"Visual Cognition and Visual
Routines", by Shimon Ullman
from "High Level Vision", Chap. 9
pp. 263-315
text published in 1996

NO DATE FOR CHAP 9

9 Visual Cognition and Visual Routines

The use of shape and spatial information is not limited to the tasks of object recognition and classification. We also use visual analysis of shape and spatial relations for other tasks, such as manipulating objects, planning and executing movements in the environment, selecting and following a path, and the like. In assembling objects and parts, for instance, we can use our visual perception to decide how different parts may fit together, and then plan and guide our movements accordingly. Problems of this type can be solved without implicating object recognition. They do require, however, the visual analysis of shape and spatial relations among parts. This visual analysis of shape properties and spatial relations is called here "visual cognition."

The visual extraction of spatial information is remarkably flexible and efficient. We can look for instance at an image such as figure 9.1 and obtain, almost at a glance, answers to a variety of possible questions regarding shape properties and spatial relations. For example, in figure 9.1a, the task is to determine visually whether the X lies inside or outside the closed curve. Observers can establish this relation effortlessly, unless the boundary becomes highly convoluted. The answer appears to simply "pop out," and we cannot give a full account of how the decision was reached. It is interesting to note that this capacity appears to be associated with relatively advanced visual systems. The pigeon, for example, that shows an impressive capacity for figure classification and recognition (Herrnstein 1984), is essentially unable to perform this task in a general manner. It can respond correctly only for simple figures, and appears to base its decision on simple local cues such as the convexity or concavity of the contour nearby (Herrnstein *et al.*, 1989).

Figure 9.1b is an example of establishing a shape property, in this case, judging the elongation of ellipse-like figures. (The terms "shape property" refer to a single item, while spatial relations, such as "above," "inside," "longer-than," and the like, involve two or more items.) The visual system is quite precise at detecting such elongations: reliable judgements of elongation can

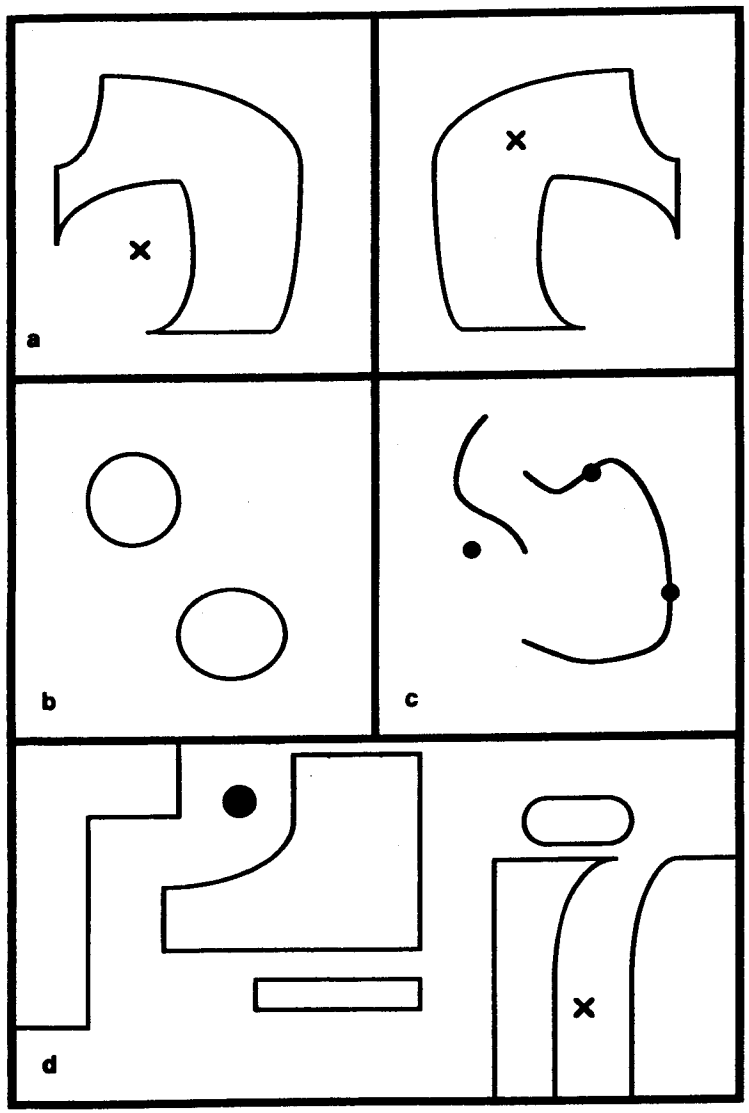

Figure 9.1
Examples of several "visual cognition" tasks involving the analysis of shape properties and spatial relations. (*a*) Inside/outside relation: it is easy to determine whether the "X" lies inside or outside the closed figures. (*b*) Elongation judgements. (*c*) The task is to determine whether two black dots lie on a common contour. (*d*) The task is to determine whether the black disk can be moved to the location of the "X" without colliding with nearby shapes.

be obtained when the major axis of the ellipse is 4 – 5% longer than the minor axes (Cave 1983). Interestingly, the judgements become more difficult when the axes themselves, without the ellipses, are present. This suggests that the judgement is not based on the extraction and comparison of the main axes, and it remains unclear what mechanisms in fact subserve this and related shape judgements. In figure 9.1c the task is to determine whether two black dots lie on a common contour. Again, a solution is obtained by "merely looking" at the figure. In figure 9.1d the task is more complex—to determine whether the black disk can be moved to the location of the *X* without colliding with any of the other shapes. We can use our visual capacities to somehow "simulate" the motion and obtain the correct answer.

These figures and tasks are artificial, but similar visual cognition problems also occur in natural settings, in the course of object manipulation, planning actions, reasoning about objects in the scene, navigation, and the like, and are solved routinely by the visual system. We also make use of visual aids such as diagrams, charts, sketches and maps, because they draw on the system's natural capacity to manipulate and analyze spatial information, and this ability can be used to help our reasoning and decision processes.

Spatial analysis of this kind does not require object recognition, that is, it does not depend on object naming or on whether we have seen the objects in the past. It does require, however, the analysis of shape and spatial relations among shapes. The visual system can perform a wide variety of visual cognition tasks with remarkable proficiency, that cannot be mimicked at present by artificial computer vision systems.

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. The apparent immediateness and ease of perceiving spatial relations is, however, deceiving: it conceals in fact a complex array of processes that have evolved to establish certain spatial relations with considerable efficiency. This makes the problems of visual cognition intriguing and challenging: What mechanisms and processes in human vision are responsible for these computations? How can machines achieve similar capacities?

