and the 'where' by the marking map.

In such a scheme, we do not maintain a full panoramic representation of the scene. After looking at various parts of the scene, our representation of it will have the following structure. There would be a retinotopic representation of the scene in the current viewing direction. To this representation we can apply visual routines to analyze the properties of, and relations among, the items in view. In addition, we would have markers to the spatial locations of items in the scene already analyzed. These markers can point to peripheral

[24]The problem considered here is not limited to the integration of views across saccadic eye motions, for which an 'integrative visual buffer' has been proposed by Rayner (1978).

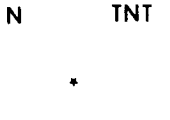
objects, and perhaps even to locations outside the field of view (Attneave and Pierce, 1978). If we are currently looking at the dog, we would see it in detail, and will be able to apply visual routines and extract information regarding the dog's shape. At the same time we know the locations of the other objects in the scene (from the marking map) and what they are (from the incremental representation). We know, for example, the location of the man in the scene. We also know various aspects of his shape, although it may now appear only as a blurred blob, since they are summarized in the incremental representation. To obtain new information, however, we would have to shift our gaze back to the man figure, and apply additional visual routines.

3.6.4. On the spatial resolution of marking and other basic operations

In the visual routines scheme, accuracy in visual counting will depend on the accuracy and spatial resolution of the marking operation. This conclusion is consistent with empirical results obtained in the study of visual counting.[25] Additional perceptual limitations may arise from limitations on the spatial resolution of other basic operations. For example, it is known that spatial relations are difficult to establish in peripheral vision in the presence of distracting figures. An example, due to J. Lettvin (see also Andriessen and Bouma, 1976; Townsend *et al.*, 1971), is shown in Fig. 19. When fixating on the central point from a normal reading distance, the N on the left is recognizable, while the N within the string TNT on the right is not. The flanking letters exert some 'lateral masking' even when their distance from the central letter is well above the two-point resolution at this eccentricity (Riggs, 1965).

Interaction effects of this type may be related to limitations on the spatial resolution of various basic operations, such as indexing, marking, and boundary tracking. The tracking of a line contour, for example, may be distracted by the presence of another contour nearby. As a result, contours may inter-

Figure 19. *Spatial limitations of the elemental operations. When the central mark is fixated, the N on the left is recognizable, while the one on the right is not. This effect may reflect limitations on the spatial resolution of basic operations such as indexing, marking, and boundary tracing.*



[25]For example, Kowler and Steinman (1979) report a puzzling result regarding counting accuracy. It was found that eye movements increase counting accuracy for large (2°) displays, but were not helpful, and sometimes detrimental, with small displays. This result could be explained under the plausible assumptions that marking accuracy is better near fixation, and that it deteriorates across eye movements. As a result, eye movements will improve marking accuracy for large, but not for small, displays.

fere with the application of visual routines to other contours, and consequently with the establishment of spatial relations. Experiments involving the establishment of spatial relations in the presence of distractors would be useful in investigating the spatial resolution of the basic operations, and its dependence on eccentricity.

The hidden complexities in perceiving spatial relationships. We have examined above a number of plausible elemental operations including shift, indexing, bounded activation, boundary tracing and activation, and marking. These operations would be valuable in establishing abstract shape properties and spatial relations, and some of them are partially supported by empirical data. (They certainly do not constitute, however, a comprehensive set.)

The examination of the basic operations and their use reveals that in perceiving spatial relations the visual system accomplishes with intriguing efficiency highly complicated tasks. There are two main sources for these complexities. First, as was illustrated above, from a computational standpoint, the efficient and reliable implementation of each of the elemental operations poses challenging problems. It is evident, for instance, that a sophisticated specialized processor would be required for an efficient and flexible bounded activation operation, or for the tracing of contours and collinear arrangements of tokens.

In addition to the complications involved in the realization of the different elemental operations, new complications are introduced when the elemental operations are assembled into meaningful visual routines. As illustrated by the inside/outside example, in perceiving a given spatial relation different strategies may be employed, depending on various parameters of the stimuli (such as the complexity of the boundary, or the distance of the X from the bounding contour). The immediate perception of seemingly simple relations often requires, therefore, selection among possible routines, followed by the coordinated application of the elemental operations comprising the visual routines. Some of the problems involved in the assembly of the elemental operations into visual routines are discussed briefly in the next section.

