Barlow, H.B. (1993). Object identification and cortical organisation. In E. Ottoson, B. Gulyas & P. Roland (Eds.), *The Functional Organisation of the Human Visual Cortex* (pp. 75-100). Oxford: Pergamon Press.

Object Identification and Cortical Organization

HORACE BARLOW

Physiological Laboratory, Cambridge CB2 3EG, UK

The starting point for what I have to say is a puzzling fact that has been evident since the 1950's. Whereas crustaceans, insects, frogs, rabbits and birds have highly selective feature detectors at early levels in their visual systems, this does not seem to be the case in cats and monkeys, which have highly developed *cortical* visual systems. Why is visual information coded differently for the cortex than for other visual systems?

This is part of a larger puzzle: In biology and comparative anatomy courses we happily say that the success of mammals, primates, and especially human primates, is attributable to the "neopallial explosion"–the vast increase in size of the neocortex that occurred during our evolutionary history. But then when we come to teach the physiology of the neocortex we talk about edge-detectors, columns, 40 Hz oscillations, and such-like, which do not go far towards explaining its spectacular evolutionary success. The puzzle deepens when we appreciate that most forms of pattern selectivity observed in the cortex are not especially complex; they can be found in the retinal ganglion cells of lower mammals or insects.

I'm going to suggest that the visual cortex identifies *objects* on the basis of the statistical regularities they cause in sensory messages, and that it provides a representation of *objects* which aids associative learning. The theoretical requirements and the physiology fit together quite well for the early steps in this process, but many questions about the later steps are still unanswered.

Innate Releasers, Bug-detectors and Object-Identification

About forty years ago Konrad Lorenz (1961) and Niko Tinbergen (1953) published semi-popular books describing *innate releasing mechanisms* that triggered off stereotyped and biologically important

responses that they called *fixed action patterns*. For instance, the chick of a herring gull opens its mouth and clearly expects to receive some food when it sees the mother gull's head or a good model of it. The herring gull's beak has a prominent red spot on it, and it turns out that this is important: without the red spot, the chick won't gape, or gapes much less often. It is therefore thought that herring gull chicks have, at hatching, a fully formed "red spot detector" in their visual system, and the mother gull develops, equally under genetic control, a spot on its beak when it matures. The sight of this spot "releases" the gaping response.

The chick will gape at a quite inappropriate object, provided it has enough "red-dottiness", though the object has to have other characteristics as well, such as "pointiness". The fact that chicks are easily fooled suggests that the neurological mechanism for detecting such an innate releasing factor is not very complicated, and this was what lay behind the idea that the frog's retina has bug-detectors (Barlow, 1953; Lettvin, Maturana, McCulloch, and Pitts 1959). This could imply that much of the psychology of perception is determined by quite simple physiological mechanisms at peripheral points in the sensory pathways – an idea that psychologists never much liked because it suggested that the physiology of a few milligrams of peripheral tissue accounted for a large chunk of their subject.

I now think they were right: perception is not about *bug-detector-like* processes, but about *object-identification*, since our learned responses are based on objects (Snyder and Barlow, 1988). Cats and primates do not have units at early levels with highly selective trigger features because they defer processing to the cerebral cortex which provides a much more comprehensive and versatile system of object-identification.

But isn't a bug just a particular example of an object? Obviously it is, but the purpose of our own perception differs from that of the sensory mechanisms for triggering fixed action patterns, and consequently a very different form of representation is required. We need to know about *all* the objects in a scene, since we need to be able to form an association with any object that happens to be present. In contrast bug-detector-like mechanisms provide a representation adequate to answer a few genetically predetermined questions such as "Is this an edible object (or a potential mate, or a dangerous enemy) I see before me?"

