Computational theories of vision

Andrew Glennerster[*]

July 26, 2002

1 Introduction

Neuroscientists study brain mechanisms at many different levels, from molecular to psychological. Despite significant progress on many of these levels, there is a disappointing lack of coherence in neuroscience research: we have yet to find an overarching theoretical framework within which to understand what the brain does.

Vision is one of the most intensively studied areas of brain function. Yet, even in this field, there are wide disagreements about the goals of cortical processing. In the second half of the twentieth century, there were two important attempts to provide a theoretical framework for understanding vision, by David Marr and James Gibson. There are now grounds for optimism that these two broad approaches can be brought together to provide a biologically plausible and yet computationally tractable framework for understanding and imitating human vision.

This is not a review of computational vision, which refers either to computer vision or, in other contexts, to any area of visual neuroscience that requires mathematical modelling. Instead, the focus is on attempts to describe vision at the level Marr called 'computational theory'.

2 Marr

Marr (1982) emphasised that vision was nothing more than an information-processing task. Any such task, he argued, could be described on three levels: (i) computational theory, (ii) specific algorithms and (iii) physical implementation. The three levels correspond roughly to (i) defining the problem and setting out how, in principle, it can be solved, (ii) designing a detailed simulation of the process and (iii) building a working system that will carry it out. Box 1 illustrates these levels using the example of binocular stereopsis. The important point is that the levels can be considered independently. As a result, it ought to be possible to mimic the algorithms underlying biological vision in robots: the only difference would be in how they were implemented physically. This concept of independent levels of explanation remains a mantra of vision research.

Box 1 about here

Marr made suggestions about algorithms and computational theory for many components of vision, such as stereopsis or colour perception. He also attempted to set out a computational theory for vision as a whole, although these ideas have survived less well. He suggested that visual processing passed through a series of stages, each stage corresponding to a different

[*] University Laboratory of Physiology, Parks Road, Oxford, OX1 3PT; ag@physiol.ox.ac.uk

representation. The stages were (i) the retinal image, (ii) a 'primal sketch' (where lines, edges and junctions are made explicit), (iii) a '$2\frac{1}{2}$-D sketch' (which includes information about surface slant and depth), (iv) a '3-D model representation' of objects (which are built up hierarchically from simple primitives such as cylinders for arms or fingers) and (v) a 'space frame centred on the observer' where the 3-D models reside. One problem with this account is that information needs to be passed continually from one coordinate frame to another. There is increasing interest in models that avoid coordinate transformations of this kind, instead using information stored in retinal coordinates for tasks such as object recognition or navigation (e.g. Mallot, 2000).

3 Gibson

Like Marr, Gibson had a powerful influence on vision research in the last century. Marr himself wrote that 'In perception, perhaps the nearest anyone came to the level of computational theory was Gibson'. Gibson promoted an 'ecological' approach to studying vision by which he meant that vision should be understood first and foremost as a tool that enables animals to achieve the basic tasks required for life: avoid obstacles, identify food or predators, approach a goal, etc. (e.g. Gibson, 1979).

This viewpoint has gained increasing influence. An emphasis on survival of the organism (or ultimately of genes) is a more promising basis for a computational theory of vision than Marr's assertion that vision is 'knowing what is where by looking'. You can blink before you know what caused you to do so, and the ability to avoid a looming object almost certainly evolved before a more sophisticated visual recognition of 'what' things are. The ecological emphasis on action is quite different from the notion of an all-purpose representation in which each object has a label and a coordinate, as Marr envisaged. Where Gibson infuriated his contemporaries was in his musings about the mechanisms by which the brain might generate these 'ecological' behaviours (Ullman, 1980). Even when he avoided any mention of brain mechanisms and stuck purely to what Marr would call the algorithmic level, his proposals were often loose or unclear (see Box 2).

Box 2 about here

Despite the criticisms, there has been continued interest in exploring the kind of 'rule-based' behavioural strategies that Gibson advocated. A recurring theme behind all these strategies is the idea that out of the myriad of potential visual signals that could be derived from a moving retinal image (or a binocular pair of images), one or two aspects of the information are especially relevant for controlling a particular motor behaviour. Recently, for example, there has been interest in recording the head and eye movements of people carrying out real-world tasks, such as driving. It is clear from these studies how a sequence of simple visuo-motor rules or sub-tasks could be linked together to achieve a higher-order goal. Take the task of making a cup of tea, for which Land et al. (1999) have recorded the entire sequence of head and eye movements. The individual sub-tasks, such as bringing the hand towards the kettle lid, tend to be straight-forward visually-guided routines when examined individually. By fixating the kettle lid, for example, the observer need only bring the image of the hand onto the fovea (in each eye, if viewed binocularly) to achieve the goal. The complexity of behaviours may therefore evolve in two ways: (i) by increasing the range of different sensory parameters that are available for controlling the motor system (which can increase the accuracy of control, e.g. by adding the second eye's view, or increase the repertoire of possible 'sub-tasks') and (ii) by storing increasingly long sequences of sub-tasks that, when strung together, achieve higher-order goals.

