# Rational Processes in Perception

Alan Gilchrist and Irvin Rock

In this paper we will give our reasons for believing that certain current attempts to explain perceptual phenomena on a lower level in terms of known sensory mechanisms are untenable. We will do this by focussing on two topics, lightness perception and the perception of apparent motion. We will summarize some older data (not all of which are sufficiently known) and will describe some recent work of our own. Finally, on a more positive note, we will try to indicate the direction that a theory must take if it is to deal effectively with these phenomena.

## Lightness Perception

We will begin with the assumption that Helmholtz was essentially wrong in his belief that an object's lightness can be inferred by interpreting the luminance reflected by it to the eye in terms of the amount of illumination falling on it. Such a process requires unequivocal information about the illumination whereas the only information directly available is the intensity of light, or luminance, reflected by each surface in the field. Each such luminance is the joint product of the reflectance property of the surface and the illumination falling on that surface. Rather we will assume that the perceived shade of gray of a surface is governed primarily by the luminance of that surface relative to the luminance of neighboring surfaces as Hering (1920) suggested and as Wallach(1948) elegantly demonstrated. There is now fairly wide agreement among investigators on this general principle.

But what is the underlying explanation of it? There is great appeal in Hering's suggestion of reciprocal interaction, i.e. that a bright region of the field would have a darkening effect on an adjacent region and a dark region would have a brightening effect on an adjacent region. We now know for a fact that the rate of discharge in one nerve fiber is attenuated when a neighboring fiber is stimulated by light. Thus such lateral inhibition can plausibly be invoked to explain why the apparent lightness of one region is governed by the extent of stimulation of an adjacent region (see Cornsweet, 1970; Jameson and Hurvich, 1964).

In fact, given lateral inhibition as a known sensory effect, one might have been able to predict the phenomenon of contrast, even if it had never been observed (although questions can be raised about the spatial distance over which such a mechanism can be expected to occur). Surrounding a gray region by a white one should lead to diminished discharging of retinal fibers stimulated by the gray region; surrounding another gray region of the same value by a black one should lead to increased discharging of retinal fibers stimulated by that gray region because of a release of inhibition. Thus one of these gray regions should look lighter than the other and so it does. An implicit assumption here is that the phenomenal shade of gray perceived in a given region is a direct function of the rate of discharging of fibers stimulated by that region.

The fact of constancy of lightness can be explained along the same lines. When the illumination falling on a surface changes, then the luminance of all adjacent regions rises and falls together. Thus, while the rate of discharging of cells stimulated by a gray region should increase when illumination increases, so too should the rate of discharging of cells from a surrounding white region increase. The latter will increase the inhibition on the former with the net result of little if any change in the absolute rate of discharging of those cells. Therefore the perceived lightness should remain more or less constant and so it does. Note again, however, the assumption, here explicit, that the phenomenal lightness is a direct function of the rate of discharge of the appropriate fibers.

Underlying this assumption is another assumption about how the visual system works that Gilchrist (1981) has called the photometer metaphor. Just as the signal produced by a photometer is a direct function of the light falling upon it, so the perceived lightness of each point in the field is assumed to be a direct function of the rate of discharging of the cells stimulated by each such point. With the knowledge that has been available about light, about the formation of the retinal image, and about photochemical processes and nerve physiology it is understandable why such a view has become so deeply ingrained as not even to be explicitly recognized as an assumption. Given this assumption, phenomena such as contrast and constancy, in which lightness does not correlate with luminance, seem to require an explanation long the lines of lateral inhibition.

There is now, however, reason to reject this approach. Evidence has been accumulating to support the theory that the perception of lightness (and chromatic color) is based on information at the edges between regions of differing luminance (or hue). Homogeneous regions between edges are then "assumed" to have the lightness or color indicated by these edges. There is overwhelming evidence (Yarbus, 1967; Whittle and Challands, 1969; J. Walraven, 1976) that the visual system responds to changes in stimulation, not to an unchanging state of stimulation. This is normally guaranteed by continuous eye movements, for vision. Whenever an image can be held stationary on the retina for a few seconds, all visual experience stops. These facts are inconsistent with the photometer metaphor and they strongly indicate the crucial nature of edges or gradients in the retinal image since this is where stimulation changes in the normal moving eye. Krauskopf (1963) has shown that when the boundary of a surface is prevented from moving on the retina, its color will disappear and be replaced by the color of the surrounding region, signalled by the boundary of that region.

It seems unlikely that any absolute luminance information would be picked up in this way and yet it now seems quite possible that the visual system achieves what it does using only relative information. Even a simple edge-relations approach goes a long way toward explaining lightness constancy since the luminance ratio between two adjacent surface colors remains the same even when illumination changes.