The mechanisms underlying such visual cognition are still ill-understood. It is still unclear what internal shape processes are used, even in simple configurations, such as comparing the length of two line segments. How is such a comparison actually accomplished? Can the system somehow apply an "internal yardstick" to the segments, shift them internally to test their overlap, or use some other method? The answer to such questions remains at present unknown. I will discuss here a possible approach, based on the notion of visual routines (Ullman 1984). Let me summarize first the basic idea, and then discuss the approach in more detail.

The approach assumes that the perception of shape properties and spatial relations is achieved by the application of so-called "visual routines" to the early visual representations. These visual routines are efficient sequences of basic operations that are "wired into" the visual system. Routines for different properties and relations are then composed from the same set of basic operations, using different sequences. Using a fixed set of basic operations, the visual system can assemble different routines and in this manner extract an essentially unbounded variety of shape properties and spatial relations. Within this framework, to understand visual cognition in general, it will be required to identify the set of basic operations used by the visual system. An explanation of how we determine a particular relation such as "above," "inside," "longer-than" or "touching," would require a specification of the visual routine used to extract the shape or property in question.

9.1 Perceiving "Inside" and "Outside"

An Example: Perceiving "Inside" and "Outside" It will be useful to examine in more detail a specific example of our perception of spatial relations in a scene: the example will make clear that our effortless perception of shape properties and spatial relations is in fact a complex process, whose exact nature remains at present quite mysterious. The example will also serve to introduce the notion of visual routines, and how they may be applied to the extraction of abstract shape properties and spatial relations.

To consider a concrete example, let us assume that the visual input consists of a single closed curve, and a small "X" figure (as in figure 9.1a), and one is required to determine visually whether the X lies inside or outside the closed curve. The correct answer appears to be immediate and effortless, and the response is usually fast and accurate (Varanese 1983). One possible reason for our proficiency in establishing inside/outside relations is their potential value in figure-ground segmentation: if the bounding contour of an object has been identified, features inside the contour belong to the objects, and features outside it to the surround (see also Sutherland (1968) and Kovás & Julesz (1994) on inside/outside relations in perception).

The immediate perception of the inside/outside relation is subject to some limitations: when the bounding contour becomes highly convoluted, the distinction between inside and outside becomes more difficult. 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. In mathematics, the "Jordan curve theorem" states that a simple closed plane curve separates the plane into two disjoint regions, its inside and outside. It may appear surprising that this is a theorem that requires an elaborate proof, since the concepts of inside and outside are intuitively clear to us, and we "see" that

the theorem must be true. This is probably based on our visual cognition capacity to deal effectively with spatial analysis.

The processes underlying the perception of inside/outside relations are as yet unknown. In the following section I will 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.

9.1.1 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 features such as short edge and line segments are detected early in the chain of visual processing by specially constructed feature-detecting units. These feature detectors are then combined to produce higher order units such as, say, corner and triangle detectors, leading eventually to the detection and recognition of complete objects. It does not seem possible, however, to construct in such a manner 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 tangent 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 figure 9.2*a*, *b*).

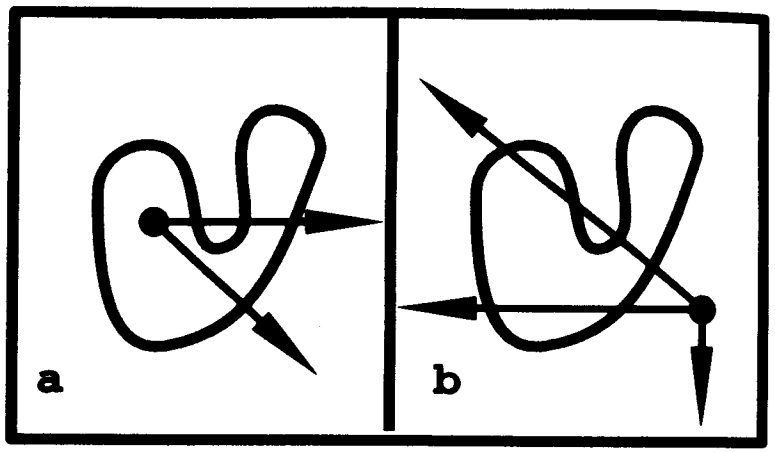

Figure 9.2
The ray-intersection method for establishing inside/outside relations. When the point lies inside the closed curve, the number of intersections is odd (*a*); when it lies outside, the number of intersections is even (*b*).

This procedure has been implemented in computer programs (Evans 1968, Winston 1977, chapter 2), and it may appear rather simple and straightforward. The success of the ray-intersection method is guaranteed, however, only if 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 figure 9.3*a*). Second, it must be assumed that the curve is isolated: in figure 9.3*b* and *c*, point *p* lies within the region bounded by the closed curve *c*, but the number of intersections is even.

These limitations of 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 figure 9.3*d* 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 figure 9.3*e* and *f* 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 figure 9.3*g* 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.

9.1.2 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.

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 termi-

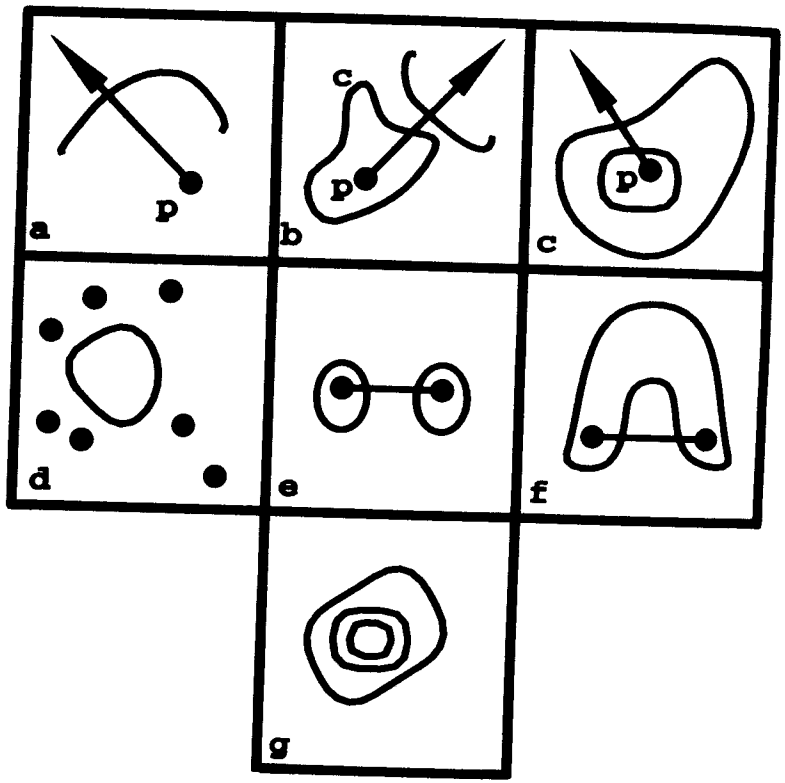

Figure 9.3
Limitations of the ray-intersection method. *a*. An open curve: the number of intersections is odd, but p does not lie inside C. *b, c*. Additional curves may change the number of intersections, leading to errors. *d* – *g*. Variations of the inside/outside problem that render the ray-intersection method ineffective. In *d* the task is to determine visually whether any of the dots lie inside C; in *e, f*, whether the two dots lie inside the same curve; in *g* to find a point that lies inside all three curves.

nation 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. A similar procedure can be used to solve some of the other problems. For example, to test whether two locations lie inside the same enclosing curve, as in 9.3*e*, *f*, the activation will start at one location, and its spread to the second location will be tested.

