The compromise between seeing spatial layout and making visual discriminations

G. Adrian Horridge

Insect vision, which is simple compared to human vision, provides many clues for the design of artificial visual-processing mechanisms. Such artificial mechanisms based on insect vision are likely to have applications in artificial-intelligence systems.

Our fascination with the marvel of human vision and our lack of information about the real mechanisms are two major obstacles to be overcome before the mechanisms of simple low-level visual discriminations can be copied into technology. The first obstacle is avoided by considering the vision available in lower animals and the second by analysis of behaviour, electrophysiology and microcircuitry of their natural visual mechanisms where we can. The appropriate level of complexity, minimal for making a visual system for manoeuvring in the three-dimensional world, is that of a generalized insect, which conveniently indicates one practical compromise between complexity, performance and ease of understanding. Unlike human vision, which relies extensively on a huge memory and rapid learning, insect visual mechanisms are relatively simple in that they certainly see the outside world very well but there is little analysis of two-dimensional pattern, and objects are not categorized into classes.

A camera, or a video camera attached to a television screen, can only transfer the image from one place to another. Vision is quite different in that it involves discrimination of certain features of the image followed by decision and action. A visual system must therefore be designed with reference to what features of the outside world are adequate cues, and what actions will follow. The design must be directed at abstracting some cues from the image, and for economy it need do no more. A good example is the automatically opening door (Figure 1, c) which detects the arrival of a shadow and opens the door for a short time. For low-level vision, there is no need to copy human vision or reconstruct the spatial relationships of every detail in the outside world.

We have little vision technology at this intermediate level. Examples are surveillance equipment that set off an alarm if an intruder enters the field of view in an otherwise stationary scene, the guidance of a heat-seeking missile by infrared radiation, the automatic landing equipment on aircraft that use radar waves, and the range-finder on some automatic cameras that uses sonar (Figure 1). All these simple kinds of eyes share the common features of eyes, namely sampling in angular space and use of some part of the total visual information.

The eye operates by having many receptors that look in different directions at the same time (Figure 2), so that information is transmitted into the eye in many channels in parallel. The video camera works quite differently, in that it scans the image one pixel at a time and transfers the raster pattern into a long time series, which can regenerate the picture on a TV screen by replay of the raster to bring together the picture. We
Insect vision

We would like to make an artificial visual processing system that reads the local spatial and temporal features into a computer code so that any desired fraction of the features can subsequently be used as cues, avoiding the combinatorial labour of reconstructing the visual world as a map of 'objects'. One way out of this difficulty is by simple steps towards a limited goal, copying low-level natural vision, such as that in insects, in which substantial useful analysis of visual mechanisms in action is possible.

Insects see nothing when their eye stands still in a stationary world. Their visual neurons are inactive unless there is relative movement. Insects in motion see very well, and they can measure the range and direction of nearby contrasting edges of objects. As insects do not have binocular vision (with a few exceptions), lens accommodation, disparity or eye-convergence mechanisms, they can measure range only by their own motion (Figure 3). As the eye moves, contrasts that are nearer appear to move further and faster than objects that are further away. Many insects deliberately scan in order to get information about range from the predetermined motion of the eye. We have shown that bees can measure the velocity and therefore the range of moving contrasts relative to the eye, irrespective of the pattern, and that bees see another effect of their own motion, namely the passage of near objects in front of contrasts that are far away, causing a local discontinuity in the flow field. Sometimes called parallax, perception of this motion across the background is a sure way of becoming aware of an object as both near and distinct.
without being able to categorize or say any more about its form, texture, shading or identity.

Vision, at this level of complexity, thus involves detecting contrasts in motion and measuring some simple features of them, such as the direction of motion, polarity of edge, and (hopefully) their velocity, together perhaps with other features such as the gradient of an edge and the magnitude of the contrast, which would enable discrimination of one contrast from another and learning its features for future recognition. Bees, which learn to come to targets to collect honey, use their eyes in motion for simple tasks of this kind, but their ability to discriminate patterns is rather limited.