The important point here is that there is no need to invoke a concept such as lateral inhibition to explain constancy. Once the photometer assumption is made explicit and in fact, is displaced by the concept of edge information, the whole edifice collapses. Of course lateral inhibition is a well-established physiological fact. It is probably part of the process whereby the ratio at an edge is determined. But we don't believe that lateral inhibition solves any of the basic problems of constancy. The concept of an exaggeration or enhancement of edge ratios seems unnecessary and illogical. If lateral inhibition exaggerated an edge ratio, it would do so in the same way every time an edge of the same value were present on the retina. A given edge ratio would be specified by a given neural signal, with or without an exaggeration function. Therefore the exaggeration function doesn't seem to add anything of explanatory value.

Certain problems turn out to be dissolvable pseudo-problems with the adoption of an edge-relations approach. One of these is constancy under changing illumination. On the other hand, other problems emerge, although they are more tractable. For example, how do we now explain the constancy of surface lightness as the surface is viewed against differing backgrounds? The luminance ratio at the edge of a surface can change

dramatically as it is placed on different backgrounds and yet lightness perception remains almost unchanged.

For example, in the classic example of lightness contrast, the gray square on the white background has an edge ratio that is radically different (even opposite in sign) from that of the gray square on the black background. Thus, under a simple edge theory they ought to appear radically different in lightness, and yet they appear almost the same. This suggests that lightness is not determined simply by the boundary of a surface, but by the relationship between that boundary and other boundaries. Presumably the boundaries of the squares themselves only signal departures from a background lightness, which in turn is signalled by the boundary of each background. Thus the edge dividing the white and black backgrounds signals the relationship between the two background lightnesses and we might expect that this edge will be as critical to the lightness of the targets as the edges of the targets themselves. In fact Gilchrist and Piantineda (unpublished experiment) have found that if that edge is retinally stabilized, the two gray squares turn black and white respectively, just what we would expect based simply on the ratios at the edges of the gray squares. We might say that the assignment of lightnesses to the various regions is the end result of a computational process in which information from all edges present is integrated. Arend (1973) and Land and McCann (1971) have proposed similar schemes.

If such computational processes occur and are governed by remote as well as local edge information, the reader may well wonder about the achievement of constancy. Consider the typical case where two gray disks of equal reflectance on the same background are unequally illuminated because one region and its immediate background are in shadow. Earlier we said that constancy could be explained on the basis of the equal ratio of each gray region to its background. That would be true in the example under consideration. But now we have also said that the presence of other edges enters into the equation. The shadow edge can easily have a ratio as great as a white-black edge and, more probably, even greater. If this is entered into the computation, constancy would fail; the disks would be seen as different shades of gray in accordance with the luminance difference between them. The equal disk-to-surround luminance ratios here logically cannot signify that the two grays are equal if the gray regions are seen as on backgrounds of <u>different</u> luminance values, or so it would seem.

Unless the perceptual system can discriminate between reflectance edges and illumination edges, that is between changes in the pigment of the surface and changes in the amount of illumination shining on the surface. If so, perhaps illumination edges would not be included in the computation of surface lightness values. There is now strong evidence of just such discrimination of reflectance and illumination edges (Gilchrist, in press). If observers view the two disks under the conditions just described, they typically do perceive the two grays as almost equal, i.e. constancy is achieved. Moreover, they perceive both sides of the background as white with one side in shadow. Thus the central edge is apparently correctly identified as an illumination edge. If, however, the observers view the display through an aperture that permits only part of the background and the two gray regions to be seen, and if the edge of the shadow is reasonably sharp, the grays no longer look equal, constancy is destroyed. Moreover, the observers now perceive the two sides of the background as unequal in lightness. Thus the edge is interpreted as separating different reflectances, not different illuminations. In this condition only then does the central edge enter into the process of computing the lightness of the gray disks.

We reported earlier that if the boundary between the black and white backgrounds in the traditional contrast pattern is made to disappear through retinal stabilization, one gray square turns black and the other turns white. This provides some of the best evidence for the concept of edge integration. A similar experiment was done by Gilchrist, et. al. (in press) that demonstrates the importance of the distinction between reflectance edges and illumination edges. When the boundary between the white and black backgrounds is made to look like the edge of a shadow, the two squares will also turn black and white respectively, just as in the stabilized-image experiment. Thus when an edge is identified as an illumination edge, it seems to drop out of the integration process for surface lightness just as if it were invisible.

It would take us too far afield to enter into a <u>full</u> discussion of precisely how the perceptual system discriminates illumination from reflectance edges. While the presence of penumbra at an illumination edge may be one source of information it is not the only one and is not necessary in the experiments just described.

