SHAPE RFPRFSFNTATTON IN PARALLEL SYSTEMS

Geoffrey F. Hinton

MFC Applied Psychology Unit
Cambridge, Fngland

APSTRACT

There has been a recent revival of interest in parallel systems in which computation is performed by excitatory and inhibitory interactions within a network of relatively simple, neuronlike units [1 2 3 4]. At the early stages of visual processing, individual units can represent hypotheses about how small local fragments of the visual input should be interpreted, and interactions between units can encode knowledge about the constraints between local interpretations. Higher up in the visual system, the representational issues are more complex. This paper considers the difficulties involved in representing shapes in parallel systems, and suggests ways of overcoming them. In doing so, it provides a mechanism for shape perception and visual attention which allows a novel interpretation of the Gestalt slogan that the whole is more than the sum of its parts.

1 INTRODUCTION

The most notorious failure of the Gestalt psychologists was their inability to specify a plausible mechanism to explain the many important and insightful perceptual phenomena that they discovered. Cognitive Science has rediscovered many of the phenomena. Can it do any better with the mechanism[9] We have the advantage of modern digital computers which can simulate any mechanism we care to invent, but what kind of mechanism should we be looking *for?* Is the digital computer itself a good analogy, or should we be investigating the computational properties of processes occuring in parallel systems of richly interconnected, neuronlike units?

The central idea of Gestalt psychology is the Gestalt itself — a coherent organisation of the parts of a figure into a perceptual whole which transcends the individual parts. The central idea of this paper is that the mechanism underlying the formation of a Gestalt is a set of competitive and cooperative interactions within a network of simple units. The interactions result in a particular subset of the units becoming active and suppressing the rest. The active subset is the internal representation of the current Gestalt.

This is not a new idea and it has many problems. How does the Gestalt represent shape independently of size, position, and orientation[9] How are successive Gestalts integrated in the temporal flow of perception[9] How is the Gestalt for the whole related to the Gestalts for its parts? How, exactly, are Gestalts encoded as activity in the units of a parallel system[9] Before discussing these issues, T shall briefly describe the historical ups and downs of parallel models in computer vision, and also give a recent example of a parallel model that illustrates many of the problems.

11 PARALLEL MODELS IN COMPUTER VISION

There is a long tradition of attempts to build neural models *of* visual perception. Much *of* the early work was unconvincing because it was based on an inadequate analysis of what a visual system must do. It ignored the main problems like segmentation or generating a 3-D representation from a ?-P image. The inadequacies of the existing neural models led people in Artificial Intelligence to abandon them and to concentrate on the problem of programming a visual system on a conventional digital computer.

Work in computer vision has now given us a much better grasp of what the real problems *are* in getting from intensity arrays to the kind of articulated internal representations of 3-D scenes that are needed for object recognition and manipulation. We have learnt, for example, that segmenting a real scene into objects is hard, and that it cannot be done properly by simply looking for edges or growing regions in the raw intensity array produced by *a* camera.

For a time, it appeared that a major problem was to develop complex heterarchical control structures that would allow high-level knowledge about objects to aid the low-level interpretation of the intensity array [5 6]. This view has now been largely superceeded by two related developments. First, people in computer vision who studied real images rather than line drawings rediscovered the Gibsonian point that there is a geat deal of available information in the intensity array, especially if sources of information like stereo and optical flow are considered. Second, David Marr [7] emphasised that low-level visual processing in the brain involves an enormous amount of parallel computation at or near the level of the intensity array. So techniques designed to economise on the number of computational operations requiresd in a sequential computer may be a poor guide to understanding how natural visual systems work.

The problem of segmenting a scene into objects is a testing ground for these new developments. Before segmentation occurs, it appears that a great deal of "low-level" visual processing must be done. The purpose of this processing is to interpret the intensities of each pixel in the image in terms of the local surface orientation, reflectance, and depth of the piece of 3-D surface thet is imaged in the pixel. These intrinsic properties of the surface *are* much more useful for segmentation than the raw intensity data, because they distinguish intensity changes caused by discontinuities in depth from similar intensity changes caused by surface markings or sharp changes in surface orientation.

