THE ROLE OF SPATIAL WORKING MEMORY IN SHAPE PERCEPTION

Geoffrey E. Hinton

MRC Applied Psychology Unit
Cambridge, England

ABSTRACT

Three demonstrations are presented and used to support a number of apparently unrelated claims about the internal representations that people have when they perceive or imagine a spatial structure. The first demonstration illustrates properties of the spatial working memory that enables us to integrate successive glimpses of parts of an object into a coherent whole. The second demonstration shows that our ability to generate a mental image is severely limited by the form of our knowledge of the shape of an object. The third shows that the shape representation which we create when we attend to a whole object does not involve creating the kinds of shape representations for the parts of the object that we would form if we attended to them and saw them as wholes in their own right. The real motivation for this medley of demonstrations and for the interpretations offered is that these phenomena can all be seen as manifestations of a particular kind of parallel mechanism which is described briefly in the last section.

I PERCEPTION THROUGH A PEEPHOLE

Fig. 1 illustrates a phenomenon called anorthoscopic perception that occurs when people perceive an object one piece at a time through a slit or peephole (Hochberg, 1968). Under suitable conditions people report that they have a perceptual experience of the whole object. They somehow integrate a number of separately perceived pieces into a single Gestalt. This means that they must be storing internal records of their perceptions of the individual pieces. The simplest theory of anorthoscopic perception is that the subject builds up an internal, picture-like representation part by part, and then uses this internal "picture" as a substitute for a retinal image in identifying the whole object. As we shall see, this theory has problems.

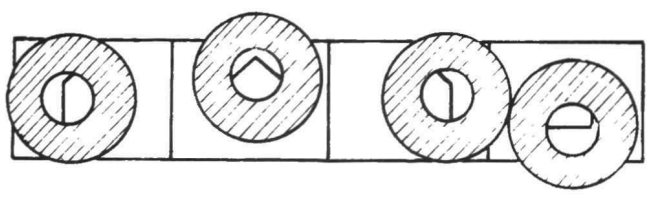

Figure 1. A cartoon strip showing a peephole moving around the outline of a shape. The fact that successive frames in the cartoon fall in different positions makes the task harder.

Retina-based versus scene-based frames

In the early stages of visual processing, the size, position and orientation of parts of the visual input are represented relative to the frame of reference defined by the retina. Anorthoscopic perception, however, cannot depend on storage in these early, "retina-based" representations because people typically fixate on the peephole, so all the different pieces of the object project to the same bit of the retina (Rock, 1981). Representations that encode the positions of the pieces relative to the retina would not allow us to perceive the whole object because the relative position of a piece within the whole is determined by where the peephole is, not by where the piece falls on the retina. It is just conceivable that as we move our eyes, the internal records of all the previously perceived pieces are correspondingly altered so that the records always encode where the piece is relative to the current retinal position, but this seems very unlikely.

What is needed is a way of representing where the pieces are that is not affected by eye-movements or even by movements of the whole person through space (Turvey, 1977). This can be achieved by using a temporary scene-based frame of reference that is defined by some larger contextual object or configuration within the external scene. If we keep a continually updated representation of the relationship between the retina and this scene-based frame, we can use it to convert from positions on the retina into positions relative to the scene before storage. These positions relative to the scene will be unaffected by subsequent eye or body movements. Obviously the scene-based frame will have to change from time to time, and it will have to have a scale that is appropriate to the scale of the parts we are attending to, but over a period of a second or two, perceptual integration of the results of successive fixations could be achieved by using a single scene-based frame of reference.

Post-categorical versus atomistic representations

In a picture-like representation, the shapes of objects are not explicitly represented -- it requires an interpretive process to extract them. Consider, for example, how a straight line is represented in an array. The line is decomposed into "atomic" fragments each of which is depicted by filling in one cell in the array. The absolute positions of the individual atomic fragments relative to the whole array are encoded directly and precisely, but there is no direct encoding of the straightness of the line, because this depends on the relative positions of the various fragments. Using this kind of atomic depiction it is impossible to represent the fact that a line is straight without representing precisely where it is relative to the whole array. It is impossible to be precise about shape and vague about position in a picture-like representation.