Before discussing the important role that depth perception plays in discriminating edges, it is worth considering the ramifications of what we have just discussed for the notion of lateral inhibition or any other theory of neural interaction which seeks to explain the important effects of remote edges on what is perceived in regions adjacent to other edges. For now in addition to other difficulties such a theory faces, in dealing with such "remote" effects it would have to be argued that such effect <u>do not occur at all</u> when those remote edges are interpreted as representing illumination rather than reflectance differences. In fact, it is interesting to note that apparently no one has noticed that when contrast effects are applied to illumination edges, they not only fail to result in constancy, but they actually <u>make matters worse</u>. Constancy requires that we explain how the perception of surface lightness could be the same on both sides of an illumination edge, given the difference in luminance. Applying a mechanism here that further exaggerates the luminance difference is a little like bringing water to a drowning man.

Recent experiments (Gilchrist, 1977, 1981) have demonstrated the role of depth perception in distinguishing an illumination edge from a reflectance edge. In one experiment an artificial interposition cue was used to make a target square appear as located either in a near plane, dimly illuminated, or in a far plane, brightly illuminated, as shown in Figure 1. In terms of the retinal image, the target square was always flanked by a surface of much lower luminance to its lower right and by a surface of much higher luminance to its upper left (note relative luminances given in Figure 1). In space, however, the low luminance surface was located in the near plane while the high luminance surface was located in the far plane. The results show that lightness is determined by the luminance ratio of the target square to its coplanar neighbor. Thus the target square looked white when it appeared in the near plane but black when it appeared in the far plane. In other terms, the edge dividing the target from its coplanar neighbor was treated as a reflectance difference while the edge dividing the target from its non-coplanar neighbor was presumably treated as representing an illumination difference. Since the retinal image was essentially the same in both conditions, this result is inconsistent with an explanation based on lateral inhibition.

In another experiment involving planes meeting to form a dihedral angle, these ideas were put to a more rigorous test in which predictions based on perceived coplanarity would be the opposite of predictions based solely on retinal ratios.

The experimental arrangements are shown in Figure 2. In the horizontal plane, a black target tab extended out into space from a larger white square. In the vertical plane, a white target tab extended upward into space from a larger black square. Thus each target

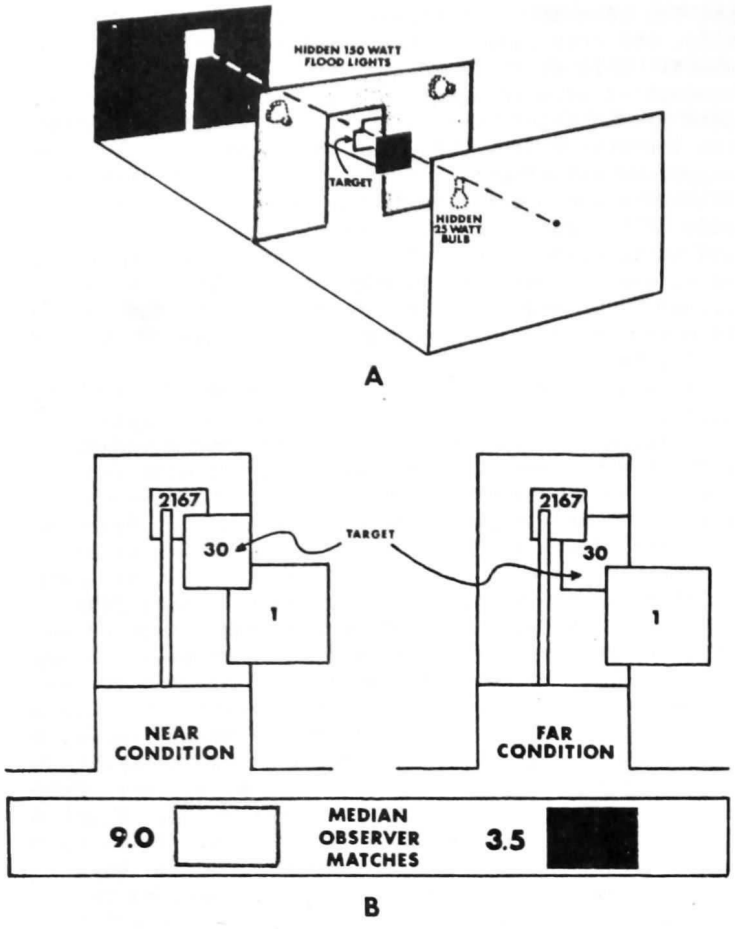


Figure 1

tab was seen against the background square that was in a separate plane. The horizontal surfaces received about 30 times as much illumination as the vertical surfaces, or just enough illumination difference to make the luminance of the black target tab equal to that of the white target tab. Given the viewing perspective of the observer, 45 degrees from each plane, the display was similar to traditional contrast displays; two targets of equal luminance on bright and dark backgrounds respectively. Thus a theory based on lateral inhibition would clearly predict that the target on the bright background, in this case the upper target, should appear darker than the other target, although the exact magnitude of the effect is harder to determine. On the other hand, if lightness is really based on luminance relationships within planes, then each tab should be compared with the larger background square that lies in the same plane, even though it is adjacent only along one edge of the tab. Thus not only would the coplanar ratio principle predict that the upper tab would appear lighter, not darker, than the lower tab, it would predict that the upper tab should look white and the lower tab black.