4. The assembly, compilation, and storage of visual routines

The use of visual routines allows a variety of properties and relations to be established using a fixed set of basic operations. According to this view, the establishment of relations requires the application of a coordinated sequence of basic operations. We have discussed above a number of plausible basic operations. In this section I shall raise some of the general problems as-

sociated with the construction of useful routines from combinations of basic operations.

The appropriate routine to be applied in a given situation depends on the goal of the computation, and on various parameters of the configuration to be analyzed. We have seen, for example, that the routine for establishing inside/outside relations may depend on various properties of the configuration: in some cases it would be efficient to start at the location of the X figure, in other situations it may be more efficient to start at some distant locations.

Similarly, in Treisman's (1977, 1980) experiments on indexing by two properties (e.g., a vertical red item in a field of vertical green and horizontal red distractors) there are at least two alternative strategies for detecting the target. Since direct indexing by two properties is impossible, one may either scan the red items, testing for orientation, or scan the vertical items, testing for color.[26] The distribution of distractors in the field determines the relative efficiency of these alternative strategies. In such cases it may prove useful, therefore, to precede the application of a particular routine with a stage where certain relevant properties of the configuration to be analyzed are sampled and inspected. It would be of interest to examine whether in the double indexing task, for example, the human visual system tends to employ the more efficient search strategy.

The above discussion introduces what may be called the 'assembly problem'; that is, the problem of how routines are constructed in response to specific goals, and how this generation is controlled by aspects of the configuration to be analyzed. In the above examples, a goal for the computation is set up externally, and an appropriate routine is applied in response. In the course of recognizing and manipulating objects, routines are usually invoked in response to internally generated queries. Some of these routines may be stored in memory rather than assembled anew each time they are needed.

The recognition of a specific object may then use pre-assembled routines for inspecting relevant features and relations among them. Since routines can also be generated efficiently by the assembly mechanism in response to specific goals, it would probably be sufficient to store routines in memory in a skeletonized form only. The assembly mechanism will then fill in details and generate intermediate routines when necessary. In such a scheme, the perceptual activity during recognition will be guided by setting pre-stored goals that the assembly process will then expand into detailed visual routines.

---

[26]There is also a possibility that all the items must be scanned one by one without any selection by color or orientation. This question is relevent for the shift operation discussed in Section 3.2. Recent results by J. Rubin and N. Kanwisher at M.I.T. suggest that it is possible to scan only the items of relevant color and ignore the others.

The application of pre-stored routines rather then assembling them again each time they are required can lead to improvements in performance and the speed-up of performing familiar perceptual tasks. These improvements can come from two different sources. First, assembly time will be saved if the routine is already 'compiled' in memory. The time saving can increase if stored routines for familiar tasks, which may be skeletonized at first, become more detailed, thereby requiring less assembly time. Second, stored routines may be improved with practice, for example, as a result of either external instruction, or by modifying routines when they fail to accomplish their tasks efficiently.

Summary

1. Visual perception requires the capacity to extract abstract shape properties and spatial relations. This requirement divides the overall processing of visual information into two distinct stages. The first is the creation of the base representations (such as the primal sketch and the 2½-D sketch). The second is the application of visual routines to the base representations.

2. The creation of the base representations is a bottom-up and spatially uniform process. The representations it produces are unarticulated and viewer-centered.

3. The application of visual routines is no longer bottom-up, spatially uniform, and viewer-centered. It is at this stage that objects and parts are defined, and their shape properties and spatial relations are established.

4. The perception of abstract shape properties and spatial relations raises two major difficulties. First, the perception of even seemingly simple, immediate properties and relations requires in fact complex computation. Second, visual perception requires the capacity to establish a large variety of different properties and relations.

5. It is suggested that the perception of spatial relation is achieved by the application to the base representations of visual routines that are composed of sequences of elemental operations. Routines for different properties and relations share elemental operations. Using a fixed set of basic operations, the visual system can assemble different routines to extract an unbounded variety of shape properties and spatial relations.

6. Unlike the construction of the base representation, the application of visual routines is not determined by the visual input alone. They are selected or created to meet specific computational goals.

7. Results obtained by the application of visual routines are retained in the incremental representation and can be used by subsequent processes.