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 is, the size of the figures can be increased considerably with only a small effect on the computation time (Varanese 1983). In contrast, in the activation scheme outlined above, computation time will 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 (Jolicoeur, Ullman & MacKay 1991). 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.

It is interesting to note, with respect to coloring, that a fast coloring operation for the purpose of visual processing was implemented in an analog VLSI device developed by Luo, Koch, and Mathur (1992). This "figure-ground" chip labels all the points inside a contour with one voltage value and all the points outside the contour with a different voltage. The basic idea is to simply let the voltage spread in a resistive network, and separate electrically the inside of a region from its outside. The structure

that carries out this operation is constructed from a regular grid of points connected by resistors and switches. The presence of an edge between two grid points causes the switch at the corresponding location to open, and in this manner the inside and the outside of the contour become disconnected. (A mechanism is also incorporated to complete small breaks in the contour.) When the voltage at a selected point is set to a relatively high "figure" value, the voltage at all connected points will rise to this value, while the background points will remain at the lower "background" value.

The goal of the above discussion was not to examine the perception of inside/outside relations in particular, but to introduce the problems associated with the seemingly effortless and immediate perception of spatial relations. The main conclusions from the discussion are, first, that the efficient and flexible computation of spatial relations is a complex task, and, second, that a natural approach to the problem involves the internal application of a sequential process, somewhat similar to a computer program, to the image representation.

The discussion so far used a specific example—perceiving "inside" and "outside" to raise some of the main issues and possible directions. 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.

9.2 Spatial Analysis by Visual Routines

In this section we will 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 met by the "visual processor" that performs such an 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).

9.2.1 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.

Intuitively, the concept of being "inside" is abstract, because it does not refer to any particular shape, but can appear in many different forms. More 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 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 simply comparing it against all the members of S stored in memory. To be more accurate, 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.

An observation by Sutherland (1960, 1968) on pattern recognition in simple animals can help to illustrate this distinction. In perceptual experiments, octopuses were trained to successfully distinguish squares from diamonds of different sizes and locations. As it turned out, however, the animals then responded to triangles as equivalent to diamonds. Apparently, what they actually used to make the distinction was the property of having a pointy top, while ignoring the rest of the shape. This property of having a pointy top depends on a small local region (at least for convex shapes), and it can be tested without analyzing the entire shape.

Without getting into detailed definitions, it is clear that because of the restricted support such properties are easier to compute.

When the set of supports is small, the recognition of even a large set of objects can still be accomplished by simple means such as direct template matching. This means that a small number of patterns is stored, and matched against the figure in question. When the set of supports is prohibitively large, a template matching decision scheme will become impossible: we cannot store, for example, all instances of closed curves in the image. The classification task may nevertheless be feasible if the set of shapes sharing the property in question 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 closure can obviously be established without storing all possible instances of closed plane curves. 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 in this case be essentially as large as the set itself (c.f. Kolmogorov 1968).

The above discussion can 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. For example, closure is an abstract property because it must be established by some process that makes use of general characteristics of closed curves. Our visual system can clearly establish abstract properties and relations of this type. 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 underly-

ing 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 already 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 above. To explain the perception of inside/outside relations it would be necessary, therefore, to unravel the constituent operations that are actually employed by the visual system, and how they are used in different situations.

9.2.2 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 establish 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," already mentioned above, 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.

9.2.3 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 detail below) implies that different 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 is determined in the primary visual cortex in our brain largely independent of other locations: the machinery for detecting a line segment of a particular orientation is duplicated many times within this visual area. 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 requirements discussed above of abstractness, open-endedness, and complexity, suggest the following conclusions:

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.

9.3 Conclusions and Open Problems

The discussion so far suggests that the immediate perception of seemingly simple spatial relations often requires in fact complex computations that are difficult to unravel. The general proposal is that using a fixed set of basic operations, the visual system can assemble visual 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 raise fundamental problems at the levels of computational theory, algorithms, and the underlying mechanisms. A general problem on the computational level is to establish which spatial properties and relations are important for different visual tasks. On the algorithmic level, the problems are how these relations are computed, and what would be a useful complete set of basic operations. These are challenging problems, since the

processing of spatial relations and properties by the visual system is remarkably flexible and efficient. When these algorithmic issues become better understood, it will also become possible to consider the construction of a "routine processor" with similar capabilities for practical use in machine vision applications. On the mechanism level, the problem is to find out how visual routines are implemented in neural networks within the visual system. To conclude this section, the main problems raised by the notion of visual routines are listed below, divided into four main categories.

- *The elemental operations.* In the examples discussed above the computation of inside/outside relations employed operations such as drawing a ray, counting intersections, boundary tracing, 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.

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.

- *Integration.* The second problem 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, and the like, 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.

- *Controlling the routines.* The questions in this category are how visual routines are selected and controlled; what triggers the

execution of different routines during the performance of visual tasks, and how the order of their execution is determined.

• *Compilation of new routines.* We have seen already that visual routines can be applied in a flexible manner that depends on the task, as well as the properties of the scene being analyzed. Problems that naturally arise, therefore, are how new routines are generated to meet specific needs, and how they are stored and modified with practice. These are interesting problems, but they will not be discussed in detail in this chapter.

In the next section, I will turn to the first of these issues, that is, the elemental operations used by visual routines.

9.4 The Elemental Operations

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 approaches can be followed. The first is the use of empirical psychological and physiological evidence. The second is computational: one can examine the kind 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 the human system in performance 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 will use below employ mainly schematic drawings rather than natural scenes. The reason is that simplified artificial figures allow more flexibility in adapting the pattern to the operation under investigation. As long 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 we use simplified drawings rather than natural scenes. In fact, our 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, by comparing the complexity of various visual tasks in the model and in human vision. 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 above 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 particular method by the human perceptual system can be ruled out. In this case, the argument is also supported by theoretical considerations showing 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 will examine a number of plausible candidates for basic operations, discuss some 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.

9.4.1 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 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 (Norton & Stark 1971). But this is clearly insufficient: many shape properties and relations 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.

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 locations in the visual input. They are listed below under three main classes.