The reason it is important to represent objects is that these are the appropriate units with which to form associations. If one formed associations with the simpler elements of sensation such as light, colour, movement, edges and so forth by which the objects are identified, one would be unable to generalize appropriately. If you have been stung by a wasp, you should try to avoid the *wasp*, not just

its sound or stripes or feel; the sound will then alert you even if you cannot see the stripes, and you will fear the feel even if the creature is invisible and silent. But for this to be the case you must recognize that the characteristic appearance, sound, and feel of a wasp belong together, and the same is true for all other objects that we can potentially form associations with. Reinforcement may be a powerful aid to grouping elements together to form the representation of an object, and it may refine the grouping and thereby change the generalization field. We may also be taught objects by example - as when told "That is a horse" - but for the moment these aspects will be ignored and we'll look at the means by which objects can be identified from the evidence they themselves provide.

Table 1 defines the computational goals of bug-detecting and object-identification. Objects are claimed to cause constellations of sensory signals that frequently occur in association with each other, that can have names attached, and that not only form the elements of our cognitive world, but are also the appropriate elements for associative learning. Fixed action patterns are genetically determined, so their trigger features can be as well. In contrast, the objects you learn about are not genetically determined (or anyway not entirely), so sensory messages must be classified and categorized in a flexible way that depends upon the actual sensory experience of the animal. The distinction is probably not as clear-cut as suggested, but many other differences follow directly from the different role the two processes play in the lives of the animals concerned, and there is little doubt that object identification is a very much harder task.

Table 1. Computational Goals of Bug-detector-like Mechanisms for Innate Releasers and Object-identification for Perception

BUG-DETECTING	OBJECT-IDENTIFICATION
Detecting one of a small range of genetically pre-specified releasers for fixed action patterns	Classifying sensory messages into categories which should be learned about separately, but within which generalization is appropriate.

Accepted neurophysiologal interpretations do not take one far towards understanding object identification and recognition, so I shall introduce some theoretical ideas that seem to me promising, and try to show how these illuminate physiological facts and give added insight into perceptual mechanisms. This is far from being a complete and fully worked out theory, but a framework for one's thoughts is seriously needed now that we are being confronted with a maze of

new facts about the organization of the visual areas, and without a theory the opportunities for informative experiments will be missed. These ideas are:-

* That the hierarchical organization of the visual system must have nodes that combine information in two logically different ways.

* That the statistical properties of the sensory messages themselves provide the initial basis for classification, prior to reinforcement or instruction.

* That non-topographic mapping creates neural images for collecting together relevant evidence and excluding irrelevant evidence – the most difficult step in any pattern recognition problem.

Selective and Generalizing Nodes in a Hierarchy

The hierarchical structure of the visual system can be recognized at three levels. First, the nerve cells themselves can hardly be arranged in any other way, for a single cell cannot make connections directly with all the sensory messages evoked by an object; second the multitude of visual areas can, as Maunsell and Van Essen (1983), Zeki and Shipp (1988) and Felleman and Van Essen (1991) have shown, be arranged in a hierarchy on the basis of the different pattern of forward and backward connections; third, the actual nature of the object recognition task suggests the need for a hierarchical structure, for the whole object must be recognized from its parts, and these parts must be identified from more primitive image features. But I don't think it has been very widely recognized that the nodes or branch points in such a hierarchy have two very different operations to perform, namely a selective one that gives specificity to the output of the node, and a generalizing one that enables the output to represent many alternative inputs.

The *selective* operation leads to an output that is active in response to a smaller class of all input patterns than any of the individual input lines. The operation corresponds logically to AND or AND-NOT, and such nodes, arranged in a hierarchy, would generally lead to a system in which the higher neurons became more and more specific in their response requirements and consequently fired less and less often. The combinatorial explosion tells us that a vast number of such high level units would be needed to give reasonably complete coverage of all inputs likely to be encountered, and in general one has the difficulties associated with "grandmother cells" or "yellow Volkswagen neurons".

The second, *generalizing*, type of operation leads to an output that responds to a larger class of input patterns than any of the individual

input lines, and in the extreme responds to any one or more of them. This is the type of operation that performs pattern completion or generalisation, and such nodes would counter the trend of decreasing frequency of firing set by the selective nodes. A hierarchy that contained both these two types of node would have the attractive property of generating elements at higher levels with both selective and generalizing attributes, which of course is what we need.