The memory used in anorthoscopic perception,

however, seems to allow just this combination of precision and vagueness. If a peephole is moved around a polygonal spiral (see Fig. 2) people often "perceive" a closed polygon. Their memory for the precise locations of the individual sides is poor and can be swayed by expectations about closed polygons, but they know that the sides are straight. This informal evidence that spatial working memory can be more precise about the shapes of pieces than about their positions implies that it contains explicit representations of shapes rather than being a picture-like collection of atomistic local features in which shapes are only implicit. A recent experiment supports this conclusion.

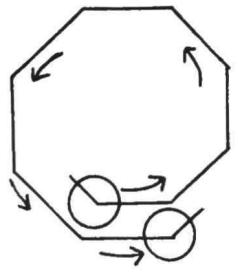

Figure 2. A peephole is moved around a polygonal spiral without revealing the free ends or the adjacent parallel sides.

Girgus, Gellman, and Hochberg (1981) have shown that it is considerably easier to "see" the shape of a whole object if the peephole is moved around the outline of the object than if the peephole jumps randomly from one part of the outline to another. The two different conditions were balanced so that the total exposure to any one part of the object was identical, so the contents of a picture-like store would be equally good in both cases. The obvious interpretation of this experiment is that when neighbouring parts of an object are exposed in succession, it is possible to form more complex chunks (shapes) and hence to reduce the number of chunks that must be stored in spatial working memory. When successive exposures are of widely separated pieces, either no chunks are formed, or chunks are created which do not correspond to the natural parsing of the whole object into parts. This type of explanation implies that the memory involved contains explicitly segmented and identified chunks.

II THE CUBE TASK

Hinton (1979) describes an apparently simple mental imagery task that people cannot do:

"Imagine a wire-frame cube resting on a tabletop with the front face directly in front of you and perpendicular to your line of sight. Imagine the long diagonal that goes from the bottom, front, left-hand corner to the top, back right-hand one. Now imagine the cube is reoriented so that this diagonal is vertical and the cube is resting on one corner. Place one fingertip about a foot above a tabletop and let this mark the position of the top corner on the diagonal. The corner on which the cube is resting is on the tabletop, vertically below your fingertip. With your other hand point to the spatial locations of the other corners of the cube."

It is fairly easy to imagine a cube in just about any orientation if the orientation is defined in terms of the natural axes of the cube. But when the diagonal is used to define the required orientation, we realise that relative to the diagonal, we have no clear idea where the various parts of the cube are. Our knowledge of the spatial dispositions of the parts of a cube is relative to the "intrinsic" frame of reference defined by the cube's own axes. Knowledge in this form is ideal for recognising the shape of a rigid object because whatever the object's actual size, position and orientation, the dispositions of its parts will always be the same relative to an intrinsic frame of reference based on the object itself (Palmer, 1975; Marr and Nishihara, 1978). So if the appropriate object-based frame can be imposed, the early retina-based representations which encode the positions of the parts relative to the retina can be recoded into object-based representations and this encoding will constitute a viewpoint-independent shape description that allows the object to be recognised.

I have now appealed to three different sorts of reference frame. The initial processing of the visual input uses representations relative to the retina; recognition of the shape of an object involves recoding these early retina-based representations into ones that are relative to an object-based frame; and anorthoscopic perception relies on storing the relationships of recognised shapes to a temporary scene-based frame.

III FRUITFACE

Fig. 3 shows a face composed entirely of pieces of fruit. Palmer (1975) reports that when subjects are shown this figure very briefly, they see it as a face without seeing the parts as fruit. The fruitface figure demonstrates that forming the Gestalt for a face does not depend on forming Gestalts for the parts. This is puzzling because to see the face we must form some representations of the parts and their relationships to the whole, since it is the relative dispositions of the parts within the whole that make it a face. One possibility which has not been much explored is that each part of the face can have two quite different internal representations. When the part is seen as a constituent of the face it receives a representation in which it is interpreted as filling the role of, say, an eye because of its crude overall shape and its relation to the whole face. When it is seen as a whole in its own right, however, it receives a quite different internal representation in which the rough shapes and dispositions of its parts cause it to be seen as a piece of fruit.