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 & 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 milliseconds at three degrees from the fovea center. Such a result may reflect

the effect of shift time, but, as pointed out by Eriksen & Schultz, alternative explanations are also possible, in particular, that the processing time in general increases with increased distance from the center of the visual field. More direct evidence comes from a study by Posner, Nissen & Ogden (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 cuing (an arrow at fixation), then their reaction time to the target in both detection and identification tasks was consistently lower (without eye movements). For simple detection tasks, the gain in detection time for a target at seven degrees eccentricity was on the order of 30 milliseconds.

A related study by Tsal (1983) employed peripheral rather than central cuing. 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 milliseconds, then by the time the letter appears the shift operation is already k milliseconds 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 precisely what the experimental results indicated. It was further found that the delay at which facilitation levels off (presumably the total shift time) increases with eccentricity, by about eight milliseconds on average per one degree 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, Remington & Mclean (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 eight degrees from the fixation point 50–100 milliseconds after the location had 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 (Sperling 1960, Shiffrin, McKay & Shaffer 1976). In such an experiment, an observer is briefly presented with a complex visual stimulus, for example, a collection of numbers arranged in a square matrix. Under such conditions, the observer can recall information from the display only in a very partial manner. If, following the visual presentation, the observers are directed to a particular location in the display, for instance, the third number in the second row, the recollection is usually close to perfect. 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 (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 a 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 above 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 requirements of the routine in question. Therefore, as far as the shift operation is concerned, it is of interest to demonstrate phenomena of attention shift under conditions that do not overwhelm the capacity of the visual system.

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 by Posner (1980) and by Shulman, Remington & Mclean (1979) suggest that it can be "programmed" to move along a straight path using central cuing. In other experiments, such as Remington's and Tsal's, the processing focus is shifted by being attracted to a peripheral cue.

Physiological Evidence Shift-related mechanisms have been explored in the monkey physiologically in the superior colliculus, and in a number of cortical areas: the posterior parietal lobe (area 7) the frontal eye fields, visual areas V1, V2, V4, MT, MST, and the inferior temporal lobe. The general idea of these experiments has been to study whether the response of units in the visual system can be modified by somehow manipulating the experimental animal's attention to different locations in the visual field. It is

of interest to examine these studies in connection with the shift operation for two reasons: first, to examine whether there is physiological evidence in support of the proposed shift operation, and second, to try to find out which brain area may be the "controller" that directs the location of processing.

In the superficial layers of the superior colliculus of the monkey, a sub-cortical structure associated with the control of eye movements, many cells have been found to have an enhanced response to a stimulus when the monkey uses the stimulus as a target for a subsequent saccadic eye movement (Goldberg & 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 & Mohler 1976s). 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 & Wurtz 1972). The intimate relation of the enhancement to eye movements, and its absence when the saccade is replaced by other responses (Wurtz & Mohler 1976a, Wurtz, Goldberg & Robinson 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 & Mohler 1976) and in prestriate cortex, probably area V4 (Fischer & Boch 1981).

Another area that exhibits a similar phenomenon of a facilitated region that can be shifted around, but not exclusively in relation 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 & Wurtz 1982.) Studies by Robinson, Goldberg & Stanton (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 Robinson *et al.* 1978, Posner 1980, Wurtz, Goldberg & Robinson 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 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.

In area TE of the inferotemporal cortex units were found whose responses depend strongly upon the visual task performed by the animal. Fuster & Jervey (1981) described units that responded strongly to the stimulus' color, but only when color was the relevant parameter in a matching task. Richmond & 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 (for instance, when the stimulus was monitored to detect its dimming). Again, it has been suggested that such responses may be linked in general to some form of selective visual attention, but they are not directly implicated in an internal shifting operation.

An elegant experiment concerning physiological mechanisms of selective visual attention is the study of Moran & Desimone (1985) in visual areas V4 and IT of the monkey. In this experiment, two different stimuli were presented simultaneously within the receptive field of a neuron. One was an effective stimulus, the other ineffective. For example, if the unit responded selectively to a red vertical bar, such a bar was used as the effective stimulus,

and a horizontal green bar could be used as the ineffective stimulus. By using a shape matching task, the animal's attention was drawn to the effective stimulus on some trials, and to the ineffective stimulus on others. The main finding was that although the physical stimulation was identical in the two conditions, the response was considerably reduced when attention was drawn to the ineffective stimulus. In the initial studies, these attentional effects were found only when the two stimuli were both within the receptive field of the recorded unit, but further studies found similar effects for stimuli with larger separations as well (Luck *et al.* 1992). These results are consistent with the possibility that the recorded units are related to the direction of visual processing to different locations. Van Essen and his collaborators (Connor, Gallant & Van Essen 1993) have tested this possibility further by directing the monkey's attention to different locations, and at the same time plotting the sensitivity profile of the receptive field to a small bar stimulus. They found that in many cases the center of gravity of the receptive field in fact shifted, in the direction of the attentional cue. They have also developed a biological model by which shifts of the center of processing can be obtained (Anderson & Van Essen 1987).

Finally, responses in the pulvinar (Gattas *et al.* 1979, Desimone *et al.* 1990) were shown to be strongly modulated by attentional and situational variables. Combined with the extensive connectivity of the pulvinar to multiple visual areas in the cortex, this led to the suggestion that the pulvinar may be (either by itself or in combination with other structures) the "controller" of location in directing visual attention.

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 by instructing the subject to "attend" to different spatial locations (e.g., van Voorhis & 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.

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: their response depends on whether the task requires shape or color discrimination. In contrast, units in area V1 for example, appear in several studies to be task-independent. Task-dependent responses suggest that the units do not belong to the purely 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. At the same time, 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 & Jervey (1981) showed task-dependent response (they 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, whether 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 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 & Davis 1977, Rezak & Beneveneto 1979, Robinson & Petersen 1992).

Selecting a Location Computational considerations strongly suggest the use of internal shifts of the processing focus, and this notion is supported by psychological evidence, and to some degree by physiological data. A related issue to consider is 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: "move

to location x = 5cm, y = 8 cm from the bottom-left corner of the screen." 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 mode of operation which seems to be used by the visual system in shifting the processing focus. This is based on the extraction of certain salient locations in the image, and then shifting the processing focus to one of these distinguished locations. The salient locations are detected in parallel across the base representations, and can then serve as "anchor points" for the application of visual routines. As an example, 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 letter proceeds in several stages. 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. As a result of this stage, visual routines can be applied to the figure. By applying the appropriate routines, the figure is identified. A similar process also played a role in the inside/outside example above. 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 is somehow sufficiently salient, it can serve to attract the processing focus, and then the execution of the appropriate routine can start immediately at that location.

To conclude, one way in which the focus of processing can be manipulated is by moving it to a salient location in the scene. A question that naturally arises at this point, is: what defines a

distinguished location, that can be used for the purpose of shifting the processing focus and applying further operations?