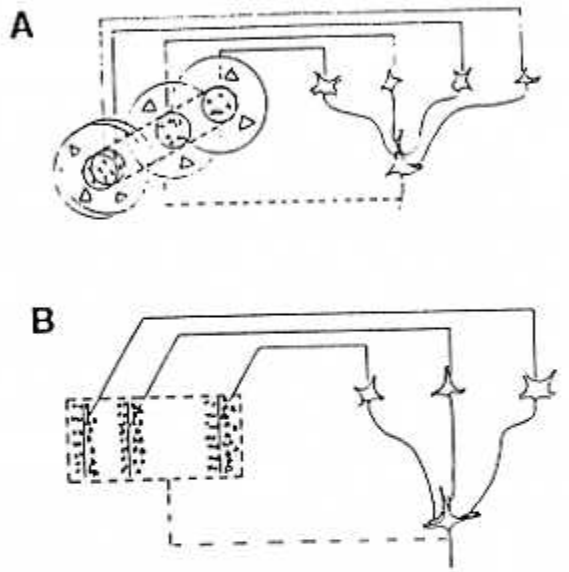

Figure 1. Hubel and Wiesel's suggested circuits for (A) simple and (B) complex cells in the cat striate cortex. According to the distinction made here, A exemplifies a selective node and B a generalizing one (from Hubel and Wiesel, 1962).

The first physiological evidence for a hierarchical organization in visual cortex came from the results on simple and complex cells obtained by (Hubel and Wiesel, 1962) and illustrated in Figure 1. The first stage was thought to produce the orientational selectivity of the simple cells by summing from circularly symmetric LGN afferents. This must be a selective node, for the simple cells are quite difficult to activate and certainly respond to a smaller class of inputs than the incoming LGN fibres, but Hubel and Wiesel, perhaps misleadingly, described them as simply summing the influences of the "on" and "off" zones of their receptive fields. The second stage then combined many of the first-type outputs, and was thus a generalizing node that generalized over a limited range of positions.

This model has been very much in people's minds ever since it was proposed, and for many of us it seemed to open the door a chink to give a glimpse of how the brain organized its work. But 30 years later, one can see some defects. First, it is now clear that complex cells receive direct inputs from the geniculate and may belong to a different stream of processing from the simple cells, so these two types of cell may not correspond to two levels of a hierarchy. And as suggested above, the operations at the two levels were not conceptually separated as clearly as they might have been.

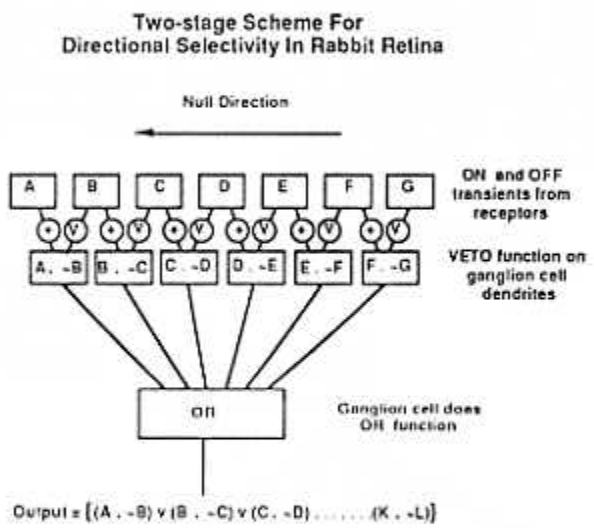

Figure 2. Scheme for directional selectivity in retinal ganglion cells of the rabbit. The connections between the two layers are excitation (+) or inhibition (V) that vetos succeeding excitation. Note also that the ON and OFF transients from the receptors seem to require parallel and independent pathways. The first step is selective and is now thought to occur at the level of the dendrites of the ganglion cells. The second stage generalizes over a limited range of positions and occurs by the ganglion cell summating inputs from the dendritic branches.