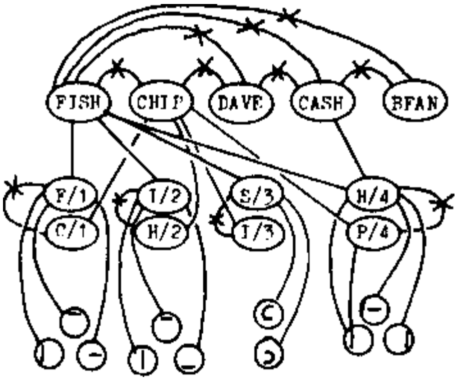
Some of the algorithms that are used for recovering intrinsic properties of surfaces from local intensities [4] or from stereo pairs of images [3], have a very interesting property. They involve local computations that can be performed in networks of interconnected simple units. Thus, for low-level processing, computer vision is moving back to models in which processing occurs in pseudo-neural networks. The current models differ from earlier neural-net models in several ways. They are rigorously specified, and the details of the computation are typically determined by careful analyses of the physics of the image formation process [8] and of the general properties of the physical world that determine how the properties of one piece of surface constrain the probable properties of neighbouring pieces [3].

This paper discusses the problems involved in extending this kind of parallel computation to higher levels of visual perception like shape recognition which was the central preoccupation of the earlier generation of neural models like perceptrons [9]. These problems are often ignored or brushed aside by parallel modelB of shape recognition 1ike recognition cones |10j or hierarchical relaxation [11]. They must be solved before this kind of model can be accepted as a plausible account of human shape perception.

III AN <u>EXAMPLE</u> OF A <u>PARALLFL SYSTEM</u>

To illustrate the kind of parallel system that 1 will be discussing, T have chosen a recent model of word perception. The model is limited to the perception of briefly presented four letter words, but it works, it fits the psychological data well, and its limitations provide a good starting point for discussing the problems that more general systems of this type will have to overcome.

When a string of letters is presented very briefly, it is easier to recognise the letters if they form a word than if if they form a nonsense string. Rumelhart and McClelland (henceforth RAM) propose a model in which many simple, neuronlike units interact to produce this effect [12 13]. For simplicity, they restrict themselves to a three-layered system, and they omit feedback from the middle layer to the bottom one (see Fig. 1).



<u>FIGURE</u> 1

Only a few of the units are shown for each of the three layers. Inhibitory interactions are marked with a cross. A single line is used to stand for connections in both directions.

The bottom layer contains "stroke" units that detect local features like the individual strokes of letters in specific positions within the word. A unit in this layer might, for example, be activated if there is a vertical stroke that could be the right hand vertical of an H, M, or N in the second-letter position within the word. Fach letter unit receives excitatory input from all the stroke units that fit it and also inhibitory input from stroke units in the same position that do not fit it.

Units in the top layer correspond to specific words. Each word unit receives excitatory inputs from all the letter units that fit it and inhibitory inputs from the rest. Word units also provide excitatory and inhibitory feedback to the letter units. In addition to these interactions between layers, there are inhibitory interactions between all pairs of word units and between those pairs of letter units that correspond to alternative letters at the same position within a word.

The activity level of a unit is a continuous variable constrained to lie between two limits, and the precise rules for the excitatory and inhibitory interactions and for the thresholds are quite complex. *They* are chosen so that when the stroke units are activated as they would be by a visually presented word, the system settles down into a stable state in which the appropriate word and letter units are highly active, and the inappropriate unite are suppressed.

Precise rules for the interactions can be chosen BO that the model is in good agreement with experimental data for a wide range of experiments. It can, for example predict the way in which the probability of correctly reporting a particular letter depends on the onset and offset times of the other letters.

IV PROBLEMS FOR PARALLEL MODELS OF SHAPE PERCEPTION

The R & M model has several interesting limitations which are characteristic of a whole class of models in which shape perception is performed by parallel computation in a network of relatively simple units:

1. The model makes no provision for variations in the size, position, or orientation of the word. It implicitly assumes that the input is somehow normalised so that the actual size, position, and orientation of the word do not affect which of the stroke units are activated by the input. To put it another way, activation of a particular stroke unit represents the existence of a stroke of a particular type in a particular position relative to the whole word. At the lowest levels of the visual system, however, it is the position, size, and orientation of features relative to the retina that determines which units are activated. How to transform from features relative to the retina to features relative to the whole word is a major problem.

2. The Gestalt for a whole word is implemented as a pattern of activity in which the active stroke, letter, and word units all support one another and suppress the rest. To perceive another word, a different pattern of activity must emerge in the very same set of units, so the representation of the previous word must be wiped out. This makes it hard to see how successively perceived Gestalts can be integrated into higher level wholes [15]. The only way to save the principle that different Gestalts are implemented as alternative patterns of activity in the very same set of units, is to introduce some kind of spatial working memory which keeps a compact record of recent Gestalts separately from the apparatus that is used for forming Gestalts. The contents of this working memory presumably act as a context that influences the formation of new Gestalts, and in extreme cases allows a new Gestalt to be formed purely on the basis of the contents of working memory without any further perceptual input (as happens when people "see" a whole object after examining it by moving a small peephole over its various parts). A comprehensive parallel model needs to specify how spatial working memory is implemented with neuronlike units, and how the contents of working memory influence the formation of new Gestalts.