In fact the latter result was actually obtained. Figure 2 shows the median Munsell matches (next to samples of those Munsell values) obtained from naive observers. Moreover, since the target tabs were actually trapezoidal in shape, they could be made to switch perceived planes when viewed monocularly. In that condition of the experiment the perceived lightnesses of the tabs also switched, with the lower tab now appearing white and the upper tab appearing black. Since these changes in perceived lightness were produced solely by a change in depth perception, with no change in the retinal image, these data raise difficulties that may be insurmountable for current theories based on lateral inhibition.

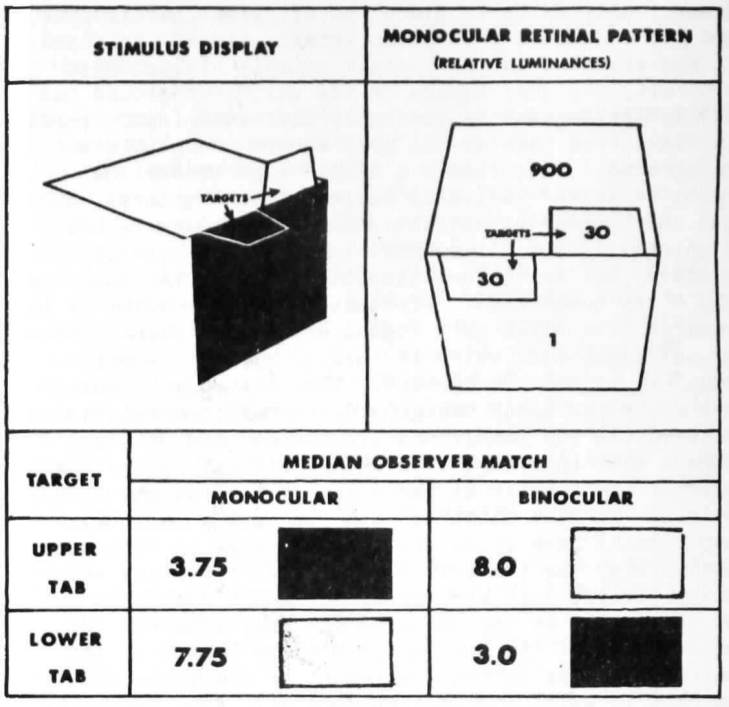


Figure 2

Apparent Motion

Although we know a good deal about the conditions that produce the illusory impression of motion referred to as apparent motion, we still do not understand why it occurs or, for that matter, why it only occurs under certain conditions. What we know about this effect is that given the sudden appearance of object a, its sudden disappearance, followed typically by just the right time interval of object b in just the right new spatial location, one tends to see motion of a to b. The currently favored explanation is that a motion-detector cell in the brain will discharge even if the appropriate receptor field of the retina is stimulated discontinuously by two points rather than by a point moving over the retina. Such cells do seem to exist in various species of animals (Grüsser-Cornehls, 1968; Barlow and Levick, 1965).

However, the fact is that it is not necessarily the case that the conditions for apparent motion perception entail stimulation of separate retinal regions. Ordinarily that is the case, since a and b are in separate spatial locations and the eye is more or less stationary. What seems to matter is the perception of a and b in separate locations in space.

To get at this question an experiment was performed in which the observers had to quickly move their eyes back and forth synchronous with the onset of a and b so that each stimulated the same central region of the retina, rather than as, more typically, two discretely different loci (Rock and Ebenholtz, 1962). Therefore, the conditions for apparent motion might be thought not to exist. Yet, the observer does locate a and b in phenomenally discrete places in the environment. The result was that although nothing was said to the observers about motion that might create an expectation of perceiving motion, most of them nonetheless spontaneously did. This experiment seems to prove that, in humans at least, it is not necessary to explain stroboscopic motion in terms of a sensory mechanism that detects sudden change of retinal location. There is neither change of retinal nor cortical locus of projection of a and b here.

An entirely different view that has been presented by Rock (1975) is that the impression of motion is a solution to the problem posed by the rather unusual stimulus sequence. First a inexplicably disappears. Then b inexplicably appears elsewhere. By "inexplicable" we

mean that when an object in the world disappears as we are looking at it, it is generally because another object moves in front of it or it is occluded by another object because of our motion. However, when a stationary object suddenly and rapidly moves to another location, it does tend to disappear from one location and to appear in another. Therefore, perhaps this state of affairs in a stroboscopic display suggests the solution of motion.