Directional selectivity in the rabbit retina provides another model, and this is illustrated in Figure 2. This distinguishes between the restrictive or selective operation that gives the neuron its pattern selectivity, and the second generalizing function. For the experimental reasons we gave (Barlow and Levick 1965) the restrictive function was thought to be logical VETO or AND-NOT rather than AND, while the second generalizing operation was logical OR. We assumed (Barlow and Levick 1965) two separate cells for the two types of operation, but this was proved wrong (Dowling, 1970; Dubin, 1970; Werblin, 1970)

and it appears probable that the first operation is done in subcompartments of the retinal ganglion cell dendrites, and the second operation by summation at the ganglion cell's cell body. There are reasons for believing that the subcompartments of the dendritic field of cortical neurons are not well enough separated electrically to allow them to perform separate logical functions (Dehay, Douglas, Martin and Nelson, 1991), but there may be other interactions inside the dendrites that would make the dendrites selective while the cell body generalized.

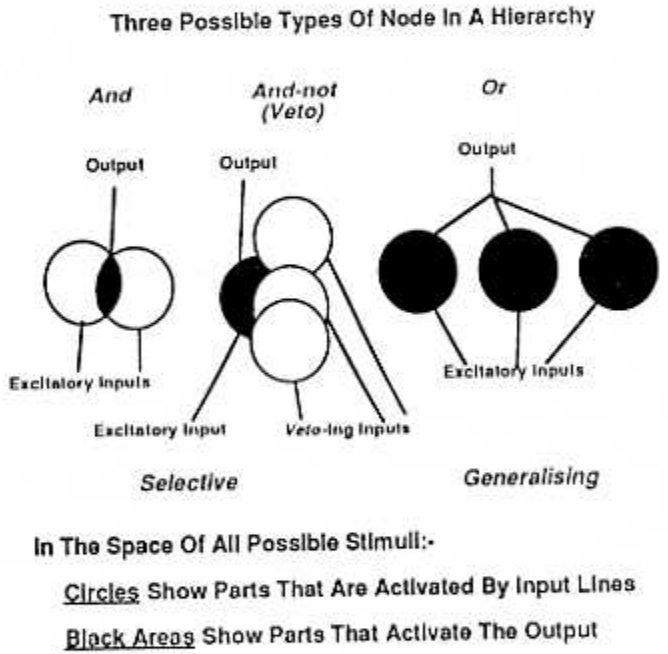

Figure 3. Selective nodes respond to a smaller class of all inputs than any of their input channels; this can be brought about by an AND or AND-NOT (veto) combination. Generalizing nodes respond to a larger class and can use an OR combination.

At all events two very different rules of combination are required and these are illustrated diagrammatically in Figure 3. The first, with a rule of combination equivalent to logical AND, or alternatively AND-NOT, restricts the response of the higher level element to some subset of the patterns to which the lower level elements it receives from respond; this is what gives a neuron selectivity. The second, roughly equivalent to OR, allows the higher level element to respond to a larger range of patterns than any of the lower ones and enables it to recognize a

pattern when all of it is not there - pattern completion or generalization.

In neurons, approximations to these logical functions can be brought about by a threshold-type non-linearity in an activation function, which can lead to an AND function if there are only two inputs but will lead to more complicated functions if there are more. The VETO operation can be mediated by shunting inhibition (Koch, Poggio and Torre, 1986) or presynaptic inhibition (Dowling, 1970), while the generalizing OR operation can be produced by a saturating type non-linearity in the activation function. However, these are only some of the possible ways of approximating these logical operations, and non-linear interactions of intra-cellular transmitters and modulators offer many other possibilities.