8. Some of the elemental operations employed by visual routines are applied to restricted locations in the visual field, rather than to the entire field in parallel. It is suggested that this apparent limitation on spatial parallelism reflects in part essential limitations, inherent to the nature of the computation, rather than non-essential capacity limitations.

9. At a more detailed level, a number of plausible basic operations were suggested, based primarily on their potential usefulness, and supported in part by empirical evidence. These operations include:

9.1. *Shift of the processing focus.* This is a family of operations that allow the application of the same basic operation to different locations across the base representations.

9.2. *Indexing.* This is a shift operation towards special odd-man-out locations. A location can be indexed if it is sufficiently different from its surroundings in an indexable property. Indexable properties, which are computed in parallel by the early visual processes, include contrast, orientation, color, motion, and perhaps also size, binocular disparity, curvature, and the existence of terminators, corners, and intersections.

9.3. *Bounded activation.* This operation consists of the spread of activation over a surface in the base representation, emanating from a given location or contour, and stopping at discontinuity boundaries. This is not a simple operation, since it must cope with difficult problems that arise from the existence of internal contours and fragmented boundaries. A discussion of the mechanisms that may be implicated in this operation suggests that specialized networks may exist within the visual system, for executing and controlling the application of visual routines.

9.4. *Boundary tracing.* This operation consists of either the tracing of a single contour, or the simultaneous activation of a number of contours. This operation must be able to cope with the difficulties raised by the tracing of incomplete boundaries, tracing across intersections and branching points, and tracing contours defined at different resolution scales.

9.5. *Marking.* The operation of marking a location means that this location is remembered, and processing can return to it whenever necessary. Such an operation would be useful in the integration of information in the processing of different parts of a complete scene.

10. It is suggested that the seemingly simple and immediate perception of spatial relations conceals in fact a complex array of processes involved in the selection, assembly, and execution of visual routines.

References

Andriessen, J.J. and Bouma, H. (1976) Eccentric vision: adverse interactions between line segments. *Vis. Res., 16*, 71–78.

Atkinson, J., Campbell, F.W. and Francis, M.R. (1969) The magic number 4 ± 0: A new look at visual numerosity judgments. *Perception, 5*, 327–334.

Attneave, F. and Pierce, C.R. (1978) The accuracy of extrapolating a pointer into perceived and imagined space. *Am. J. Psychol., 91(3)*, 371–387.

Barlow, H.H. (1972) Single units and sensation: A neuron doctrine for perceptual psychology? *Perception, 1*, 371–394.

Barlow, H.B. (1981) Critical limiting factors in the design of the eye and the visual cortex. The Ferrier Lecture 1980. *Proc. Roy. Soc. Lond. B, 212*, 1–34.

Bartlett, F.C. (1932) *Remembering*. Cambridge, Cambridge University Press.

Beck, J. and Ambler, B. (1972) Discriminability of differences in line slope and in line arrangement as a function of mask delay. *Percep. Psychophys. 12(1A)*, 33–38.

Beck, J. and Ambler, B. (1973) The effects of concentrated and distributed attention on peripheral acuity. *Percept. Psychophys., 14(2)*, 225–230.

Beneveneto, L.A. and Davis, B. (1977) Topographical projections of the prestriate cortex to the pulvinar nuclei in the macaque monkey: an autoradiographic study. *Exp. Brain Res., 30*, 405–424.

Biederman, I., Glass, A.L. and Stacy, E.W. (1973) Searching for objects in real–world scenes. *J. exp. Psychol., 97(1)*, 22–27.

Donderi, D.C. and Zelnicker, D. (1969) Parallel processing in visual same–different decisions. *Percep. Psychophys., 5(4)*, 197–200.

Egeth, H., Jonides, J. and Wall, S. (1972) Parallel processing of multi-element displays. *Cog. Psychol., 3*, 674–698.

Engel, F.L. (1971) Visual conspecuity, directed attention and retinal locus. *Vis. Res., 11*, 563–576.

Eriksen, C.W. and Hoffman, J.E. (1972) Temporal and spatial characteristics of selective encoding from visual displays. *Percep. Psychophys., 12(2B)*, 201–204.

Eriksen, C.W. and Schultz, D.W. (1977) Retinal locus and acuity in visual information processing. *Bull. Psychon. Soc., 9(2)*, 81–84.

Estes, W.K. (1972) Interactions of signal and background variables in visual processing. *Percep. Psychophys., 12(3)*, 278–286.