J. The RAM model requires a separate unit for each possible relationship of a stroke or letter to the whole word. •This duplication of feature units over all discriminable relationships requires a lot of units. It is not too bad in the case of word perception where the number of possible positions of a letter within a word is small and the number of letter types is also small, but for other kinds of shape perception it could prove very expensive to use a different unit for each possible relation of a feature type to the whole object. The kind of model being proposed would be more plausible if there was some encoding scheme which achieved the effect of having separate units for each possible relation of a feature to the whole without requiring as many units as this seems to imply.

In most simulations, the number of units is not a problem, because only a very small fraction of the possible features are present at once, and they can be represented by data-structures containing numerical values that code the relation of the feature to the frame of reference. The interactions between feature representations can be implemented by using a general procedure which takes these numerical values into account. Unfortunately, this way of coping with the huge number of possible features relies on the ability of the digital computer to perform arithmetic, and it therefore hides a very real problem for truly parallel systems.

In a network of neuronlike units, the interactions between features are achieved by direct connections [14 ' rather than by repeated application of a single parameterised procedure that determines the effect of one feature on another as a function of the numerical parameters of the two features. But it is the use of a single general procedure that enables a simulation program to avoid keeping data-structures for all the possible but currently absent features. If all the required interactions between features are coded by connection strengths between hardware units, rather than by a general procedures, it appears that all the units for all possible features must be present all the time. So how can we avoid having a very large number of units for each type of feature?

These three problems — normalisation, integration of successive Gestalts, and efficient encoding of relative features are the topics of the rest of this paper.

V VIEWPOINT AND SHAPE CONSTANCY

We see an object from different viewpoints on different occasions. On each occasion it has a different retinal image, and yet we generally recognise it as having the same shape. To do this, we have to cope with two quite different difficulties. First, parts of an object may be hidden or partially hidden due to self-occlusion or occlusion by other objects. So we must be able to recognise the object from the subset of its parts that is visible and their interrelations. Second, the metrical properties of the parts and relationships that are visible in the image depend on the viewpoint. The size, orientation, and position of an edge in the image depends as much on the viewpoint as on the properties of the corresponding edge in the external object. I shall focus on the second of these difficulties.

Artificial Intelligence has been dominated by a particular approach to these problems that can be traced back to Roberts [16] and is probably-most widely known in the theory of shape representation proposed by Minsky in his frames paper [l7]. The variations in the metrical properties of the images of parts of an object are handled by using "topological" categories. If, for example, an object has a fully visible flat surface with three straight sides, then its image will contain a triangular region. The shape of the triangle in the image depends on the precise viewpoint, but the fact that it is a triangle does not. So what is meant by "topological" in this context is not the usual mathematical sense (invariant under any continuous transformation), but the somewhat stronger property of being invariant under projection, and hence not affected by viewpoint. Relationships between the different parts of an image are likewise handled by using discrete category labels like "connected to"" or "above" or "behind". Again these categories are typically unaffected by small changes in viewpoint.

By using categorical labels for parts and their relationships, an image can be reduced to a relational network that is then matched against stored models. Since relational labels like "behind" *are* relative to the viewer, and since different topological features are visible from different viewpoints, several different models are typically needed for each object. The advantage of this approach is that the great wealth of metrical information in an image is reduced to a compact description which can be matched against similarly compact Btored representations. Its disadvantage is that this reduction of information faisB to utilise a powerful constraint on the interpretation of an image -- the single viewpoint constraint.

The relationship between an object and the viewer determines how each part of the object appears in the image. Conversely, when part of an image is interpreted as depicting part of an object, this puts constraints on the relationship between the object and the viewer. Since every retinal or TV image is formed from exactly one viewpoint, the interpretations assigned to the various parts of an image must agree on what that viewpoint is.

Some computer vision programs [16] make use of the single viewpoint constraint as a final check on the interpretation of an Image. They first extract a relational network of topological features and use it to suggest a particular 3-D model. Then they compute the relationship between the viewer and the object by using precise metrical information about a few points in the image and in the stored 3-D model. Finally, they use this computed relationship to project the 3-D model back onto the image. The fit of the projected model with the original image acts a check on the interpretation. This approach is rather sensitive to inaccuracies in the image, but it has been refined by [18] who describe a neat way of discovering the optimal viewpoint, i.e. the one which gives the best overall fit between the original image and the image produced by projecting the stored 3-D model.