The next problem presented by the visual world is the enormous range of light intensity in which vision is performed, from bright sunlight at $10^{14}$ effective photons per cm$^2$ per sec down to moonlight at about $10^7$ effective photons per cm$^2$ per sec. This is a range of at least $10^7$. Natural visual systems cope with this wide intensity range in three ways: they adapt, i.e. they cut down receptor and neuron sensitivity at high average intensity; they sum over larger fields at low intensity; and they look at changes in intensity, ignoring absolute intensity. These mechanisms have the effect of shifting the working range of the second-order neurons to correspond with the best positive or negative signal.

At this level in the insect eye, just behind the retina, there is also a high synaptic gain, and the second-order neurons (Figure 4, d e) give a rapidly adapting response in one direction to a local increase in intensity and in the other direction to decrease in intensity. These signals are effectively the temporal derivatives of those at the receptors and they retain the full spatial resolution of the receptor array.

### Artificial visual-processing system

To copy this mechanism in our model, we take the responses of a horizontal line of say 100 pixels (Figure 5), then we subtract from each the new response when the eye has moved in a horizontal direction to a new position, and repeat this, to get successive rows of intensity differences as the row of pixels is moved along. We now have a measure of temporal contrast at each pixel as the line of pixels moves along the scene parallel to its own length. Copying the high gain, which rapidly reaches saturation in the natural system, we take a threshold positive or negative temporal contrast at an arbitrary low level of near-1%. This value is near the natural lower threshold for contrast in insects and also suits the noise level in our optoelectrical equipment. With the line of pixels plotted horizontally on a diagram, and the successive one-dimensional scans plotted downwards, we now have a spatiotemporal map of temporal contrast in three states, namely increasing contrast, decreasing contrast, and no observable change (Figure 5, b). Having only two states would make a visual system that is too simple for the required performance; having more than three states improves the discrimination of intensity changes, gradients and shading, but adds no new principles for understanding vision, while causing a large increase in the number of possible combinations in subsequent processing.

Such a 'primitive eye' consisting of one line of pixels can scan along its own linear direction and generate a spatiotemporal map of thresholded temporal contrast.

---

**Figure 3.** Range $(\text{nc})$ is inversely proportional to the angular velocity induced on the eye by (a) forward motion or (b) lateral scanning.

**Figure 4.** Summary of retina-lamina-medulla relations in the insect optic lobe. a, Light is absorbed in the rhabdom which acts as a coloured light guide in each ommatidium. b, Single ommatidium in transverse section, with six retinula cells. c, The microvilli, which contain the visual pigment. d, Optic lobe in longitudinal section showing the successive stages of neural processing. e, The responses to light in (1) the receptors and (2) second-order lamina cells which produce the postsynaptic temporal derivative.

**Figure 5.** Diagram of retinal lamina relations in the insect eye (a) Longitudinal section of the eye showing the relation between the lamina-retinal-microvilli (b) Longitudinal section of a single lamina cell showing the postsynaptic response to the retinal cell.
Figure 5. The processing of the spatiotemporal map by the templates. a. The sloping shaded areas represent contrasts seen at different ranges by an eye moving forward. b. The same with three thresholded states of temporal contrast as follows: i.e., increasing contrast; ii, decreasing contrast; blank, no temporal contrast. c. The same with the responses of six templates shown in Figure 7. For further explanation, see Figure 6.

(Figure 5, b). Vertical lines on this map represent contrasts that are changing in intensity but are very far away, and sloping lines represent contrasts that move as a result of the scanning, and the slope is a measure of range. At this stage we remember that the motion induced by rotation of the eye must later be eliminated by a special process.

Next, we consider all possible spatiotemporal interactions between inputs from each pair of adjacent visual axes at each pair of successive instants, to begin the search for correlations at the resolution of the retina. First we write out all possible spatial pairs of three states, for one instant, namely: (--), (↓), (↑), (→), (←), (↑↓), (↓↑), (↑→) and (→↑), where ↑ = increasing contrast, ↓ = decreasing contrast, and = no temporal contrast. For every spatially adjacent pair of inputs along the line of motion, one of these combinations must exist.