From psychophysical studies it appears that certain odd-man-out locations that are sufficiently different from their surroundings can attract the processing focus directly, and eliminate the need for lengthy scanning. For example, differences in orientation and direction of motion can be used for this purpose, while more complex distinctions, such as the occurrence of the letter "A" among similar letters, cannot define a distinguished location.

By using visual search and other techniques, Treisman and her collaborators (Treisman 1977, Treisman & Gelade 1980, see also Beck & Ambler 1972, 1973, Pomerantz *et al.* 1977) have shown that color and simple shape parameters can define distinguished locations. 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 distinguished 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 distinguished by certain shape parameters, such as orientation and curvature.

The notion of a 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 support 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, color, and binocular disparity, were found to be extracted in parallel by units that cover the visual field. These units appear to be driven in a bottom-up manner, and their responses are almost unchanging when the animal is awake, anesthetized, or naturally

sleeping (Livingston & Hubel, 1981). On physiological grounds these properties are suitable, therefore, for defining distinguished locations prior to the application of visual routines.

It is interesting to note that apparently only simple differences in these early-computed properties can be used to define distinguished locations, prior to the application of visual routines. For example, several studies by Treisman and her collaborators examined the problem of whether different properties measured at a given location can be combined to define a distinguished odd-man-out location. 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 distinguished if it has a unique color or shape. The question now was whether the conjunction of two such properties is also immediately distinguished. The empirical evidence indicates that items cannot be immediately distinguished 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 in such studies (Treisman & Gelade 1980) were consistent with a serial self-terminating search in which the items are examined sequentially until the target is reached.

In summary, one way of shifting the processing focus around in the course of applying visual routines is by first extracting a set of distinguished locations in the scene, and then shifting the processing focus towards one of these locations. In defining the distinguished locations, a small number of elementary properties such as orientation, contrast, color, motion, binocular disparity, and perhaps a few others, are computed in parallel across the early visual representations, prior to the application of visual routines. Simple differences in these properties can then be used to define distinguished locations. These locations can then be used in visual routines by moving the processing focus directly to one of the

distinguished locations, without the need for extensive search or systematic scan.

9.4.2 Bounded Activation (Coloring) and the Incremental Representations

The bounded activation, or "coloring" operation, was suggested above in examining the inside-outside relation. It consisted of the spread of activation over a surface in the visual representation emanating from a given location or contour, and stopping at discontinuity boundaries. I will discuss below the coloring operation together with some general notion of creating "incremental representations" for subsequent processing.

The results of the coloring operation may be retained for further use by additional routines. Coloring provides in this manner one method for defining larger units in the initial visual representations: the "colored" region becomes a unit to which routines can be applied selectively. An example along this line is illustrated in figure 9.4*a*. The visual task here is to identify the subfigure marked by the black dot. One may have the subjective feeling of being able to concentrate on this subfigure, and "pull it out" from its complicated background. It is easily seen in the figure that the marked subfigure has the shape of the letter G. The area surrounding the subfigure in close proximity contains a myriad of irrelevant features, and therefore identification would be difficult, unless processing can be directed to this subfigure. The suggestion is, then, that the figure is first separated from its surroundings by using the area activation operation. Recognition routines could then concentrate on the activated region, ignoring the irrelevant contours. This example uses an artificial stimulus, but the ability to identify a region and process it selectively is equally useful for the recognition of objects in natural scenes.

The use of coloring to define a region of interest to which subsequent processing can be applied provides an example of the distinction between the early visual representations (also called "base representations"), and the subsequent, or "incremental" represen-

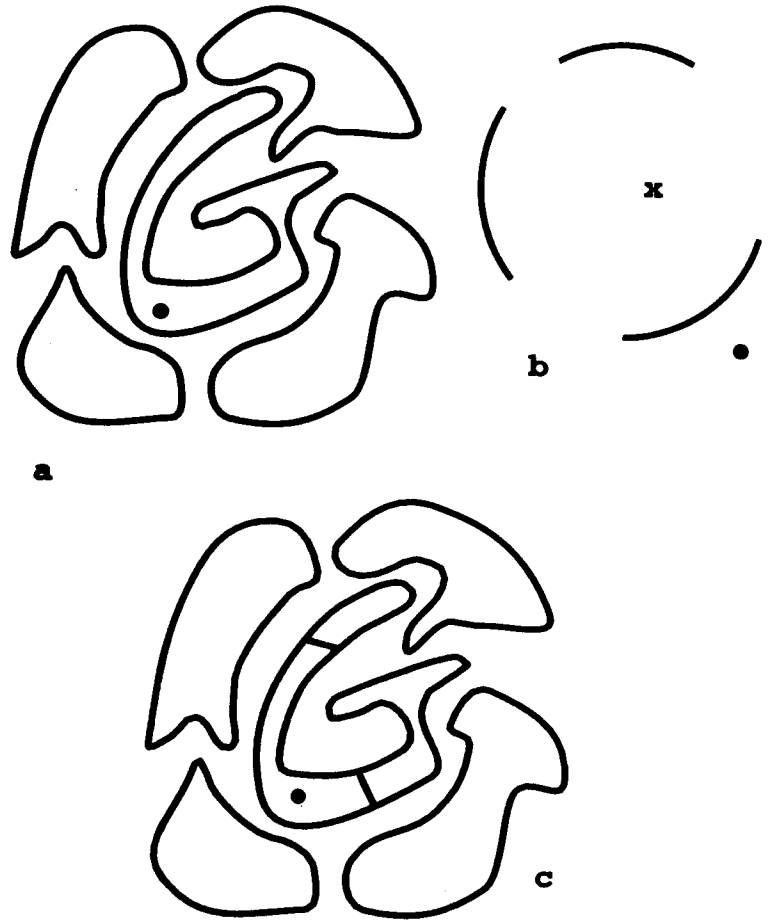

Figure 9.4
Examples of the "coloring" operation. In *a*, the visual task is to identify the subfigure containing the black dot. This figure (the letter "G") can be recognized despite the presence of confounding features in close proximity to its contours. The capacity to "pull out" the figure from the irrelevant background may involve the bounded activation operation. In *b* the boundaries are fragmented: the curve is defined by a dashed line, but inside/outside judgements are still immediate. In *c*, 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.

tations (Ullman 1984). The earlier representations are produced prior to the application of visual routines. They are produced in an unguided bottom-up manner, determined by the visual input and not by the goal of the processing. They are also spatially uniform, in the sense that, with the exception of a scaling factor, similar processes are applied across the visual field, or large parts of it. Examples of early visual representations include the extraction of edges and lines from the image, the computation of motion, disparity, and color.