Using the single viewpoint constraint as a final check after a particular 3-D model has been hypothesised is better than not using it at all, but it would be more efficient to make use of the constraint to prevent inappropriate 3-D structures from being hypothesised in the first place. To show how this can be done, I need to introduce the concept of an object-based feature.

A. Object-based Features

One way of ensuring that the underlying representation of the shape of an object is independent of viewpoint is to impose a canonical frame of reference on the object and to describe the sizes, positions, and orientations of the parts of the object in terms of this object-based frame. This technique allows an object to be described in terms of a constant set of object-based features and hence to be recognised whatever its size, position, and orientation. If a different object-based frame is imposed, a different set of object-based features will be obtained. This explains why a single object can have several phenomenal shapes. An upright diamond, for example, may also be seen as a tilted square.

A considerable amount of early processing must occur before object-based features are extracted, because an object must be segmented out from the rest of the image before a frame of reference can be imposed on it. In normal circumstances, the problem of getting from a 2-D image to a 3-D representation may be solved by this early processing before segmentation occurs and hence before object-based features are extracted. Put simply getting 3-D features does not solve the problem of shape constancy. The 3-D features generated by early processing are retina-based. In other words, their sizes, positions and orientations are defined relative to the frame of reference of the retina (or camera). If the viewpoint is changed, the 3-D retina-based features produced by an object also change, so they do not constitute a shape representation.

The relationship between the imposed object-based frame and the viewer determines the optical mapping from features of the object to features on the retina. Hence an internal representation of this relationship can be used to govern the mapping from retina-based to object-based features (see Fig. 2). Each possible viewpoint specifies a particular set of pairings between retina-based and object-based features. Conversely, each consistent set of pairings specifies exactly one viewpoint, retails of one possible scheme for implementing the structure shown in Fig. 2 in a network of simple units are given in [19]

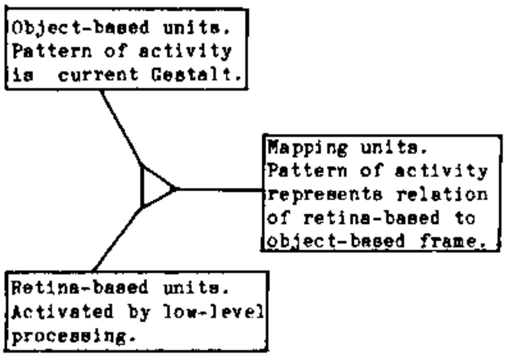


Figure 2

One interesting aspect of this way of achieving shape constancy is that it requires an extension to the normal way of thinking about the global structuring of parallel systems. Instead of allowing groups or layers of units to interact directly with other groups or layers, we have introduced a three-way interaction in which activity in one group controls the way in which two other groups interact. The idea of a variable mapping between feature sets recurs later. Again, the feature sets involved are features relative to different frames of reference, and the mapping is controlled by a representation of the spatial relationship between the two reference frames.

VI HIERARCHICAL STRUCTURAL DESCRIPTIONS

So far, I have been assuming that people only impose one object-based frame of reference at a time. *This* appears to conflict with the widely held view tha.t people use hierarchical structural descriptions in which there is a node for each object that is linked to lower-level nodes for its parts. These lower-level nodes, in turn, are linked to nodes for their parts, and so on until a level of primitive entities like edge segments is reached. Each node in a structural description haB its own associated object-based frame of reference, and each link between two nodes is labelled with the spatial relationship between their two object-based frames [20 21]. The great value of hierarchical structural descriptions as spatial representations is demonstrated by their use in computer programs for graphics [22] visual recognition, spatial manipulation, and spatial reasoning with innaccurate data [23].

Structural descriptions seem to explain many interesting effects in human perception and imagery [24]. However, there is little evidence that the whole of a complex structural description is actively represented at the same time. It may well be that our attention flits between levels and that at each moment, we only *focus on* one node, i.e we impose the object-based frame appropriate for this node and form a Gestalt for it. This sequential theory, raises several problems: How can there be a Gestalt for the whole without Gestal ts for the parts also being present, and how can successively perceived Gestalts be integrated into a larger wholes[9]

Before answering these questions T need to correct the common misapprehension that a hierarchy of active object-based features is equivalent to, or is an implementat ion of, a structural description.