Evans, T.G. (1968) A heuristic program to solve geometric analogy problems. In M. Minsky (ed.), *Semantic Information Processing*. Cambridge, MA, M.I.T. Press.

Fantz, R.L. (1961) The origin of form perception. *Scient. Am., 204(5)*, 66–72.

Fischer, B. and Boch, R. (1981) Enhanced activation of neurons in prelunate cortex before visually guided saccades of trained rhesus monkey. *Exp. Brain Res., 44*, 129–137.

Fuster, J.M. and Jervey, J.P. (1981) Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science, 212*, 952–955.

Gattas, R., Osealdo Cruz, E. and Sousa, A.P.B. (1979) Visual receptive fields of units in the pulvinar of cebus monkey. *Brain Res., 160*, 413–430.

Glass, L. (1969) Moire effect from random dots. *Nature, 243*, 578–580.

Glass, L. and Perez, R. (1973) Perception of random dot interference patterns. *Nature, 246*, 360–362.

Goldberg, M.E. and Wurtz, R.H. (1972) Activity of superior colliculus in behaving monkey. II. Effect of attention of neural responses. *J. Neurophysiol, 35*, 560–574.

Holtzman, J.D. and Gazzaniga, M.S. (1982) Dual task interactions due exclusively to limits in processing resources. *Science, 218*, 1325–1327.

Humphreys, G.W. (1981) On varying the span of visual attention: evidence for two modes of spatial attention. *Q. J. exp. Psychol., 33A*, 17–31.

Johnston, J.C. and McClelland, J.L. (1973) Visual factors in word perception. *Percep. Psychophys., 14(2)*, 365–370.

Jonides, J. and Gleitman, H. (1972) A conceptual category effect in visual search: O as a letter or as digit. *Percep. Psychophys., 12(6)*, 457–460.

Johnson, R.B. and Kirk, N.S. (1960) The perception of size: An experimental synthesis of the associationist and gestalt accounts of the perception of size. Part III. *Q. J. exp. Psychol., 12*, 221–230.

Julesz, B. (1975) Experiments in the visual perception of texture. *Scient. Am., 232(4)*, April 1975, 34–43.

Julesz, B. (1981) Textons, the elements of texture perception, and their interactions. *Nature, 290*, 91–97.

Kahneman, D. (1973) *Attention and Effort*. Englewood Cliffs, NJ, Prentice-Hall.

Kolmogorov, A.N. (1968) Logical basis for information theory and probability theory. *IEEE Trans. Info. Theory, IT-14(5)*, 662–664.

Kowler, E. and Steinman, R.M. (1979) Miniature saccades: eye movements that do not count. *Vis. Res., 19*, 105–108.

Lappin, J.S. and Fuqua, M.A. (1983) Accurate visual measurement of three-dimensional moving patterns. *Science, 221*, 480–482.

Livingstone, M.L. and Hubel, D.J. (1981) Effects of sleep and arousal on the processing of visual information in the cat. *Nature, 291*, 554–561.

Mackworth, N.H. (1965) Visual noise causes tunnel vision. *Psychon. Sci., 3*, 67–68.

Marr, D. (1976) Early processing of visual information. *Phil. Trans. Roy. Soc. and B, 275*, 483–524.

Marr, D. (1980) Visual information processing: the structure and creation of visual representations. *Phil. Trans. Roy. Soc. Lond. B, 290*, 199–218.

Marr, D. and Nishihara, H.K. (1978) Representation and recognition of the spatial organization of three–dimensional shapes. *Proc. Roy. Soc. B, 200*, 269–291.

Marroquin, J.L. (1976) Human visual perception of structure. MSc. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Milner, P.M. (1974) A model for visual shape recognitio. *Psychol. Rev. 81(6)*, 521–535.

Minsky, M. and Papert, S. (1969) *Perceptrons*. Cambridge, MA and London: The M.I.T. Press.

Minsky, M. (1975) A framework for representing knowledge. In P.H. Winston (ed.), *The Psychology of Computer Vision*. New York, Prentice Hall.

Mountcastle, V.B. (1976) The world around us: neural command functions for selective attention. The F.O. Schmitt Lecture in Neuroscience 1975. *Neurosci. Res. Prog. Bull., 14*, Supplement 1–37.