In contrast with these early processes, the application of visual routines, and the properties and relations they extract, are not determined by the input alone. For the same visual input different aspects will be made explicit at different times, depending on the goals of the computation. Unlike the base representations, the computations by visual routines are not applied uniformly over the visual field (for example, not all of the possible inside/outside relations in the scene are computed), but only to selected objects. Another distinction between the two stages is that the construction of the early representations is essentially fixed and unchanging, while visual routines are open-ended and permit the extraction of newly defined properties and relations.

For various visual tasks, the analysis of visual information therefore 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 only in the image but also in three-dimensional space. Many spatial judgements we make naturally depend in fact primarily on three-dimensional relations rather than on projected, two-dimensional ones (see, for example, Joynson & Kirk 1960, Lappin & 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. Since the base representations already contain three-dimensional information, the visual routines applied to them can also establish such properties and relations in three-dimensional space.

Discontinuity Boundaries for Coloring The activation operation is supposed to spread until a discontinuity boundary is reached. This raises the question of what constitutes a discontinuity boundary for the activation operation. In figure 9.4*a*, lines in the two-dimensional drawing served for this purpose. In more natural scenes, 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 figure 9.4*b*, 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. If activation is to be used in this situation as well, then incomplete boundaries should have the capacity to block the activation spread. It is interesting to note that inside/outside judgements using dashed boundaries appear to require somewhat longer times compared with continuous curves, suggesting that fragmented boundaries may indeed require additional processing (Varanese 1983).

Finally, the activation is sometimes required to spread across certain boundaries. For example, in figure 9.4*c*, which is similar to figure 9.4*a*, the letter G is still recognizable, in spite of the internal bounding contours. To allow the coloring of the entire subfigure in this case, the activation must spread across internal boundaries.

In conclusion, the bounded activation, and in particular, its interactions with different contours, is not a simple 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.

A Possible Mechanism for Bounded Activation 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. A second network containing a map of the discontinuity boundaries can 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 such a multi-scale scheme, the time to color a shape becomes almost independent of the shape's size. Such schemes were developed and implemented by Shafrir (1985) and by Mahoney and Ullman (1988).

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 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.

We still know very little about the kind of computations that are taking place in the visual system in the course of performing visual cognition tasks, but an interesting point to consider is that if such networks for executing and controlling basic operations are in fact 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. The testing and interpretation of units in these networks would depend not on finding the optimal stimulus conditions for the unit, but rather on the visual tasks that cause the unit to be active. Elucidating the basic operations would be helpful in providing clues for understanding the activity in such networks and their patterns of interconnections.

9.4.3 Boundary Tracing

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 an internal visual representation. 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 & 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 the traced contour.

Another simple example which illustrates the role of boundary tracing is shown in figure 9.5*a*. The question here is whether

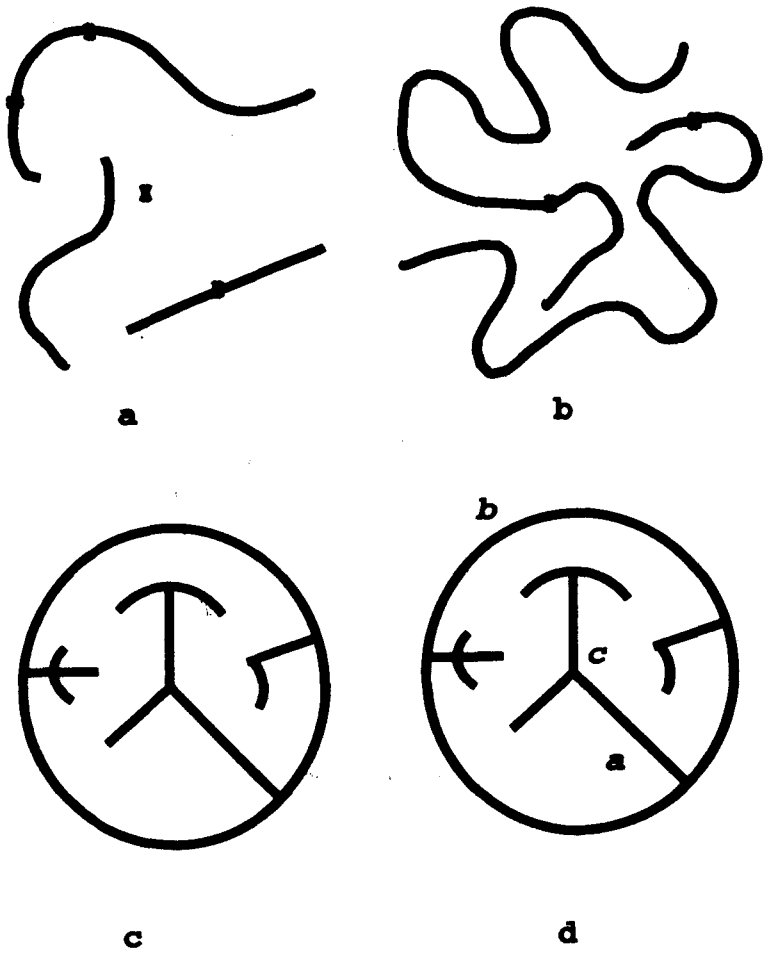

Figure 9.5
Examples of the boundary tracing operation. In *a*, the task is to determine visually whether two small X's lie on the same curve. This simple task requires in fact the application of a visual routine that is likely to include the use of a tracing operation. In *b* the task is similar, but the figure, after the study of Jolicoeur, Ullman & Mackay (1986), was designed to change the separation of the two X's when they appeared on the same curve, while keeping their direct distance fixed. In *c*, the task is to determine visually whether there is a part connecting the center of the figure to the surrounding circle. In *d* the solution is labeled. The interpretation of such labels also relies upon the use of common natural visual routines.

there are two small X's lying on a common curve. The answer seems immediate and effortless, but how is it achieved? Clearly, it cannot be mediated by a fixed array of two-X's-on-a-curve detectors. This simple perception conceals in fact a more elaborate chain of events that involves the application of a visual routine. In response to the question, a routine has been compiled and executed. An appropriate routine for this task can be constructed if the repertoire of basic operations included a shift to the X's and the tracking of curves. The tracking provides in this task a useful role of integrating information from different parts of the same contour, in the presence of other contours nearby.

Another example of the possible use of boundary tracing goes back to the inside/outside example discussed above. Tracking can be used in conjunction with the area activation operation to establish inside/outside relations, by moving along a boundary, coloring only one side. If the curve is closed, its inside and outside will be separated. Otherwise, the fact that the curve is open will be established by the coloring spread, and by reaching a termination point while tracking the boundary.