A Structural rescriptions and Feature Hierarchies

One important difference between a hierarchical structural description and a hierarchy of active object-based feature units is that each link between nodes in the structural description is labelled with an explicit spatial relationship, whereas there are no explicit representations of the spatial relationships between the various object-based features. An object-based feature unit is activated by the combination of a particular feature type with a particular relationship to the global object-based frame of reference. The type of a feature and its relationship to the global reference frame are not separately encoded. This means that higher-level feature units can be activated directly by combinations of lower-level ones. They do not need to check the relationships between these lower-level features, because the relationships are implicitly encoded by which of the lower-level units are active.

The absence of explicitly represented spatial relationships may seem like a rather minor point, but it allows hierarchies of object-based features to avoid the computational complexities of graph matching. The cost, of course, is that for each type of feature, there must be a separate unit for each discriminable relationship of a feature of this type to the global object-based reference frame. The duplication of object-based units of a given type for all different positions, orientations, and sizes can be viewed as a way of using parallel hardware to avoid the graph-matching problem by avoiding representations of relationships that are separate from the things being related.

VIT VHOLFS ANP PARTS

The Gestalt psychologists were fond of saying that the whole is more than the sum of its parts. Most information processing theories have interpreted this slogan to mean that in addition to the representations of the parts, there is a higher-level representation for the whole that is separate from, but connected to, the representations for the parts (as in a hierarchical structural description). There is, however, a far more radical interpretation of the Gestalt siogan: When we attend to a whole we do not see its parts as wholes because the representation of the whole does not in any way involve or require the representations of the parts as <u>wholes</u>. When a part is seen as a constituent of a larger whole it is given a quite different internal representation from the one it has when it is seen as a whole in its own right.

The view that there are two quite different ways of representing an object, as a whole or as a constituent of a larger whole, is a surprising reoult of considering a problem that is peculiar to parallel systems: How is the representation of a shape related to the representations of the particular parameter values (e.g. its size and position) that distinguish different instances of the same shape. In a conventional computer, this is not a problem because a data-structure can be created for the instance containing separate fields for the shape and for each parameter value. The inapplicability of this method to parallel systems has already been discussed at the end of section IV.

In a parallel system like the brain, there appear to be two main ways of relating the representation of a shape to the representations of the parameter values that distinguish different instances of the shape. If only one instance is represented at a time, the values of properties of the instance, like its size and position, can be associated with the shape of the instance by simply activating separate representations for the shape and for each of its specific property values all at the same time. The only thing that binds the separate representations together is their simultaneous activation. This method has the great advantage that if different instances of the same shape are presented on different occasions, the very same set of active units will be used to encode the shape information. When an object is seen as a Gestalt, simultaneous activation can be used to bind a representation of its shape to separate representations of properties like its size and posi tion.

The method of simultaneity has the advantage that the *very* same representation of the shape is active whatever the values of the other properties. So this representation explicitly captures what it is that all instances of the same shape have in common, and it therefore explains how learnt associations like the name of the shape can be generalised from one instance to other instances with different sizes, positions, and orientations. Unfortunately, the method of simultaneity has the disadvantage that it will not work if more than one instance must be represented at a time, and that is a major motivation for the "one Gestalt at a time" principle. If, for example, the representations for "large", "circle", "small", and "square" are all active at once, mere simultaneity cannot indicate which size goes with which shape.

The second method of binding shapes to their property values involves using multi-dimensional units, each of which responds to a conjunction of a particular shape with a particular set of property values. This kind of representation is used in the RAM model at the letter level. For each combination of a particular letter with a particular position within the word, there is a particular dedicated unit. This method allows many instances to be represented at the same time, but it requires a large number of units, and by coding different instances of the same shape as activity in different units, it fails to capture what is common to ell the instances of the shape. For example, in the RAM model the letter H is encoded quite differently in the two words FISH and CHIP. This difference, however, is a positive advantage because it allows the two instances of the H to have quite different effects at the word level. One supports the word FISH and the other supports CHIP. Thus multi-dimensional coding allows the effects of different instances of the same shape to be tailored to the particular property values of the instance (relative to the global object-based frame) . This is the primary motivation for thinking that when instances are perceived as constituents of a Gestalt they are encoded by multi-dimensional units.

To summarize, there are two quite different ways of binding together the shape and other properties of a particular instance in a network of neuronlike units. When an instance is perceived as a Gestalt, the method of simultaneity can be used. This allows the very same active units to be used to represent the shape of an instance whatever its other properties. When an instance is seen as a constituent of a larger Gestalt, however, the multi-dimensional method is used. This allows many constituents to be coded at once, and it allows the effects of each constituent to depend on its particular parameter values relative to the whole. The representation of an instance when it is seen as a Gestalt is therefore quite different from its representation when it is seen as a constituent of some larger whole. The Gestalt for the whole does not in any way involve the Gestalts for its parts.