Then for successive instants we can make all possible 2 x 2 spatiotemporal templates from these nine, making 81 templates (Figure 6). One of these combinations must turn up at each instant at every visual axis along the line of motion, so all the resolved image is coded as spatiotemporal correlations. The fully determinate relation between template responses and the spatiotemporal map of moving contrasts is shown in Figure 7. A possible location for the templates is to have one set in each column of the optic medulla (Figure 4, d).

Parallel processing

A single template response alone is not particularly informative, and templates do not necessarily represent unique features of the outside world, but together they make up the whole picture. This dependence of templates upon each other mimics neurons which have evolved together and is an essential feature of a parallel-processing mechanism. Templates in parallel can be
much simpler than (say) zero-crossing edge detectors, which are individually more informative but require more computation and do not make use of component interdependence. Because each template gives a yes/no answer and all pixels are in parallel, the mechanism has no numerical calculations and would operate much faster than the scanning system or the rate of collection of photons by the receptors, especially if everything up to this stage could be built in hardware.

It is clear that the resolution of the retina is fully used and that form or structure on a larger scale is represented by the groups and combinations of template responses in the spatiotemporal diagram (Figure 5,c). The polarity of edges, the direction of their motion, and the amounts of temporal and spatial contrast are immediately obvious from the templates that respond. Insects see these features very well, but they also measure range\(^8\) and discriminate simple form\(^3\). But what is the artificial processing mechanism designed to see?

Always ready to cut down the number of template types, we now see what clues the visual scene holds to make possible the discriminations. They we need employ only templates for the task in hand. For example, at a very simple level, there are two ways to measure the velocity or range of a contrast. By the first method\(^7\), we construct elongated fields for directional templates that lie along different slopes in the spatiotemporal map. These templates are like neurons looking for particular velocities in the patterns of any primary templates, and illustrate that velocity or range is just another aspect of spatiotemporal form as seen by this mechanism.

**Taking ratios**

The second method of measuring velocity is to take the ratios of the total number of symbol pairs indicating temporal contrast, e.g. \( (\downarrow/\uparrow) = (\downarrow/\downarrow) \) and divide by the total number of pairs indicating spatial contrast, \( (\uparrow) \) and \( (\downarrow/\uparrow) \) in the same medium-sized spatiotemporal region. This calculation is similar to the familiar gradient algorithm for angular velocity at an artificial seeing system\(^6\), in which \( \delta I/\delta t \), where \( I \) is intensity, \( t \) time, and \( \theta \) the angle on the eye. This operation suggests that taking ratios is a way of measuring many other qualities of the features independently of their number or intensity (Figure 8).

Bees and some other animals can discriminate a wide range of colours although they possess only three or four visual pigments with broad overlapping absorption spectra in separate photoreceptor types (Figure 8,a). The photoreceptors count photons captured by each receptor type and the mechanism is well represented by a colour triangle (Figure 8,b) or tetrahedron, each corner of which has a receptor type. Each region within the diagram represents a different ratio of the colour mixture (with white somewhere in the middle, for man). We divide the colour space into regions called colours and give them separate output lines (Figure 8,b). Because ratios are involved, colours can be discriminated or recognized independent of intensity, shape or frequency of presentation.