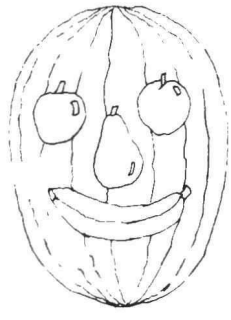

Figure 3. A face composed entirely of pieces of fruit. (After Palmer, 1975)

The idea that an object receives a quite different internal representation when it becomes the object of focal attention does not fit the popular view of attention as a kind of internal spotlight which can illuminate any one of a number of otherwise unconscious shape representations. However, the idea is very compatible with "early selection" theories (Triesman and Gelade, 1980) in which focal attention is constructive and is necessary for the generation of a shape representation.

The internal spotlight metaphor for visual attention is a powerful one, but I believe it is based on a mistaken analogy between external perception and introspection. Normally our attention moves rapidly and smoothly from one level to another and we do not realise that at any instant we are attending at just one level. Only when the information at the different levels is made inconsistent, as in the fruitface, does it become obvious that the Gestalt for the whole cannot coexist with the Gestalts for its parts. Introspection is of little use for deciding what is in our minds at one brief instant because it does not allow us to decide between two possibilities. Either there are shape representations that lurk outside focal attention, or shape representations are generated or regenerated the moment we ask ourselves whether they are there. Our fundamental epistemological assumption that the existence of objects is independent of our awareness of them cannot be applied to the contents of our own minds.

An obvious objection to any theory which claims that people only see one shape at a time is that the shape of an object is determined by the shapes of its parts and their dispositions relative to the whole. This kind of recursive definition of a shape in terms of the shapes of its parts leads to a regress that only terminates at hypothetical "primitive" features. The fruitface figure is important because it suggests an alternative way out of the regress. The representations of the parts that are used in perceiving the shape of the whole may be different in kind from the representations used to perceive the shapes of the parts when we attend to them. Naturally, different shape representations must be able to influence one another. Having recognised an eye it should be easier to see the whole face, but this influence could be mediated by spatial working memory. Although only one Gestalt can be formed at a time, records of many previous Gestalts can be kept in working memory and used to influence the formation of the next Gestalt.

IV WHAT THE DEMONSTRATIONS SHOW

The demonstrations have been used as evidence for the following claims:

1. We integrate the information obtained in successive glances by storing records of the shapes that we identify and their relationships to a temporary scene-based frame of reference. We can use these stored records to generate new shape representations.

2. The process of recognising a shape (forming a Gestalt) involves imposing an object-based frame of reference and representing the size, position, and orientation of each part of the object relative to this frame.

3. The representation that an object receives when it is seen as a Gestalt and its shape is recognised is completely different from its representation when it is seen as a constituent of a larger Gestalt. Only one Gestalt can be formed at a time, but many separate records of previous Gestalts can be stored in spatial working memory.

V A MECHANISM FOR SPATIAL REPRESENTATIONS

There is not space here to discuss all the various kinds of mechanism that have been suggested for representing spatial structures. I shall simply describe one possibility which is designed to make use of parallel interactions between very large sets of features. This kind of computation seems to be a natural way of harnessing the computational power provided by a system like the brain in which a large number of richly interconnected units all compute in parallel (Anderson and Hinton, 1981). The mechanism is based on four related assumptions:

1. A perceptual feature must always be represented relative to some frame of reference because properties like the length, position, and orientation of a feature implicitly assume a reference frame.

2. At any moment during perception we use three different frames of reference -- retina-based, object-based, and scene-based -- so our perceptual apparatus has three different sets of units, each of which represents features relative to one of these frames of reference.

3. The meaning of features relative to one frame of reference in terms of features relative to another depends on the relationship between the two frames. So the way in which units in one set affect units in another set must be controlled by a representation of the spatial relationship between the frames of reference used by the two sets. A particular spatial relationship pairs each unit in one set with one unit in the other set, and allows activity in one of these units to cause activity in the other.

4. Different Gestalts correspond to alternative patterns of activity in the very same set of object-based units. So only one Gestalt can be formed at a time, though records of many previous Gestalts can be stored as activity in the scene-based units.