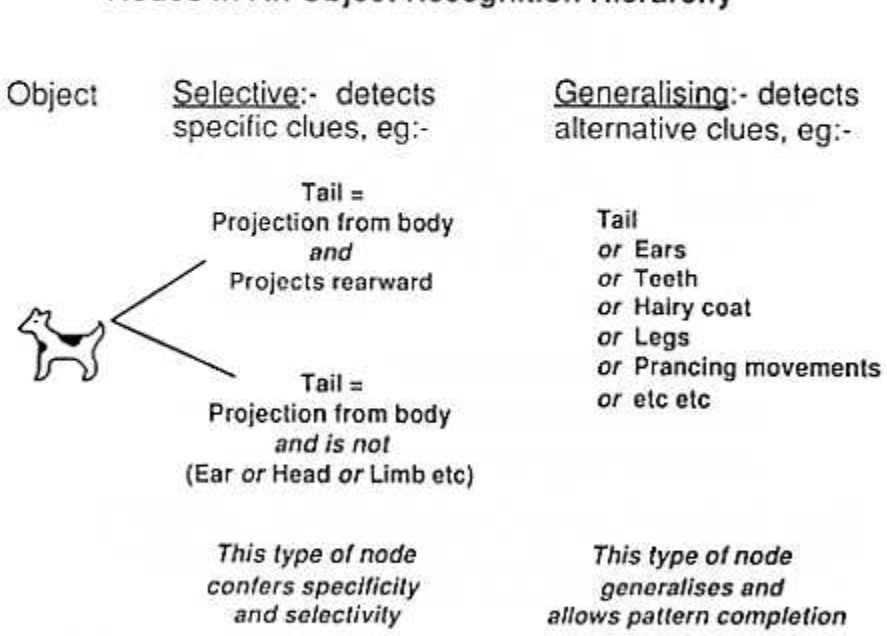

Figure 4. The role of selective and generalizing nodes in recognizing a dog from its image. Any of a number of alternative clues such as tail, ears, or prancing movements can be used, but the individual clues require a selective operation. It is clear that other criteria than those shown are required to distinguish a dog from a horse, for example.

The caricature of Figure 4 shows the part that might be played by these two operations in recognizing a dog. "Tail" is an example of a specific clue to the higher level element "dog", and the diagnostic requirement for "tail" is supposed to be that it is both a projection from

the body, and that it projects rearward. Alternatively, something would qualify as "tail" if it was a projection from the body and is not a head or an ear or limb. This veto method has other merits I'll return to.

The second, OR, type of combination is appropriate when any of a number of alternative clues can be accepted as establishing the presence of an object. This is caricatured in Figure 4 by the list of alternatives at the right. This is the type of combination where pattern completion is appropriate, and it leads to generalization of learned associations that would be functionally beneficial.

Those familiar with Fukushima's cognitron and neocognitron (Fukushima, 1975; Fukushima and Miyake, 1982) will recognize the similarity with the organization of his recognition system, which has the two logically different operations performed in alternating layers and was inspired by the neurophysiological model of simple and complex cells (Hubel and Wiesel, 1962).