Given that potential solution, the question arises as to whether it is acceptable. Motion from a to b does account for the brief stimulation by a and b, but isn't the absence of any visible object between the locus a and b a violation of the requirement that a solution be supported by what is present in the stimulus? If the solution is "a moving across space to b" doesn't this call for stimulus support in the form of continuously visible motion across that spatial interval? Ordinarily that would be true, but it is a fact that has been demonstrated that for very rapid motion of an actually displacing object, little more than a blur can be seen in the region between the terminal locations (Kaufman et al, 1971). In fact it was shown that if the terminal locations are occluded, no motion of a moving object is seen. Therefore when the spatial and temporal intervals between a and b in a stroboscopic display are such as would correspond with the real motion of a rapidly displacing object, the absence of continuously visible movement need not act as a constraint against perceiving movement. In fact, this analysis may explain why slow rates of alternation do not lead to the impressions of motion. By "slow rate" we mean a condition with a relatively long interval between the disappearance (offset) of a and the appearance (onset) of b. Such a rate would imply a slowly moving object and a slowly moving object would normally be seen throughout the spatial interval between a and b. Therefore the absence of object motion over that interval at slow rates of alternation is a violation of the requirement of stimulus support. Hence the movement solution is not acceptable at slow rates even if the offset and onset tend to suggest this solution.

While on this topic of rate we might briefly comment on the case where the alternation is very rapid, i.e. a zero or only a minimum interval between the offset of a and onset of b. If the "on" time of a and b is itself very brief, this state of affairs will result in a and b being visible simultaneously by virtue of neural persistence. But if a is visible when b appears, the solution that a has moved to b is not supported or one might say, is contradicted. This deduction was tested by using rates of alternation that ordinarily do produce the stroboscopic effect but with the following variation: First a appears, followed by the usual blank interval; when b appears so does a (in its original location). Therefore the sequence of events is: a; a and b,: a; a and b, etc. If the presence of a during the exposure of b violates the requirements of the motion solution, then observers should not achieve a stroboscopic effect under these conditions. Our observers did not. If, however, the display is changed so that the a object that appears concurrently with b but in the same location as a is somewhat different than the a that appears alone, observers do perceive a moving to b. The sequence is a; a' and b,: a; a' and b; etc.

It was noted above that in the typical experiment on stroboscopic motion, a and b inexplicably disappear and appear. What was meant was that no rationale is provided to the observer of why they appear and disappear such as is the case when things in the environment suddenly appear or disappear because another object in front suddenly moves out of the way or in the way. This suggested the following kind of experiment. Suppose we cause the retina to be stimulated by a and b in just the right places at just the right tempo, etc., but by a method in which we move an opaque object back and forth, alternately covering and uncovering a and b.

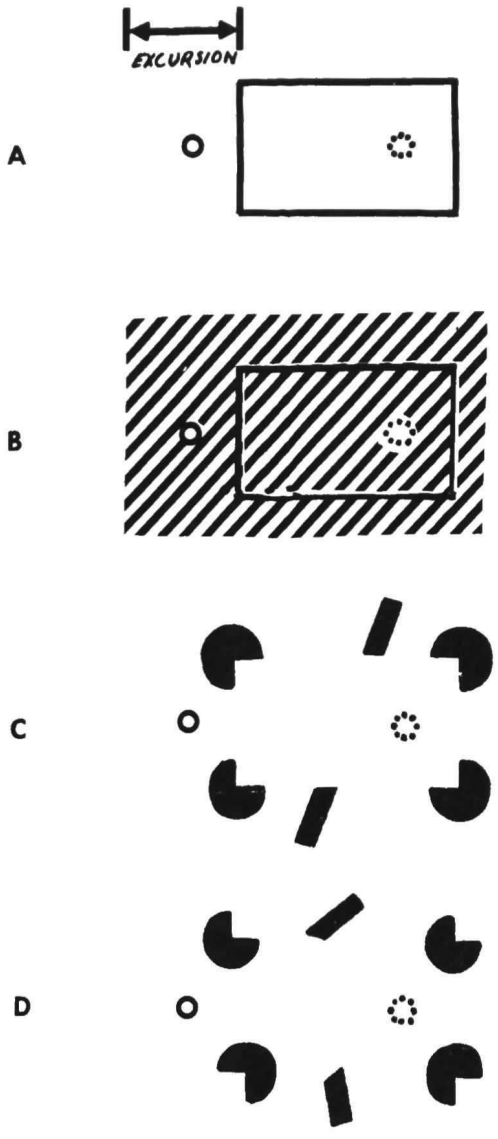


Figure 3

(See Figure 3B) As far as the sensory theory of apparent motion is concerned there is no obvious reason why these conditions should not produce an impression of a and b moving. But from the standpoint of problem solving theory, we have now provided an explicable basis for the alternate appearance and disappearance of a and b, namely, that they are there all the time but undergoing covering and uncovering. Therefore the perceptual system may prefer this solution or at least we are offering it a viable alternative not usually available (see Stoper, 1964; Sigman and Rock, 1974).