Mountcastle, V.B., Lynch, J.C., Georgopoulos, A., Sakata, H. and Acuna, A. (1975) Posterior parietal association cortex of the monkey: command functions for operations within extrapersonal space. *J. Neurophys., 38*, 871–908.

Navon, D. (1977) Forest before trees: the precedence of global features in visual perception. *Cog. Psychol., 9*, 353–383.

Neisser, U., Novick, R. and Lazar, R. (1963) Searching for ten targets simultaneously. *Percep. Mot. Skills, 17*, 955–961.

Neisser, U. (1967) *Cognitive Psychology*. New York, Prentice-Hall.

Newsome, W.T. and Wurtz, R.H. (1982) Identification of architectonic zones containing visual tracking cells in the superior temporal sulcus of macaque monkeys. *Invest. Ophthal. Vis. Sci., Suppl. 3*, 22, 238.

Nickerson, R.S. (1966) Response times with memory–dependent decision task. *J. exp. Psychol., 72(5)*, 761–769.

Noton, D. and Stark, L. (1971) Eye movements and visual perception. *Scient. Am., 224(6)*, 34–43.

Pomerantz, J.R., Sager, L.C. and Stoever, R.J. (1977) Perception of wholes and of their component parts: some configural superiority effects. *J. exp. Psychol., Hum. Percep. Perf., 3(3)*, 422–435.

Posner, M.I. (1980) Orienting of attention. *Q. J. exp. Psychol., 32*, 3–25.

Posner, M.I., Nissen, M.J. and Ogden, W.C. (1978) Attended and unattended processing modes: the role of

set for spatial location. In Saltzman, I.J. and H.L. Pick (eds.), *Modes of Perceiving and Processing Information*. Hillsdale, NJ, Lawrence Erlbaum.

Potter, M.C. (1975) Meaning in visual search. *Science, 187*, 965–966.

Rayner, K. (1948) Eye movements in reading and information processing. *Psychol. Bull., 85(3)*, 618–660.

Regan, D. and Beverley, K.I. (1978) Looming detectors in the human visual pathway. *Vis. Res., 18*, 209–212.

Rezak, M. and Beneveneto, A. (1979) A comparison of the organization of the projections of the dorsal lateral geniculate nucleus, the inferior pulvinar and adjacent lateral pulvinar to primary visual area (area 17) in the macaque monkey. *Brain Res., 167*, 19–40.

Richards, W. (1982) How to play twenty questions with nature and win. *M.I.T.A.I. Laboratory Memo 660.*

Richmond, B.J. and Sato, T. (1982) Visual responses of inferior temporal neurons are modified by attention to different stimuli dimensions. *Soc. Neurosci. Abst., 8*, 812.

Riggs, L.A. (1965) Visual acuity. In C.H. Grahan (ed.), *Vision and Visual Perception*. New York, John Wiley.

Riley, M.D. (1981) The representation of image texture. M.Sc. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Robinson, D.L., Goldberg, M.G. and Staton, G.B. (1978) Parietal association cortex in the primate: sensory mechanisms and behavioral modulations. *J. Neurophysiol., 41(4)*, 910–932.

Rock, I., Halper, F. and Clayton, T. (1972) The perception and recognition of complex figures. *Cog. Psychol., 3*, 655–673.

Rock, I. and Gutman, D. (1981) The effect of inattention of form perception. *J. exp. Psychol.: Hum. Percep. Perf., 7(2)*, 275–285.

Rumelhart, D.E. (1970) A multicomponent theory of the perception of briefly exposed visual displays. *J. Math. Psychol., 7*, 191–218.

Schein, S.J., Marrocco, R.T. and De Monasterio, F.M. (1982) Is there a high concentration of color-selective cells in area V4 of monkey visual cortex? *J. Neurophysiol., 47(2)*, 193–213.

Shiffrin, R.M., McKay, D.P. and Shaffer, W.O. (1976) Attending to forty-nine spatial positions at once. *J. exp. Psychol.: Human Percep. Perf., 2(1)*, 14–22.

Shulman, G.L., Remington, R.W. and McLean, I.P. (1979) Moving attention through visual space. *J. exp. Psychol.: Huma. Percep. Perf., 5*, 522–526.

Sperling, G. (1960) The information available in brief visual presentations. *Psychol. Mono., 74*, (11, Whole No. 498).

Stevens, K.A. (1978) Computation of locally parallel structure. *Biol. Cybernet., 29*, 19–28.