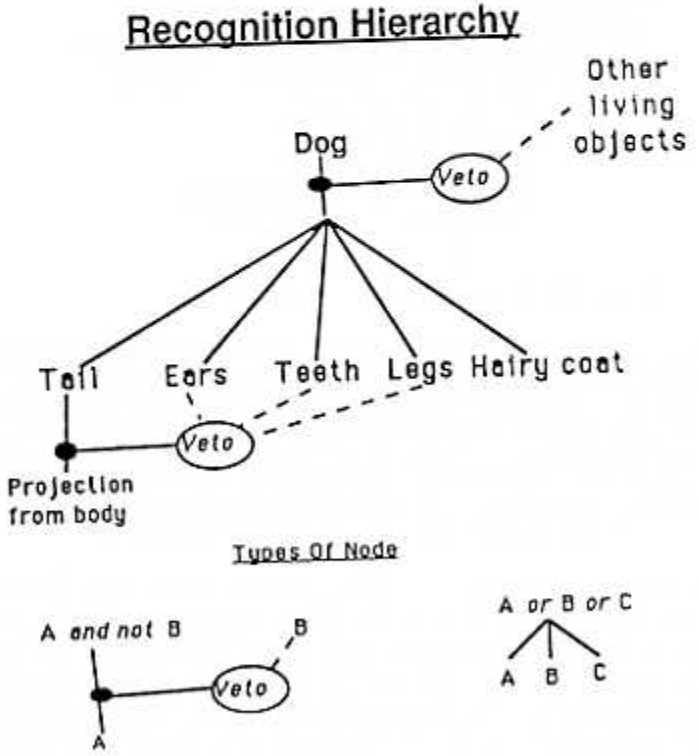

Figure 5. A recognition hierarchy may work better using the veto combination because it is robust and can be easily modified.

Figure 5 shows a hierarchy using AND-NOT as the restrictive function instead of AND. As well as being more robust than an AND brought about by a threshold type non-linear activation function, this type of restrictive function should be rather easily modifiable by adding to the list of alternatives in the part following AND-NOT. Furthermore, these alternatives should be already available, perhaps from higher levels, in a parallel hierarchical system where other elements are likely to be concerned with heads, ears, and such-like. Though there are important differences in these idealized models, the AND-NOT part might correspond neurally to the anti-Hebbian, decorrelating part of the element described by Földiák (1990), or the mechanism might resemble competitive inhibition (Rumelhart and Zipser, 1985).

Can one point to evidence for these two types of node at higher levels in the visual system? Certainly a neuron that responds selectively to faces (Gross, Rocha-Miranda and Bender, 1972; Perrett, Rolls and Caan, 1982) must have benefitted from both types of operation, particularly if it is selective for a single individual's face (Perrett, Smith, Potter, Mistlin, Head, Milner and Jeeves, 1984), and it has been suggested that there are cells in STS that generalise by summating the outputs of other, more selective, cells in their neighbourhood (Perrett, Harries, Mistlin, and Chitty, 1990). In extra-striate cortex one has evidence of cells that have types of selectivity not found at earlier levels both with respect to colour (Zeki, 1980; Desimone, Schein, Moran, and Ungerleider, 1985) and motion (Allman, Meizin, and McGuiness, 1985; Movshon, Adelson, Gizzi and Newsome, 1985; Newsome, Britten and Movshon, 1989). It would be interesting if one could link the selective and generalizing operations with intra- and inter-area operations, but it is more probable that the collection of information in an area facilitates both types of operation by cells in that area.

I think a greater awareness of the conceptual difference between the two types of operation might lead to experiments that would clarify what is happening. The distinction is also important in considering how the statistics of the sensory input can be used to help object identification, but this must wait until other aspects of the statistics have been considered.

Statistics of Sensory Messages as the Basis for Object Identification

The basic idea is that objects cause certain types of statistical regularity or redundancy in visual messages, and it is the task of sensory and perceptual mechanisms to detect these regularities and use them to help categorize the messages. The argument is that this

categorization will correspond to object identification, and that the processes required to do this will shed light on the physiological mechanisms in visual centres. This approach does not deny the importance of reinforcement and teaching by example, but the expectation is that categorization of the input using internal evidence alone will make these other factors much more effective.

The message that redundancy or regularity in sensory messages is important can certainly be traced back to Attneave (1954) and possibly to Mach (1886) and Helmholtz (1925). I've suggested that some physiological processes can be regarded as redundancy reducing codes (Barlow, 1961), or more recently as detectors of "suspicious coincidences" (Barlow, 1985), or decorrrelating mechanisms (Barlow and Földiák, 1989; Barlow, 1990b). In computer vision Witkin and Tenenbaum (1983) introduced the idea of non-accidental occurrences as pointers to important features of the image connected with objects, and this is the main idea in Lowe's (1985) scheme for perceptual organization. I've argued recently (Barlow, 1990a; 1991) that this gives a very different view of the task of perception from that of 3-D reconstruction of the scene, as advanced by Marr (1982) and his followers.