The subjects rarely perceived motion of the dots here. Some may object that the presence of the actually moving rectangle interfered in some way with perceiving stroboscopic motion. It is, after all, an unusual, atypical, way of studying such motion. The rectangle may draw the subjects' attention or otherwise inhibit motion perception of the dots. For this reason a slight change was introduced, one that had another purpose to it as well. Suppose the rectangle moves, but a bit too far, far enough no longer to be in front of where the dot had been. But by a method, the details of which need not be discussed here, things were so arranged that when the rectangle is in its terminal location, the dot is nonetheless not visible.

Now it is no longer a fitting or intelligent solution to perceive a and b as two permanently present dots that are simply undergoing covering and uncovering. For it can be seen that in fact the rectangle is not covering the spot in its terminal location and yet the spot is not visible (violation of the stimulus-support requirement). Therefore the best solution is again one of

movement and that is what the subjects perceived. Note that this experiment serves as a control for the objection raised to the first one; the moving rectangle here does not interfere with perceiving motion of the dots.

Another variation performed is based on the idea that for the covering-uncovering solution to be viable, the covering object must appear to be opaque. If it does not, it can hardly be covering anything. This factor was manipulated in an experiment illustrated in Figure 3B. The actual stimulus conditions are very similar but in one case, because the oblique lines within the rectangle are stationary and aligned with all the others, the rectangle looks like a hollow wire perimeter. In a control condition the lines inside it moved with the rectangle, and it looked like an opaque object. The difference in results is very clear: when the rectangle appeared to not be opaque, subjects by and large perceived movement whereas in the case of the opaque-appearing rectangle, they did not. A hollow rectangle is in contradiction of the property of opacity required by the covering-uncovering solution.

In a final experiment, conditions were such that no physical contours at all moved back and forth in front of the dots. There was, however, a phenomenally opaque object that moved, one based on illusory contours, as illustrated in Figure 3C. The great majority of subjects did not perceive movement. In a control experiment, illustrated in Figure 3D, the orientation of the corner fragments was changed so that no subjective rectangle was perceived and this array was moved back and forth. Now the majority of subjects did perceive movement.

It should be noted that in all these cases where a covering-uncovering effect is perceived there is no reason why movement of the dots could not have been perceived as well. That is to say, if the observer were to see an opaque rectangle moving back and forth and, simultaneous with this, a dot stroboscopically moving in the opposite direction, such a solution would also account for the stimulus sequence. Conversely everything implied by that solution is represented in the stimulus, and no contradictory perception is occurring. Therefore the tendency to perceive dots undergoing occlusion and disocclusion rather than dots moving, represents a preference for one solution over the other. The preferred solution is obviously related to a very basic characteristic of perception, namely, object permanence, the tendency to assume the continued presence or existence of an object even when it is momentarily not visible for one reason or another. But given the very strong predilection we have to perceive apparent motion even under the most unlikely conditions, it remains a problem as to why it is not perceived in this situation and the object-permanence solution is preferred. A possible answer is that the covering-uncovering solution accounts for all stimulus change by one "cause": a moving rectangle covering and uncovering spots that are continuously present. The other solution entails two independent events that are coincidentally and unaccountably correlated; a rectangle moving in one direction and in anti-phase to spots moving in the opposite direction.

There is another line of evidence that also strongly supports a problem-solving interpretation of stroboscopic motion. If the stimulus consists of more than a single dot or line, the problem arises of what in a is seen moving to what in b. To make the point clear, suppose that a and b each consist of a two-by-three matrix of dots. What will be seen here is the rectangular grouping moving as a whole (Ternus, 1926). Apparently the perceptual system seeks a movement solution that will do justice to the object as a whole. Indeed, were this not the case, the motion perceived in moving pictures would be quite chaotic, because it is typically objects consisting of many parts that change location from frame to frame (and often many such objects are simultaneously changing locations in either the same or varying directions). Yet this outcome is not predictable at all in terms of the other kinds of sensory theories mentioned earlier.

A related example is the perception of motion of complex stimuli such as the line drawings of three-dimensional cuboid figures that Shepard and his associates have used in the mental rotation studies. Shepard and Judd (1976) presented two perspectives of such figures in a stroboscopic motion paradigm and showed that, at the appropriate rate of alternation, observers perceive these objects rotating through the angle necessary to account for the change in perspective from a to b. This effect clearly implies that the perceptual system deals with the problem of accounting for the differences in a and b by an intelligent motion solution. A further finding of interest is that the optimum rate of alternation for achieving a continuous coherent rotation of a rigid whole object was an inverse function of the angular difference as implied by the two perspectives views. In other words, the greater the angle through which rotational motion was seen, the slower the rate of alternation had to be.