VIII <u>SPATIAL WORKING MEMORY</u>

If we accept the principle of one Gestalt and one object-based frame at a time, there is a serious problem of piecing together successive Gestalts. This problem is *at* its most severe when the parts of an object are observed sequentially through a peephole, and a new Gestalt for the whole object is formed from these fragmentary glimpses.

The role of the hierarchy of object-based feature units is to allow a Gestalt to be formed. Once this has been done, a more compact record of the shape of the Gestalt and of its else, position and orientation can be kept in the form of activity in a different set of hardware units which I shall call the "scene-buffer". A number of these records may be accumulated in the scene-buffer, and they can act as a context which influences the formation of new Gestalts from the perceptual input. If, for example, one part of an object has been seen as a Gestalt in its own right, the corresponding record in the scene buffer will facilitate certain of the object-based feature units when the Gestalt for the whole object is formed.

The position, orientation and size of a Gestalt must be represented relative to some frame of reference. One possibility is the retinal frame of reference. The relation of the Gestalt to the retinal frame is needed anyway to determine the mapping from retina-based to object-based features. The retinal frame, however, is not very useful for the perceptual integration of Gestalts formed at different times because the retina moves around in the world. What is needed is a stable contextual frame of reference defined by the scene (this argument is elaborated in [25].

The combination of the shape of a Gestalt and its relation to the scene can be represented by activating a particular "scene-based" feature unit. Records of many different Gestalts can be stored at the same time provided the units in the scene-buffer use the multi-dimensional method for binding the parameters of a Gestalt to its shape.

The mapping from the higher object-based features to the scene-based features can be handled by just the same kind of mapping apparatus as was used for relating retina-based and object-based features. 3y allowing the mapping to work in both directions, it is possible to implement the contextual effects of existing scene-based features on the creation of new Gestalts. Pig. 3 summarises the various sets of features that have been invoked and the interactions between them.

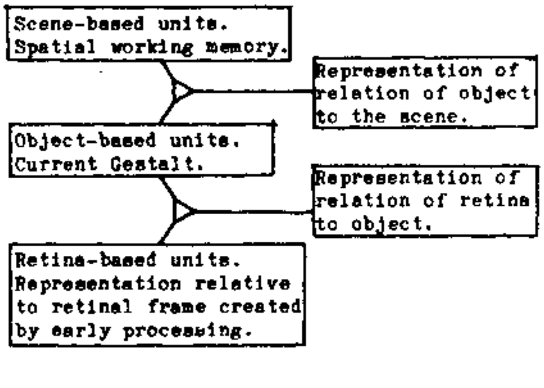

Scene-based units.
Spatial working memory.
Representation of relation of object to the scene.
Object-based units.
Current Gestalt.
Representation of relation of retina to object.
Retina-based units.
Representation relative to retinal frame created by early processing.

Figure 3

IX ENCODING MULTI-DIMENSIONAL FEATURES

If variations in size, position, and orientation are taken into account, the number of possible features is enormous. The relationship of a 3-D feature of a particular type to a frame of reference can vary along 7 dimensions (3 for position, 3 for orientation, 1 for size). So if there *are*, say, 10^2 discriminable values along each dimension, there are 10^{14} possible particular features. Is there any way of achieving the same accuracy with less united

In information theoretic terms, it is very inefficient to have a unit for each possible feature if only a very small fraction of the possible features are present at any one time. It would be much more efficient to use an encoding in which a much larger fraction of the units were active at any moment. This can be done if we abandon the naive idea that each specific feature is represented by activity in exactly one unit. Instead each unit can be more *coarsely* tuned so that it is activated by a range of possible features, and the ranges of different unite can be made to overlap so that each feature activates many different units. The representation of a particular feature then becomes a pattern of activity in many units, and similar features are represented by similar patterns of activity. Even though each unit is coarsely tuned and therefore rather imprecise about the *exact* parameters of the feature that activated it, the whole set of units activated by a particular feature codes the parameters of the feature very accurately. To get an idea of the efficiency of this "coarse-coding" scheme as compared with the naive method in which each discriminable feature is coded by its own unit, we need to jump into hyperspace.