Fig. 4 incorporates these assumptions. Unlike many box diagrams in psychology, the separate boxes really are intended as separate collections of hardware units. Every unit continually recomputes its activity level as a function of the input it receives from other units. In the short term (i.e in about 100 msec), the whole system computes by settling into a state of activity that is temporarily stable. This kind of settling process is described in more detail in Hinton (1981b) where it is shown that the process of assigning an appropriate object-based frame of reference can be implemented by the three-way interaction between retina-based units, object-based units and the units for representing the spatial relationship between

the retina and the object. This kind of three way interaction is what the triangular symbols in Fig. 4 depict. After each settling, control processes (unspecified here) can reset the pattern of activity in any set of units, and thereby initiate a new process of settling. Not all the units in a set need be involved in the interactions with other sets. For example, the object-based units that are directly affected by retina-based units probably code fairly simple features, whereas the object-based units that directly affect the scene-based ones probably code complex conjunctions of the simpler features.

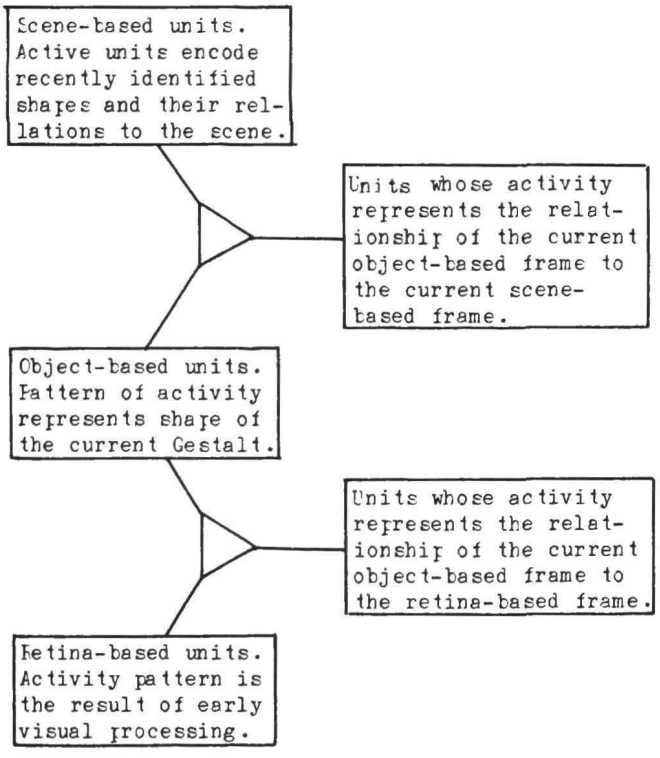

Figure 4. A parallel mechanism.

This kind of mechanism raises many interesting issues, some of which are discussed elsewhere (Hinton, 1981a). The following section focusses on what the scene-based features are like, and how they influence the the formation of a new Gestalt, i. e. how they affect the formation of temporarily stable pattern of activity in the object-based units.

Scene-based features

Once the general approach of implementing spatial working memory as activity in a set of scene-based units is accepted, quite a lot can be deduced about the nature of the units from their function. One important function of spatial working memory is to allow previously identified Gestalts to aid in the formation of related Gestalts. Having recognised an eye, the whole face should be easier to see, and vice versa. The kind of precisely located, atomistic features that would be needed for a picture-like representation would not be of much value in spatial working memory, because they would not explicitly represent the identities of objects, and so their effects could not be made to depend on these identities. It is more useful to make each active scene-based unit represent the existence of an object of a particular type with a particular relationship to the current scene, as the following examples show.

Suppose that as a result of previous perceptual analysis, activity in a scene-based unit, S_i, represents the existence of an eye with the relationship F_{is} to the scene. Suppose also that the system is now attempting to settle on an interpretation of a larger object (a face) with the relationship F_{fs} to the scene. F_{is} and F_{fs} determine F_{if}, the relationship of the eye to the face, and so they determine which object-based unit, C_i, should be activated to represent the eye as a constituent relative to the frame of reference of the whole face. This influence of the contents of working memory on perception can be implemented (see Fig. 4) by having an explicit representation of F_{fs} which governs the interaction between scene-based and object-based units and ensures that activity in S_i provides excitatory input to C_i.