This finding can be considered to be in keeping with one of Korte's Laws which states that optimum apparent motion is preserved when the spatial separation between presentations of a and b is increased by increasing the time interval between presentation of a and b. This law makes sense if one assumes that the perceived speed of rotation is constant. If therefore the mental representation of the object has to rotate through a greater angle, more time is required.

Further support for this interpretation is provided by an experiment which asked the following question: Is it the retinal spatial separation or the perceived spatial separation that governs Korte's Law? Perceived separation was varied by creating conditions in which a and b appeared at differing distance but were always located so as to project to the eye in the same retinal loci (Corbin, 1942; Attneave and Block, 1973). The experiments demonstrated that it was the perceived spatial separation, not the retinal separation that enters into Korte's Law.

A problem-solving theory can account for these facts. It offers an explanation of why motion is seen. Unlike other theories, it takes as a point of departure and is quite compatible with the fact that the conditions leading to motion perception entail change of perceived location rather than change of retinal location. It offers a rationale for the known facts about alternation, i.e. why movement is perceived only within a certain range of middle values of inter-stimulus interval. It can deal easily with the kinds of perceived transformations or movements that occur when a and b are more than single dots or lines, such as groupings or forms with sub-parts, or complex three-dimensional figures in differing orientations. Finally it permits us to predict instances where no motion will be perceived despite the maintenance of the spatial and temporal parameters that ordinarily produce the stroboscopic effect.

On the other hand this theory does not as yet explain all the known facts. It does not explain the reported findings that motion is seen more readily if both a and b are placed so that their projections fall within one hemisphere of the brain (Gengerelli, 1948); nor does it explain why the effect is more readily obtained if a and b stimulate one eye compared to the case where a stimulates one eye and b the other (Ammons and Weitz, 1951). However, these findings have never been replicated and warrant careful re-examination. And finally a problem-solving theory might be considered to be inappropriate as an explanation of the stroboscopic effect that seems to occur in decorticated guinea pigs (Smith, 1940) or newly born lower organisms such as

fish or insects (Rock, Tauber, and Heller, 1965).

However there now seems to be fairly good evidence that there are two kinds of apparent motion (Broddick, 1974; Anstis, 1980). One kind, referred to as the short-range process, occurs over very small angular separations of a and b. There is reason for believing that this kind may be based on motion-detector neurons responsive to a small shift in stimulation on the retina. The other kind, referred to as the long-range process, occurs over larger angular separation of a and b. This process is probably not based on the activation of motion-detector neurons. Most if not all of the evidence discussed above pertains to this long-range process. The short-range process thus seems to have a direct sensory basis whereas the long-range process seems to have a cognitive basis. In the light of this distinction, it is possible that the findings referred to in the previous paragraph are explicable in terms of the short-range process.

## Conclusion

At this point we should step back from these empirical studies and see what general lessons can be drawn as to the nature of theories of perception. If the visual system is to achieve a faithful representation of the physical world then the organization of its own processes must somehow mirror the organization of the world. Any theory of perception that does not take this point into account will ultimately fail.

In certain theories of perception, constancy and veridicality are fortuitous outcomes that occur only under some circumstances. This is not good enough. Both the logic of what the perceptual system must accomplish and the emperical evidence of what it does achieve demand a theory in which constancy is inevitable, not accidental.

Herein lies the danger of theories based on simple and limited physiological findings. Unless the physiological finding can be seen as part of a larger process that "homes-in" on reality in an inevitable way, that physiological finding is likely to be misunderstood. This is the problem with viewing lateral inhibition as an exaggeration or distortion process. If we cling to the photometer metaphor, to the assumption that fundamentally the visual system measures the intensity (and perhaps the wavelength) of the light at each point in the image, then it is not surprising that some kind of distortion process will be required to transform the array of photometer readings into something vaguely representing visual experience.

There is no need to talk as though the intensity of light at point A in the image is "affected" by the intensity of light at point B. The fact is that the light at point A, seen by itself, would be perceptually meaningless. Having a second intensity of light present in the visual field doesn't merely change the first amount of light, it literally establishes a relationship and the apprehension of this relationship produces lightness perception in its simplest form.

Contrast theories are usually thought to be relational theories, but they are not. As Koffka (1935) has correctly pointed out, the ultimate correlate of lightness perception in a contrast theory is still an absolute amount of light, not a relationship. The contrast process only allows the absolute value of one region of the field to be changed as a function of other values. But it is still that absolute value that reigns. And the reason that it has to be changed is that as an absolute value, it will always be out of touch with visual experience, which involves relationships. In fact the history of theories of lightness perception is the history of different correction factors designed to bring the local luminance into correlation with perceived lightness. This has never worked and it is time we recognize that no theory based on absolute amounts of light can work. What is constant about a white surface, for instance, is its relationship to the rest of its environment.