For a given type of feature, the possible relations to a frame of reference form a seven-dimensional space. Fach particular feature corresponds to a point in this space. The naive encoding is equivalent to dividing the space into small, non-overlapping zones, and using one unit for each zone. *The* coarse-coding scheme divides the space into larger, overlapping zones. For simplicity, I shall assume that the zones are hyperspheres, that their centers have a uniform random distribution throughout the space, and that all the zones used by a given encoding scheme have the same radius. What we *are* interested in is how accurately a feature is represented as a function of the radius of the zones. Is it better to have large zones with each feature point falling within many zones and hence being coded by activity in many units, or is it better to have the same number of smaller zones so that a feature is represented by activity in fewer but more finely tuned units?

One way of expressing the accuracy with which the parameters of a particular feature are encoded is to ask what the probability is that two similar features (presented on different occasions) will receive different encodings. For the encodings to be different, there must be at least one zone that contains one feature point and not the other. If the zones have a radius of r, then the centres of all the zones that contain a given point fall within a hypersphere of radius r centered on that point. So for points P and 0 in Fig. 4 to receive different encodings, there must be at least one zone whose center falls in one of the hypersheres around P and 0 but not in the other, i.e. there must be a zone with its center in one of the two shaded "hypercrescents".

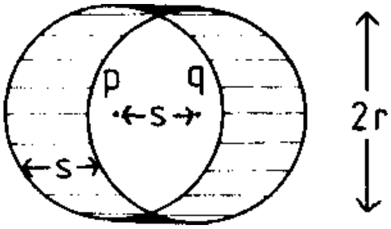


Figure 4

The probability of there being at least one zone center within the hypercrescents is completely determined by the expected number of zone centers within the hypercrescents. This number is the product of the volume of the hypercrescents and the density of zone centers throughout the space. As the volume of the hypercrescents is increased, the density of zone centers can be decreased proportionately without affecting the probability that the two features receive different encodings. Hence, the number, N_r , of zones of radius r that is required to achieve a given accuracy is inversely proportional to the volume of the hypercrescents.

If the separation of the features under consideration is small compared with the zone radius, then the solid areas in Fig. 4 are negligible and, in two-dimensional space, the area of each crescent is approximately the same as for a rectangle of height 2r and width s, because the horizontal distance between the sides of a crescent is exactly s except at the very top and bottom. So in 2-D the area of a crescent is proportional to r. In 3-D, the two surfaces bounding the 3-D "crescent" again have a separation of s in the direction of the line joining the two feature points. So the 3-D "crescent" can be divided into many narrow rods of length s. These rods can be rearranged into a disk in same way as the horizontal strips are rearranged into a rectangle in the 2-D case. So the volume of each 3-D crescent is the same as that of a disk of thickness s and radius r. This is proportional to r^2 . In k dimensions, each hypercrescent has a volume of s times the k-1 dimensional cross-section, which is the volume of a k-1 dimensional hypersphere. Hence, in k dimensions N^r. <* $1/r^{k-1}$, provided s<<r. (it is hard to turn this argument into a formal proof because the extent by which r must exceed s to make the solid area in Fig. 4 negligible depends on the dimensionality of the space).

This unexpected result makes it much more difficult to dismiss models because they require too many units. By encoding features as patterns of activty in many coarsely tuned units, it is possible to have many more discriminable features than there are units. Similarly, the representations of the mappings between reference frames can be economically encoded by using coarse-coding in the space of possible mappings.

It is probably no accident that sensory neurons are typically much more broadly tuned than might be expected from the accuracy of an animal s perception. Far from causing inaccuracy, this broad tuning is a way of increasing the accuracy of a representation given a fixed number of available units.

Apart from boundary effects, there are two factors that set upper limits on the sizes of the zones. If many similar features occur at the same time, their encodings may overlap. This is not fatal if the activity level of a unit reflects the number of features that fall within its zone, but generally nearby features will affect each others encodings. So zone sizes should be chosen so that not more than a few features fall within a zone at any one time. Thus the value of the coarse-coding technique relies on the features being relatively sparse.

The other limit on zone sizes stems from the fact that the representation of a feature must be used to affect other representations. There is no point using coarse-coding if the features have to be recoded as activity in finely tuned units before they can have the appropriate effects on other representations. The details of this argument are complex, and there is not space for them here, but the conclusion is that coarse-coding can be used provided the required effects of a feature are approximately the average of the required effects of its neighbours. At a fine enough scale this is nearly always true. *The* scale at which it breaks down determines an upper limit on allowable zone sizes.

X CONCLUSION

This paper has explored the issues that arise from the assumption that perceiving a shape as a whole involves a cooperative computation in which a stable pattern of activity emerges in a network of units as a result of the external input and the interactions between the units.