The possible use of an internal tracing process has been investigated in a number of studies (Jolicoeur & Ingleton 1991, Jolicoeur, Ullman & Mackay 1986, 1991, McCormick & Jolicoeur 1991, 1992, Pringle & Egeth 1988), and the results clearly support the use of a contour tracing operation. For example, in a study by Jolicoeur, Ullman & Mackay (1986), subjects were presented with stimuli composed of two separate curves. In all trials there was a small 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 the same, (1.8 degree 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 degrees 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, proceeding along the curve from the location of the first X to the second. The short presentation time (250 milliseconds) precluded the tracing of the curve using eye movements, hence the tracing operation must have been performed internally. Similar results were obtained for a range of different stimuli, including simple configurations such as isolated arcs (Pringle & Egeth 1988) and straight lines (Jolicoeur, Ullman & Mackay 1991).

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 milliseconds 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.

Other studies have revealed additional interesting properties of the tracing operation. In a study by Jolicoeur, Ullman & Mackay (1991), it was found that the speed of tracing depended on properties of the curve, in particular its curvature and the proximity to other contours in the scene. The average speed of tracing in these experiments was about 40 degrees of visual angle, measured along the curve, per second, but it could be higher or lower, depending upon the conditions in the scene (Pringle & Egeth 1988). Tracing speed was highest for straight contours, and decreased systematically with the contour's curvature. The tracing speed was also higher when the curve was isolated in the visual field, free of nearby clutter, and slowed down in the presence of nearby contours. Finally, it turns out that within a broad range the tracing time is invariant to the absolute size of the pattern (Jolicoeur & Ingleton 1991, Pringle & Egeth 1988). If the entire display is simply scaled by a factor of, say, two, the tracing time does not double, but remains essentially unchanged. As mentioned above, the tracing time does depend on the curvature and proximity to other curves, but the relevant factors are the relative rather than the absolute distances. If the curve length is doubled, but at the same time all the other parameters, such as the distances to nearby curves, are also doubled, then the tracing time will not be affected.

It seems from these studies that internal boundary tracing is a fairly sophisticated operation, that attempts to perform the tracing as efficiently as possible. Perhaps the simplest implementation of a boundary tracing operation would be to march in a fixed step-size from one point to the next along the curve. Such a process will be insensitive, however, to parameters such as curvature and proximity to other curves. A more efficient process would adjust itself to the properties of the scene. One could use a coarser tracing mechanism for relatively straight, isolated contours, and a finer mechanism for a highly curved contour or a cluttered environment. This could be likened to the use of an adjustable beam moving along the curve. When the traced curve is relatively straight and isolated, one could use a larger beam, moving in larger steps along the curve. In the presence of nearby clutter, for instance, a smaller beam, moving in smaller steps, will be used (see Jolicoeur, Ullman & Mackay, 1991, McCormick & Jolicoeur 1991, for further discussion). Mahoney and Ullman (1988) proposed a similar mechanism, in which the curve is first divided into "chunks" of optimal size, and curve tracing subsequently proceeds along a succession of such chunks. If internal tracing is indeed one of a small set of basic operations, it is perhaps not surprising to find out that a fairly efficient and sophisticated mechanism is used for the task.

It is also interesting to note that when the same task employed colored curves, such as a green and a red one, the distance effect, and the effects of curvature and proximity, all but disappeared. The task could now be solved by using a simplified strategy, of checking whether the two X's lie locally on a curve of the same

color. It appears that the visual system makes use of this shortcut, and adjusts the strategy used, even without deliberate or conscious planning, to the available information.

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 figure 9.5*c*, 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 easily obtained by looking at the figure, but again, it must involve in fact a complicated chain of processing. To cope with this seemingly simple problem, visual routines 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.) In 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 figure 9.5*d*, 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 graphical material. To be unambiguous, such notations must rely 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.

Finally, it should be noted that the examples illustrated above used contours in schematic line drawings. However, if boundary tracking is indeed a basic operation in establishing properties and spatial relations, it is expected to be applicable not only to such contours, but also to the different types of contours and discontinuity boundaries in the early representations, such as boundaries defined by discontinuity in depth, motion, and texture. It should also be able to deal with incomplete boundaries, and to cope with the presence of intersections and branching points along the contour.

In summary, the tracing and activation of boundaries are useful operations in the analysis of shape and the establishment of spatial relations. Psychophysical studies provide strong support for the employment of such internal tracing by the visual system. This is a complicated operation since flexible, reliable tracing should be able to cope with breaks, crossings, and branching, and with different resolution requirements.

9.4.4 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 visited.

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 9.6a 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, one can use the above strategy of marking the X and tracking the curve. When the tracing is completed, we know that we have reached the same X, as opposed to a second one. Again, the problem cannot be solved by pre-existing detectors specialized for the task. Instead it is suggested that this simple perception of the X on the curve involved the application of visual routines that employ operations such as marking and tracing.

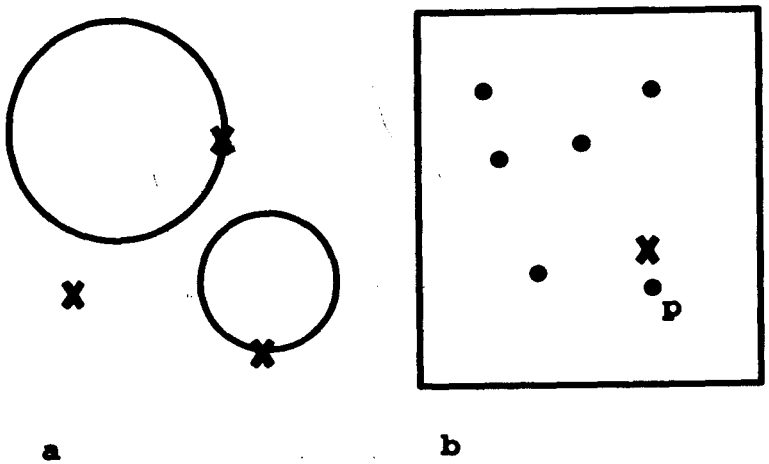

Figure 9.6
The use of marking. The task in *a* is to determine visually whether there are two *X*'s on a common curve. The task could be accomplished by employing marking and tracing operations. *b*. The use of an external reference: the position of point *p* can be defined and retained relative to the predominant *X* nearby.

Many other tasks may benefit from the marking of more than a single location (Pylyshyn 1988). A simple example is visual counting, that is, determining as fast as possible the number of distinct items in view (Atkinson, Campbell & Francis 1969, Kowler & 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 conjectures 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, Campbell & Francis 1969).

It is in fact possible, in principle, to construct special counting mechanisms of this type. For example, in their book "Perceptrons," Minsky and Papert (1969, chapter 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 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 by itself, responding directly to the presence of M items in view (Ullman 1980). Unlike the shifting and marking operations, the computation is performed by these networks uniformly and in parallel over the entire field.

Counting can also be performed, however, not by elaborate networks, constructed specifically for this task, but by simple visual routines that employ elementary operations such as shifting and marking. It could be achieved by shifting the processing focus among the items of interest without scanning the entire image systematically. In more complicated displays, shifting and marking can also be used for visual counting by scanning the entire scene in a fixed predetermined pattern.