We can, instead of colours, our thresholded symbols or spatiotemporal templates (Figure 8,c) at the corners of a triangle, tetrahedron (Figure 8,d) or multidimensional hedron and identify regions in form space by the same mechanism that we know can be done in colour vision, whatever the synaptic mechanism may be in real nervous systems. By taking ratios, therefore, we can measure many attributes of regions of the visual world without actually seeing form in spatial layout, i.e. without a spatial representation at the full motion resolution of the retina. The spatiotemporal resolution of these discriminations depends on the size of the fields over which the ratios are counted. Our knowledge of insect visual behaviour suggests that this could be all they discriminate. Such a system would be sufficient for classification of contrast groupings, and could be copied into artificial seeing systems. This new technology would be useful for mobile vehicles, aids for the blind, and detecting errors in quality control and dirt or abnormal textures in manufacturing, but it would have poor resolution for 'reconstructing the picture' for categorizing objects such as dogs, chairs or written words.
In this mechanism based on ratios of numbers of responses of particular templates with overlapping sensitivities, the individual responses are of little account. The field over which the ratio is taken cannot cover too many pixels or else the spatial resolution is lost, nor too few or the sensitivity is sacrificed. Similarly the time over which the ratio is measured or temporal or spatial frequency. In fact, in the bee, this accurate than those when simpler patterns are presented'.

This compromise in the spatiotemporal field size cannot be too long, or temporal resolution and lost, nor too few or the sensitivity is sacrificed. Moving objects are 'blurred', nor too short or sensitivity is raised to the power (number of members that combine). A $2 \times 3 \times 3$ two-dimensional template with three states (Figure 10,b) has $3^{18}$ possible combinations to be

Evolution of early visual processing

Natural visual systems must have evolved many times in different animal groups, implying that the mechanisms of evolution of visual processing must be relatively straightforward. Let us consider how this process occurs. We remember that the spatially adjacent pairs of inputs and the temporally successive states at each pixel are the minimum combinations that retain the retina's full resolution for contrast in space and in time. Subsequent logical processing mechanisms can operate on these 81 combinations which contain all of the available information after thresholding to three states. In practice, however, only about half of the templates turn out to respond to natural moving scenes.

As growing neurons come together in the peripheral layers of a visual system, they must form combinations of excitatory and inhibitory impulses with convergence of neighbours and some persistence of transmitters. Therefore one or another of the 81 possible basic templates is likely to be formed, at first accidentally. Some of these convergences would be selected on Darwinian principles, and, if growth conditions are uniform, a regularly repeated array would form. Even a random convergence into adjacent pairs would produce some templates, and with only 81 possibilities there is better than 1% chance that any particular desired template would turn up. We can see that evolution towards a collection of useful templates can proceed relatively rapidly if the templates are simple, and the performance can be progressively improved by the addition of further templates. The resulting model circuit (Figure 9) looks something like a real nervous system, operating with ensembles of units in parallel.

This early evolution is much less likely with more complicated templates because the number of possible combinations rises rapidly as the number of states raised to the power (number of members that combine). A $2 \times 3 \times 3$ two-dimensional template with three states (Figure 10,b) has $3^{18}$ possible combinations to be

![Figure 9. Ensemble processing by an array in the insect visual system. The inputs 1, 2, 3 are the lamina monopolar cells (LMC) carrying temporal contrast at the resolution of the retina. In each column these diverge to a number of types of intrinsic neurons $a$, $b$, $c$, etc., which abstract aspects of temporal contrast in various spatio-temporal combinations (templates). A $2 \times 2 \times 2$ template (a) is insufficient for discrimination of corner from an edge, and $2 \times 3 \times 3$ templates (b and c) offer too many possible combinations.](image)

![Figure 10. Constructing templates for two dimensions in space. Corners that move in predictable ways generate patterns of intensity change. A $2 \times 2 \times 2$ template (a) is insufficient for discrimination of corner from an edge, and $2 \times 3 \times 3$ templates (b and c) offer too many possible combinations.](image)
sorted out by natural selection, and a large number of combinations slows up the sorting process during normal functioning. With this number, both evolution of templates and visual processing would be impossible.