Shape representations that are independent *of* viewpoint can be achieved by using two different sets of features, one relative to the retina and the other relative to a frame of reference imposed on the object. The interactions between features in the two sets are controlled by a representation of the relation between the fremes.

Two ways of binding a shape to its parameter values (e.g. size, position) are described. One method can only be used for one shape at a time, and so it is suitable for the Gestalt, but not for its many conetituents. This leads to the idea that when an object is seen as a constituent of a larger whole, it receives a quite different internal representation from the one it has when it is seen as a Gestalt in its own right.

The stable pattern that represents a Gestalt can be recoded as activity in a different set of scene-based features, thus freeing the object-based features *for* the formation of a new Gestalt. This recoding again involves a flexible mapping between sets *of* features relative to different frames of reference. The scene-based features act as a spatial working memory which influences the formation of new Gestalts.

Finally, a coding scheme is presented which allows efficient and accurate encoding of sparse, multi-dimeneional features by using patterns of activity in coarsely-tuned units.

FFFFFFNCFS

[1] Minsky, M. K-lines: A theory of memory. Cognitive Science, 1980, 4, 117-133

[2] Hinton, G. F. A Anderson J. A. (Eds.) Parallel models of associative memory Hillsdale, NJ: Fribaum , 1981.

[3] Marr P. A Poggio T. Cooperative computation of stereo disparity. Science 1976, 194, 283-287.

[4] Barrow, H. G. A Tenenbaum, J. M. Hecovering intrinsic scene characteristics from images. In A. P. Hanson A F. M. Piseman (Eds.) Computer vision systems. New York: Academic Press, 1978.

[5] Shirai, Y. A context sensitive line finder for recognition of polyhed ra. Artificial Intelligence, 1973, 4, 95-119.

[6] Freuder, F. C. "A computer system for visual recognition using active knowledge." AI-TF-345, A.I. Laboratory, MIT, 1976.

[7] Marr, P. Early processing of visual information. Phil. Trans. Poy. Soc. Series B, 1976, 275, 483-524.

[8] Horn, B. K. P. Understanding image intensites. Artificial Intelligence, 1977, 8, 201-231.

[9] Rosenblatt, F. Principles of neurodynamics. Washington P. C.: Spartan, 1961 .

[10] Uhr, L. "Fecognition cones", and some test results; In A. P. Hanson A F. M. Piseman (Eds.) Computer vision systems. New York: Academic Press, 1978

[1 0 Davis, L. S. A Rosenfeld, A. Hierarchical relaxation for waveform parsing. In A. F. Hanson A F. M. Piseman (Eds.) Computer vision systems. New York: Academic Press, 1078.

[1?] McClelland, J. I,. A Rumelhari, P. F. "An interactive activation model of the effect of context in perception: Part 1", Technical Report 91 , Center for Human Information Processing, Univ. California, San Piego, 1980.

[13] Fumelhart, P. F. A McClelland, J. 1,. "An interactive activation model of the effect of context in perception: Part 2", Technical Report 95 Center for Human Information Processing, Univ. California, San Pi ego, 1980.

[14] Feldman, J. A. A connectionist model of visual memory. In [2] above.

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

[16] Roberts, L. G. Machine perception of three-dimensional solids. In J.T. Tippett et . al. (Fds.) , Optical and electro-optical information processing. Cambridge, MA: MIT Press , 1965.

[17] Minsky, M. A framework for representing knowledge. In P. H. Winston (Ed.), The Psychology of Computer Vision. New York: McGraw-Hill, 1975.

[18] Barrow, H. C. , Tenenbaum, J. M. , Polios, P. C. , A Wolf, H. C. "Parametric correspondence and chamfer matching: Two new techniques for image matching." In Proc. IJCAI-77. Cambridge, MA, August, 1977, 659-66?.

[19] Hinton, G. F. "A parallel computation that assigns canonical, object-based frames of reference." This proceedings.

[20] Palmer, P. F. Hierarchical structure in perceptual representation. Cognilive Psychology, 1977, *9,* 441-474

[21 J Marr, P. A Nishihara, H. K. Representation and recognition of the spatial organisation of three-dimensional shapes. Proc. Roy* Soc. Series B, 1978, 200, 269-294.

[22] Newman, W. A Sproul 1 , P. Principles of interactive computer graphics. New York: McGraw-Hill, 1979.

[23] McPermott, P. Spatial inferences with ground, metric formulas on simple objects", Research Report 173, Computer Science Pept. Yale University, New Haven, CT, Jan 1980.

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

[25] Hinton, C. F. Paper to appear in the Proceedings of the 3rd Annual Conference of the Cognitive Science Society, Berkeley, 1981