Now consider what is required of spatial working memory if the face is seen first and attention is then focussed on one eye. The fact that this part had the role of an eye within the whole face should facilitate its interpretation as an eye when it becomes the focus of attention. This effect can be achieved if the Gestalt for the whole face activates scene-based units that represent the major constituents of the Gestalt as well as the whole. So the mapping from object-based to scene-based units operates simultaneously on units that represent the identity of the whole Gestalt and on units representing its major constituents.

VI CONCLUSION

Three demonstrations have been used to illustrate aspects of our internal representations of spatial structures. Particular attention has been given to the spatial working memory that allows people to integrate their perception over time. It has been argued that this memory contains compact records of the rich perceptual Gestalts that are formed when a person attends to an object. The interactions between spatial working memory and the apparatus in which Gestalts are formed allows previous Gestalts to influence (or entirely determine) the formation of the current Gestalt even though only one Gestalt can be present at a time. This view of the role of spatial working memory supports "early selection" theories in which focal attention is required to synthesize a shape, and only one shape can seen at a time. It also supports the view that different Gestalts correspond to alternative patterns of activity in a set of units that encode features relative to a frame of reference imposed on the object.

Finally, a few provisos. The demonstrations are well known but the interpretations of what they show are probably contentious, and the mechanism I suggest is speculative and underspecified. There has not been space to elaborate on many interesting issues like how the mechanism might account for the experimental data on mental rotation (Cooper and Shepard, 1973) or spatial working memory (Broadbent and Broadbent, 1981; Phillips and Christie, 1977). Nor has it been possible to discuss crucial

theoretical issues like the number of units that would be required by the mechanism, or the problems of encoding novel shapes in working memory.

ACKNOWLEDGEMENTS

I thank Steve Draper, Ed Hutchins, Tony Marcel, Don Norman, Dave Rumelhart, Tim Shallice, Joanne Sharp and Aaron Sloman for useful discussions. Many of the ideas presented here were developed while I was a Visiting Scholar at the Program in Cognitive Science at the University of California, San Diego, supported by a grant from the Sloan Foundation.

REFERENCES

Anderson J. A. & Hinton, G. E. Models of information processing in the brain. In G. E. Hinton & J.A. Anderson (Eds.) Parallel models of associative memory. Hillsdale, NJ: Erlbaum, 1981.

Broadbent, D. E. & Broadbent, M. H. P. Recency effects in visual memory. Quarterly Journal of Experimental Psychology, 1981, 33A, 1-15.

Cooper, L. A. & Shepard, R. N. Chronometric studies in the rotation of mental images. In W. G. Chase (Ed.), Visual information processing. New York: Academic Press, 1973.

Girgus, J. S. Gellman, I. H. & Hochberg, J. The effect of spatial order on piecemeal shape recognition: A developmental study. Perception and Psychophysics, 1980, 28, 133-138.

Hinton, G. E. Some demonstrations of the effects of structural descriptions in mental imagery. Cognitive Science, 1979, 3, 231-250.

Hinton, G. E. Shape representation in parallel systems. To appear in Proc. IJCAI-81 1981a.

Hinton, G. E. A parallel computation that assigns canonical object-based frames of reference. To appear in Proc. IJCAI-81. Vancouver, Canada, 1981b.

Hochberg, J. In the mind's eye. In R. N. Haber (Ed.) Contemporary theory and research in visual perception. New York: Holt, Rinehart and Winston, 1968

Marr, D. & Nishihara, H. K. Representation and recognition of the spatial organisation of three-dimensional shapes. Proc. Roy. Soc. Series B, 1978, 200, 269-294.

Palmer, S. E. Visual perception and world knowledge: Notes on a model of sensory cognitive interaction. In I. A. Norman & D. E. Rumelhart (Eds.), Explorations in cognition. San Francisco: Freeman, 1975.

Phillips, W. A. & Christie, D. F. M. Components of visual memory. Quarterly Journal of Experimental Psychology, 1977, 29, 117-134.

Rock, I. Anorthoscopic perception. Scientific American, March 1981, 244, 103-111.

Triesman, A. M. & Gelade, G. A feature-integration theory of attention. Cognitive Psychology, 1980, 12, 97-136.

Turvey, M. T. Contrasting orientations to the theory of visual information processing. Psychological Review, 1977, 84, 67-88.