Sutherland, N.S. (1968) Outline of a theory of the visual pattern recognition in animal and man. *Proc. Roy. Soc. Lond. B, 171*, 297–317.

Townsend, J.T., Taylor, S.G. and Brown, D.R. (1971) Latest masking for letters with unlimited viewing time. *Percep. Psycholphys., 10(5)*, 375–378.

Treisman, A. (1977) Focused attention in the perception and retrieval of multidimensional stimuli. *Percep. Psychophys., 22*, 1–11.

Treisman, A. and Celade, G. (1980) A feature integration theory of attention. *Cog. Psychol., 12*, 97–136.

Tsal, Y. (1983) Movements of attention across the visual field. *J. exp. Psychol.: Hum. Percep. Perf.* (In Press).

Ullman, S. (1979) *The Interpretation of Visual Motion*. Cambridge, MA, and London: The M.I.T. Press.

Varanese, J. (1983) Abstracting spatial relations from the visual world B.Sc. thesis in Neurobiology and Psychology. Harvard University.

van Voorhis, S. and Hillyard, S.A. (1977) Visual evoked potentials and selective attention to points in space. *Percep. Psychophys., 22(1)*, 54–62.

Winston, P.H. (1977) *Artificial Intelligence*. Reading, MA., Addison-Wesley.

Wurtz, R.H. and Mohler, C.W. (1976a) Organization of monkey superior colliculus: enhanced visual response of superficial layer cells. *J. Neurophysiol., 39(4)*, 745–765.

Wurtz, R.H. and Mohler, C.W. (1976b) Enhancement of visual response in monkey striate cortex and frontal eye fields. *J. Neurophysiol., 39*, 766–772.

Wurtz, R.H., Goldberg, M.E. and Robinson D.L. (1982) Brain mechanisms of visual attention. *Scient. Am., 246(6)*, 124–135.

Zeki, S.M. (1978a) Functional specialization in the visual cortex of the rhesus monkey. *Nature, 274*, 423–428.

Zeki, S.M. (1978b) Uniformity and diversity of structure and function in rhesus monkey prestriate visual cortex. *J. Physiol., 277*, 273–290.

Reference Note

1. Joliceur, P., Ullman, S. and Mackay, M. (1984) Boundary Tracing: a possible elementary operation in the perception of spatial relations. Submitted for publication.

Résumé

Cet article porte sur le traitement de l'information visuelle après la création des premières représentations. La capacité de déterminer visuellement les propriétés formelles absraites et les relations spatiales est un prérequis à ce niveau. Cette capacité joue un role majeur dans la reconnaissance d'objet, dans les manipulations guidées par la vision ainsi que dans la pensée visuelle plus abstraite.

Pour le système visuel humain, la perception des propriétés spatiales et des relations complexes au point de vue calcui apparait trompeusement immédiate et facile. L'efficacité du système humain pour analyser l'information spatiale surpasse de loin les capacités des systèmes artificiels utilisés pour l'analyse spatiale de l'information visuelle.

La perception des proprietés de forme abstraite et des relations spatiales soulève des difficultées fondamentales avec des conséquences importantes pour le traitement général de l'information visuelle. Les auteurs défendent l'idée que le calcul des relations spatiales sépare l'analyse de l'infformation visuelle en deux stades principaux. Au cours du premier se créent, de bas en haut, certaines représentations de l'environnement visible. Au cours du second des processus dits 'routines visuelles' s'appliquent aux représentations issues du premier stade. Ces routines peuvent revèler des propriétés et des relations qui n'étaient pas représentées de façon explicite dans les représentations initiales.

Les routines visuelles sont composées de séquences d'opérations élémentaires conjointes pour les différentes propriétés et relations. En utilisant une série fixe d'opérations de base, le système visuel peut assembler différentes routines pour extraire une suite illimitée de propriétés de forme et de relations spatiales.

A un niveau plus détaillé, on suggère un certain nombre d'opérations de base, en se fondant essentiellement sur leur utilité potentielle et, en partie, sur des preuves empiriques. Ces opérations incluent le changement du centre de traitement, l'indexation à une localisation d'un observateur extérieur, des' activations limitées, le tracage de frontières et des marquages. Les auteurs posent le problème de l'assemblage de ces opérations élémentaires en routines visuelles signifiantes.