A two-dimensional system

The criticism that the model presented operates only in one dimension can be met by two arguments. First, even the simplest two-dimensional template with three states (Figure 10.a) on each pixel has $3^a$ combinations, which seems a lot from which to evolve specificity of any one combination. Such templates are also rather unspecific when applied to natural scenes unless they are used for only certain directions of motion across the scene, in which case they may as well be one-dimensional. Secondly, vision depends on motion, and at any one time and location on the retina the motion must be in one dimension, with components in the horizontal and vertical planes. Therefore vision can remain one-dimensional in the horizontal plane with the addition of some one-dimensional vision in the vertical plane so that the directional vectors of local motion can be found if the output requires it. This is what we find in insects, where vision appears to be one-dimensional in the horizontal and vertical planes, but the two axes appear not to be combined to form two-dimensional detailed pictures in space. Bees can discriminate the inclination of edges on a vertical target (relative to the direction of gravity), perhaps by a ratio, but the unit motion detectors in the two planes do not appear to be combined to form two-dimensional pictures in space. The known examples in which insects discriminate between forms in two dimensions can apparently all be explained by one-dimensional vision along the line of motion in vertical and horizontal planes separately. The problem is not how pattern analysis in general can be performed, but how simple the minimum cues with which bees perform the observed discriminations can be.

To end this discussion of evolution and dimensionality, we might conclude that vision evolved as one-dimensional because the templates had to be simple, and therefore the eye had to nod, scan or move during locomotion. We could equally well conclude that locomotion in the horizontal plane came first and therefore the eye evolved one-dimensional vision in the horizontal plane, adding the vertical later. Either way, a complex visual processing system can be evolved in small steps behind an eye that has already progressively evolved many small photoreceptors in the pursuit of resolution of spatial contrast. The templates depend upon group action because they evolve together, and the ratio mechanisms allow discriminations irrespective of intensity or frequency of presentation.

Conclusion

Basically, receptors see changes in intensity; second-order neurons see positive or negative temporal contrasts. In the front-end processing mechanism, we need lateral and temporal correlations, so in the model we specify all possible spatial and temporal thresholded contrast pairs and count how many of these templates actually respond in local areas of a moving scene. Vision always involves a convolution of the fields of the seeing units (neurons in real systems) with the visual scene, not of the pattern of contrasts with itself. The quantized thresholded templates reduce this convolution to a yes/no response in many parallel labelled pathways. Finally, ensembles of units, each of which means little in isolation, depend upon each other's activity at every level right through the system to the final task-oriented response.

Many aspects of science are touched upon in the study of vision, and conversely a good model of vision is one that draws upon many different principles and explains the known properties of visual systems in a simple, comprehensive way. Vision has evolved separately in lower animals such as marine worms, insects, octopus, spiders and fish. Mechanistic models for these varied visual systems are still lacking for want of experimental data on how they work. The existence of many kinds of vision, adapted for different visual worlds and different visual behaviours, suggests that we shall find new mechanisms of natural vision that might be useful for 'simple' robots that perform visual tasks, even if we cannot yet copy man's visual system, which categorizes objects before it presents them to our consciousness.

Vision in man can be described as a continual hallucination derived from memory which is 'topped up' by the wandering fovea. By this means we can explain visions in dreams, visual illusions, filling in at the blind spot, failures to notice errors, size constancy, colour constancy, sharpness of edges in motion, and categorization, preceding consciousness, however, as in rapid reading. Even with existing supercomputers it is at present not possible to construct a general-purpose artificial seeing system with anything like this marvellous performance.

Acknowledgements. I would like to thank those who have worked on these and related questions in Canberra over the past few years while these ideas have been fermenting. The comprehensive theory has depended on research on the visual behaviour of the bee, much of it done by Myrcam Lehrer, on the electrophysiology of insect optic lobes by Ljerka Marcelja over many years, and on computer modelling done recently by Peter Sobey. Especially, however, I would like to thank my colleague, Mandjyam Srinivasan, for his unfailing support, and for the ideas and critical comments that emerge from our discussions. Numerous drafts of the article have been typed by Margaret Donohue, the illustrations have been prepared by Jenny

692

CURRENT SCIENCE, VOL. 60, NO. 12, 25 JUNE 1991
Vaughan. The work on object vision in insects is supported by a grant from the Fujitsu Company of Japan, by funds from the Centre for Visual Sciences and the Centre for Information Science of ANU.