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 perceptron-like networks, the process of determining the number of items by visual routines can be decomposed into a sequence of elementary operations. The problem of decomposing perceptual processes into simpler components is an interesting issue that lies at the heart of the controversy concerning Gibson's notion of "direct perception" (Ullman 1980). 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. Although we do not know how counting of this type is in fact achieved by the visual system, counting by visual routines appears 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 by itself an elementary operation. It is more plausible 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.

Marking and the Integration of Information in a Scene

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 position of point q in figure 9.6*b* 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.

To be useful in the natural analysis of visual scenes, the marking map should also 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. 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. Roughly speaking, 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 (Rayner 1978). 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 objects, and perhaps even to locations outside the field of view (Attneave & Pierce 1978). If we are currently looking at the dog, we would see it in fine 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.

We have examined above a number of plausible elemental operations including shift, bounded activation, boundary tracing, 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 a comprehensive set, but it appears that a small set of operations will still be sufficient to perform rather elaborate visual computations.

Visual routines that are much more complex than the simple tasks used in the discussion above have been used in computational studies by Chapman (1991, 1992) in the context of a sophisticated computer system that interprets its environment visually in the course of playing an interactive video game. The computer program developed by Chapman simulates the activities of a human player of the game. The program is presented with a screen of the kind used in many computer games. The screen is in fact not presented visually, but its content is made available to the computer program. The task of the program is to produce commands for moving the game's characters on the screen and for producing various actions, such as picking up objects or shooting at an opponent. Producing the appropriate commands at any given instant in the game requires the analysis of the visual display to determine, for instance, a path to a desired target, obstacles

to avoid, directions of potential threats, possible hiding places, and the like. This is performed by the program by applying appropriate visual routines, determined by the various tasks, to an internal image of the changing game display. Although the tasks were varied and sometimes quite complex, different combinations of a small and fixed set of basic operation proved sufficient to perform all the visual analysis regarding shapes and their relations required for playing the game successfully. Another example is a system developed by Romanycia. Using a somewhat augmented set of basic operations, the system can respond to queries about the image, and extract a variety of shape properties and spatial relations, for example, finding line crossings, concave and convex regions, closed curves, isolated triangles, vertical rectangles, triangles inside squares, vertical bars that are parts of triangles, convex shapes inside closed curves, and so on.

The discussion in the last few sections of the basic operations and their use in establishing spatial relations illustrates that in perceiving spatial relations the visual system accomplishes with intriguing efficiency highly complicated tasks. There are two main sources for the complexity of these computations. 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 boundaries. In addition to the complications involved in the realization of the individual 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 spatial 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.

9.5 The Assembly and Storage of Routines

The use of visual routines allows a variety of properties and relations to be established using a fixed set of basic operations. We have discussed above a number of plausible basic operations. In this final section I will raise some of the general problems associated 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 will be more efficient to start at some other locations, or from the bounding contour. As another example, suppose that we are trying to locate an item defined by the combination of two properties, such as a vertical red item in a field of vertical green and horizontal red distractors, as in (Treisman 1977, Treisman & Gelade 1980). There are at least two alternative strategies for detecting the target: one may either scan the red items, testing for orientation, or scan the vertical items, testing for color. The distribution of distractors in the field determines the relative efficiency of these alternative strategies. In such cases it will 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.

We have also seen in discussing the tracing operation that when additional information was available, such as differently colored curves, the strategy for solving the two-x's-on-a-curve changed. This appears to be a general situation--the same goal can often

be reached by different routines, and various parameters of the scene, such as the density and distribution of objects and their properties, will determine which routine is more appropriate.

This introduces the "assembly problem" of visual routines, that is, the problem of how routines are constructed in response to specific goals, and how this generation is controlled by aspects of the scene to be analyzed. In the above examples, a goal for the computation was set up externally, and an appropriate routine was applied in response. In the course of performing visual cognition tasks, routines are usually invoked in response to internally generated goals. Some of these routines may be stored in memory rather than assembled anew each time they are needed. These stored visual routines constitute "perceptual programs" somewhat analogous to stored "motor programs" for executing movements. The repeated execution of a familiar visual task 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 fill in details and generate intermediate routines when necessary. The perceptual activity will be guided by setting pre-stored goals that the assembly process will then expand into detailed visual routines.

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 in fact from two different sources. First, assembly time will be saved if the routine is already "compiled" in memory. Second, stored routines may be improved with practice, as a result of either external instruction, or by modifying routines when they fail to accomplish their tasks efficiently. In the final chapter of the book we will discuss cortical mechanisms that could be used in the assembly, storage, and application of visual routines.

9.6 Routines and Recognition

Visual routines are useful for a variety of visual tasks that arise in the course of reasoning about objects in the scene, visual search, the manipulation of objects and planning of actions, navigation in the environment, the use of visual aids such as diagrams and maps, and the like. What about visual recognition—are visual routines also used in the course of visual object recognition?

Recognition and visual routines are both important parts of high level vision, but they are generally separate processes. As we have seen in discussing object recognition, the general capacity to establish abstract shape properties and spatial relations is usually not required in the recognition of specific 3-D objects. To recognize a familiar 3-D object, we typically use specialized recognition processes that compare the current view with stored models, as discussed in previous chapters, rather than the processes of visual routines discussed in this chapter. The specialized recognition mechanisms are useful for the specific task of object recognition, but are not suitable for general visual cognition tasks. To inspect a map, for instance, and locate the largest city between the lake and the highway, the mechanisms of object recognition are no longer sufficient (although they may be employed as a part of performing the task), and we will use other processes of visual routines. Biologically, the processes related to visual routines may be associated primarily with the so-called dorsal system of visual processing, and the recognition process with the ventral processing stream (Ungerleider & Mishkin 1982, Ungerleider & Haxby 1994).

Under some cases, however, visual routines could also serve a useful role for the purpose of object recognition. This is not surprising since, as discussed in the introduction, recognition is a general term, and more than a single process can be used for the task of identifying or classifying an object. Routines can be employed, for example, to scan a large object that cannot be perceived effectively in a single glance, or we may use boundary tracing to

trace the stream of connected characters in recognizing cursive handwriting. As was mentioned in the previous chapter, some of the processes of image segmentation,in particular selection and completion, also involve the use of visual routines. As another example, visual routines can be used to distinguish between two closely similar objects. To distinguish between two similar cars, for instance, or between highly similar faces, we may first perform a more general recognition, and then execute a "disambiguating routine" that inspects special distinguished locations, looking for the shape of a specific part, or a special marking and the like. In such cases final recognition is obtained by the combination of general recognition mechanisms, of the type discussed in previous chapters followed by the application of appropriate disambiguating routines.