There is no need to dwell on the first steps of signal processing in which the dynamic range of the receptor mechanisms is adjusted to suit the amplitude, and perhaps the range of contrasts, of the incoming signals that have been received over the recent past. These adjustments are certainly made in response to statistical properties, and are obviously important for a sensory system to function properly, but they have little to do with the properties of objects; instead they are concerned with compensating for changes in the illumination level, viewing conditions, and distance of the objects being looked at, etc. I think lateral inhibition would also be widely accepted as a mechanism to compensate for the universal tendency of visual images to have strong autocorrelations. Again, the plausible functional interpretation is that it reduces the dynamic range of the signals and thereby adapts them to the limited dynamic range of neurons. However, the autocorrelation clearly results from the image being composed of surface patches of relatively uniform luminance, and this may be the simplest example of an important property imposed on the input by objects.

Table 2 Processing Driven by Statistics of the Input

Statistic	Main determining factor	Processing step	Where done	Refs
Mean	Illuminant	Light adaptation	Retina	1,2
Range	Conditions Distance	Contrast gain control	V1	1,3 4
Spatio-temporal correlations	Motion of eye or object	Motion selectivity	V1	5,6 7
Spatial autocorrelation	Object surfaces	Lateral inhibition	Every-where	8,9
Translational symmetry	Object edges	Orientational selectivity	V1	5
Simultaneous correlations	Object attributes	Decorrelation	? V1	10
Clustering of:- motion vectors colour texture disparity	Objects' spatial coherence and uniformity of material	Project information to new area	? Extra-striate areas	11

References

1) Werblin (1973) 2) Barlow (1972) 3) Barlow and Levick (1969) 4) Ohzawa, Sclar and Freeman (1985) 5) Hubel and Wiesel (1959; 1962) 6) Barlow (1960) 7) Barlow, Hill and Levick (1964) 8) Barlow (1950; 1953) 9) Kuffler (1953) 10) Barlow and Földiák (1989); Barlow (1990b) 11) Barlow (1990a; 1991)

Table 2 summarizes the relations one can trace between the statistical properties of the sensory input and mechanisms in the visual system to deal with them. The interesting, and perhaps contentious,

ones from the present point of view are those in the lower part of the table, where statistical characteristics are listed that are held to be caused by objects. Objects come in such a wide variety of shapes, sizes, origins and purposes that one cannot say, of any particular small part of an image by itself, "That signifies the presence of an object." If objects all tended, say, to be red, then one could do this, but there is no such simple property they all share. But they do tend to share more complex properties: they show clustering of many different attributes, such as direction of motion, disparity, or particular colours or textures, and they also tend to be bounded by edges. As a result clusters of any attribute, and borders, give an indication of the presence of an object.

The Gestalt movement should be credited with discovering the properties of objects that make them segregate from their backgrounds perceptually. Table 3 lists their best known principles and suggests that they correspond to properties that objects impose on the visual input. For example, objects are usually rigid and cohesive, so that their different parts are often close together in space, and stay close together even when they move. Thus neighbouring points in an image are likely to be related to the same object, and so are regions showing the same direction and velocity of movement, or the same disparity. Brunswik (1956) has given some evidence that "proximity" does in fact pick out manipulable objects in a scene. In this light Gestalt laws of segregation are seen as rules the visual system follows in order to group together the parts of the image that result from a single object. The Gestalt laws about boundaries can also be related to the fact that objects often have well-defined borders or edges which lie in front or behind other borders and edges, and completely enclose the object.

Table 3. Gestalt Grouping Principles as Cues to Objects

Gestalt principle	Object property
Proximity	Spatial coherence
Similarity	Often made of one material
Continuation	Have edges
Closure	Have continuous boundary
Symmetry	Often symmetric
Familiarity	Cause conjunctions that would not occur by chance

Thus, although they were not formulated as such, the Gestalt laws seem to represent the perceptual system's ways of identifying the statistical properties imposed on sensory messages by objects. When doing a jig-saw puzzle one sorts the pieces according to colour, texture,