The approaches advocated by Marr and Gibson are not mutually exclusive. Advances in computer vision may help to bring the two together. Mathematical rigour and computational theory are brought to bear here on tasks that, increasingly, must be carried out in unpredictable, ‘natural’ environments. One current research theme that exemplifies that process of resolution is Bayesian inference.

4 Bayes

Bayesian inference is sometimes couched in fearsome mathematical terms, but the basic idea is both straight-forward and highly relevant to understanding animal behaviour.

The brain receives signals from afferent (sensory) fibres. On the basis of these, and of the information it has stored previously, the brain must generate a response (ultimately, a motor response). A reasonable model for this process is that one response is chosen out of a list of possibilities by choosing the most appropriate in the organism’s current context (e.g. most probably rewarded or least probably punished).

It is here that Bayes’ formula is useful. Bayes pointed out that the probability of state S being the case (here, some state of the world, e.g ‘food straight ahead’ or ‘there is a kettle over to my right’) given information I (here, the sensory information the brain receives) is directly proportional to two quantities that can, in principle, be estimated in advance (and hence, in the context of the brain, stored in memory). The first quantity is the ‘prior’ probability of state S occurring, i.e. $\mathsf{P}(S)$. This makes sense intuitively: if you are forced to guess what the current state of the world is and you have no evidence (or highly inconclusive evidence) at the moment, you should guess a likely rather than an unlikely state (these prior probabilities being determined on the basis of previous experience). For example, if the kettle was on your right the last time you looked, it is a reasonable assumption that the fuzzy grey shape on the periphery of your vision is (still) the kettle.

The second quantity incorporates the actual data, I, and gives an indication of how conclusive it is. It is the probability of receiving evidence I given that the current state of the world really is S, normalised by the total probability of getting information I (i.e. summed over all possible states). Again, this makes sense intuitively. For example, fuzzy grey shapes are common in peripheral vision and do not always arise from kettles, so the evidence on its own is inconclusive. Fixating the kettle is a good way to improve the evidence. The higher resolution image is richer (and rarer) and more specific to kettles. In general, if sensory input I is both rare (i.e. $\mathsf{P}(I)$ is low) and also characteristic of state S (i.e. $\mathsf{P}(I|S)$ is high), then information I is good evidence that the world is in state S. Put more succinctly:

$$\mathsf{P}(S|I) = \frac{\mathsf{P}(S)\mathsf{P}(I|S)}{\mathsf{P}(I)} \quad .$$ [1]

Many perceptual phenomena can be explained parsimoniously using a Bayesian approach. Box 3 illustrates one example, in which it helps to predict the pattern of mistakes that people make when judging the direction of motion of a low-contrast object. There are many other examples (e.g. a similar analysis of a 3D motion illusion, Hogervorst and Eagle (1998), and a review by Knill and Richards (1996)).

Box 3 about here

Bayesian inference fits well with all of Marr’s levels of description. It is a useful tool

[1] Bayes’ rule can be derived from an assertion that the probability of S *and* I is equal to that of I *and* S. If the two joint probabilities are expressed in terms of conditional events, this becomes $\mathsf{P}(S|I)\mathsf{P}(I) = \mathsf{P}(I|S)\mathsf{P}(S)$, from which the expression for $\mathsf{P}(S|I)$ can be obtained.

in describing a problem at the level of computational theory, making explicit what is to be computed and the constraints that are to be used to derive output from input. It can be the basis of specific, working models or algorithms and it can be implemented in a number of ways, such as in neural networks. Following Marr's notion of independence between levels, theories of neural architecture in the brain that might carry out this kind of inference can be developed to deal with the generic quantities $\mathsf{P}(S), \mathsf{P}(I)$ and $\mathsf{P}(I|S)$, without reference to specific stimuli (see review by Barlow, 2001).

At the same time, Bayesian approaches fit well into the evolutionary or ecological perspective that Gibson advocated. A simple organism, with a simple behavioural repertoire, needs only to divide information about the organism's state with respect to the world into a small number of categories. It can use its motor system to move between these categories (this is, after all, the only way it can know that a motor movement has been successful). A more complex behavioural repertoire requires a greater number of states to be discriminated reliably. This means that sensory systems must evolve to help an organism discriminate between the contexts in which it will generate different motor outputs. At various stages in a task, the sensory parameters that are most helpful in discriminating (and hence controlling) movements will be quite different, as discussed in section 3. This leads to a view of the cortex as a pool from which evidence can be drawn. From moment to moment, the neurons containing the most relevant information may be located in quite different parts of the cortex, according to the demands of the task.

5 Time

Proposals about how neural signals are combined to give rise to visual percepts include some that do not provide a description of the solution at the level of a computational theory. Examples are the idea that synchronised oscillations in the firing rate of neurons in different parts of the brain (Singer, 1998) or even quantum fluctuations in tubulin molecules (Penrose, 1994) could account for perceptual phenomena. As they stand, such theories have little explanatory power.