It is not surprising that the visual system gets its critical information from the edge. This is the point in the image where the relationship between two amounts of light is represented. In more complex scenes each local edge relationship, or ratio, has to be seen in relation to other edge ratios. The concept of edge integration that we have discussed does not involve any distortion or exaggeration process. Rather it involves the proper organizing of certain local relationships in order to make explicit a more global relationship that was only implicit in the local relationships. At the level of cognition the same function is served formally by the syllogism.

Fundamentally the visual system must be logical because the world is logical. The world is not put together in a random or capricious way. If a rectangular object is incapable of obscuring a set of diagonal lines it will also be incapable of obscuring a luminous spot. How inefficient it would be if local percepts were allowed to coexist with other local but contradictory percepts. A system that excludes contradictions from its global relationships has the tremendous advantage of reducing the ambiguity of its local relationships.

Perception and cognition seem to share this quality of excluding contraditions within their own domains. Of the two, however, perception seems to be the more successful. Of course it is possible to construct figures such as impossible triangles, or Escher drawings, which surprise us by the extent to which visual contradictions are tolerated. But it is the rareness of such visual contradictions that leads to our delight at such figures. Examples in which the cognitive system fails to exclude contradictions are unfortunately too numerous to mention. One only needs to turn to political speeches, or the Bible, or journal articles to find a gold mine of examples.

Seeing, then, is like thinking, at least in many of its formal properties, and this may be because thinking is like seeing. That is, seeing may be the primitive form of thinking, the basic prototypical form that shows how relationships are to be integrated in order to correctly represent the world. Seeing, of course, had to come first, and it must be there even in mosquitoes. Thinking, however, allows us to integrate relationships that extend beyond the time and space limitations of the visual system.

Perhaps the world looks the way it looks because we are what we are physiologically. But it should not be forgotten that the world existed before we did and thus, as we learn from evolution, we are what we are because the world is what it is.

# 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.

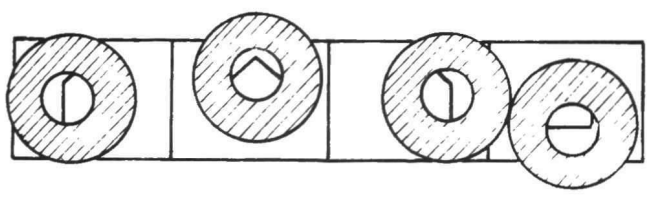

Figure 1. A cartoon strip showing a peephole moving around the outline of a shape. The fact that successive frames in the cartoon fall in different positions makes the task harder.

### 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.

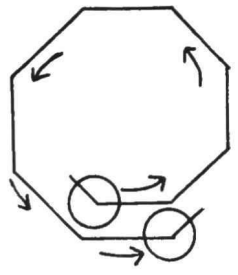

Figure 2. A peephole is moved around a polygonal spiral without revealing the free ends or the adjacent parallel sides.

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.

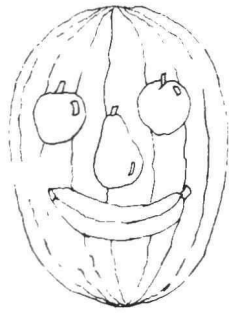

Figure 3. A face composed entirely of pieces of fruit. (After Palmer, 1975)

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.

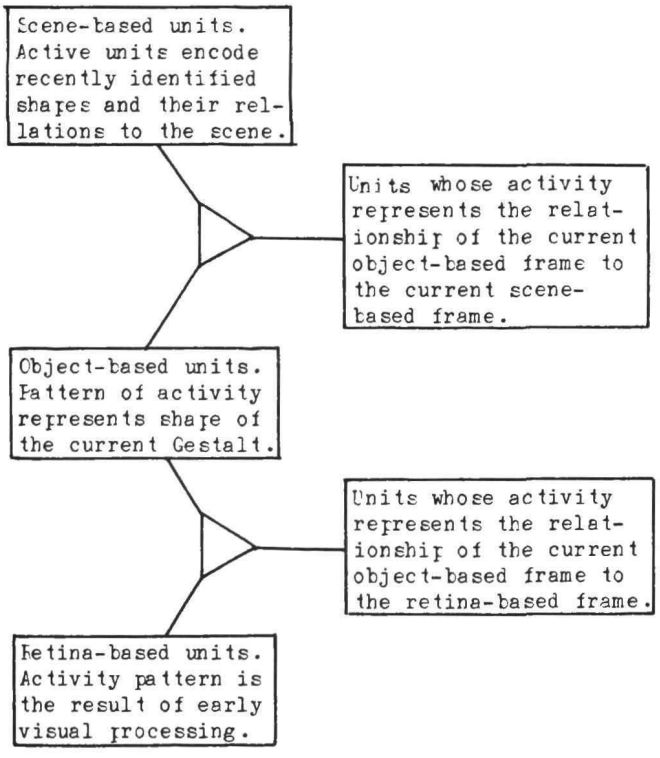


Figure 4. A parallel mechanism.

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.