A general weakness of these and many other theories is the failure to consider how representations could be built up over time. There is a tendency to assume that neural responses could somehow be combined to generate a vivid reconstruction of the scene in an instant. Not only is this a daunting prospect, but the computational problem is amplified as time is brought into the equation. Having reconstructed the world, heaven forbid that the observer now move their eyes or their head! That would entail a new reconstruction and a new problem of relating it to the one created a moment ago. More promising approaches focus on the small, discrete goals of one epoch (e.g. a period of fixation) and how these could be combined, like the pieces of a jigsaw, into a richer representation (e.g. Ballard et al., 1997; Land et al., 1999; Rensink, 2002).

This view brings Marr and Gibson's ideas together in another way. Gibson emphasised the role of vision as a tool for action. One of the things that makes human vision special is our ability to carry out tasks involving long sequences of movements, each one simple if considered in isolation, to achieve our goals. The processes involved in building up a vivid, detailed visual representation are perhaps best seen as a bi-product of that ability, taking time and being divisible into purposeful steps. But as Marr emphasised, whatever the processes turn out to be, emulating them computationally is the best way to understand them fully.

Acknowledgements

Supported by the Royal Society. I am grateful for comments from Bruce Cumming, Tam Curnow, Miles Hansard and Anya Hurlbert.

References

Ballard, D. H., Hayhoe, M. M., Pook, P. K., and Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20:723–742.

Barlow, H. B. (2001). Redundancy reduction revisited. *Network: computation in neural systems*, 12:241–253.

Cumming, B. G. and DeAngelis, G. C. (2001). The physiology of stereopsis. *Annual Review of Neuroscience*, 24:203–238.

Cutting, J. E., Springer, K., Braren, P. A., and Johnson, S. H. (1992). Wayfinding on foot from information in retinal, not optical, flow. *Journal of Experimental Psychology-General*, 121:41–72.

Dodd, J. V., Krug, K., Cumming, B. G., and Parker, A. J. (2001). Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. *Journal of Neuroscience*, 21:4809–4821.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Glennerster, A., Hansard, M. E., and Fitzgibbon, A. W. (2001). Fixation could simplify, not complicate, the interpretation of retinal flow. *Vision Research*, 41:815–834.

Hogervorst, M. A. and Eagle, R. A. (1998). Biases in three-dimensional structure-from-motion arise from noise in the early visual system. *Proceedings of the Royal Society of London, B, Biological Sciences*, 265:1587–1593.

Knill, D. and Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge University Press.

Land, M. F., Mennie, N., and Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28:1311–1328.

Lappe, M., Bremmer, F., and van den Berg, A. V. (1999). Perception of self-motion from visual flow. *Trends in Cognitive Sciences*, 3:329–336.

Mallot, H. A. (2000). *Computational vision: information processing in perception and visual behaviour*. 2nd edition, MIT Press, Cambridge, Massachusetts.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. New York: W.H. Freeman and company.

Penrose, R. (1994). *Shadows of the mind: a search for the missing science of consciousness*. Oxford University Press, Oxford, UK.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology*, 53:245–277.

Singer, W. (1998). Consciousness and the structure of neuronal representations. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences*, 353:1829–1840.

Ullman, S. (1980). Against direct perception. *Behavioral and Brain Sciences*, 3:373–415.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5:598–604.

Box 1

Level	**Definition**	**Example**
Computational theory	Sets out the goal of a process and an outline of how it can be achieved in principle. This includes defining the input, the output and establishing the constraints that will be used in computing one from the other.	For binocular stereopsis, the input is (at least) the left and right eyes' images; the output could be a depth map of features as viewed from a point midway between the two eyes. An example of a constraint is that a point in an image corresponds to one and only one scene point (since the scene point must be opaque for it to be visible).
Algorithm	Shows how a process is to be carried out. Gives details of how the input and output are represented and a set of rules for the transformation between the two.	Stereo algorithms often use sparse sets of features (such as light/dark boundaries) as their input. Coarse-to-fine algorithms compute depth in blurred versions of the input images and use the results to help solve ambiguities at finer scales.
Implementation	Specifies the physical method for carrying out an algorithm, e.g. in computer hardware or using neurons.	Disparity sensitive neurons have been identified at important stages in the process of stereopsis in animals, some closer to the 'input' stage, others apparently closer to the 'output' stage. However, we do not yet know the algorithm or even the computational theory underlying the process these cells are involved in.

Figure 1: **Marr's three levels for understanding any visual process.** Each level is illustrated using the example of binocular stereopsis (seeing depth with two eyes).

Recent neurophysiological investigations have improved our knowledge of the first stages of processing, in area V1 of the visual cortex, after input from the two retinae have been combined but before the depth of points has been computed unambiguously (reviewed by Cumming and DeAngelis, 2001). Other studies have identified neurons that appear to be beyond the stage at which depth is computed, because firing of these neurons is remarkably predictive of the perceptual choice an animal will make in a depth-related task (e.g. Dodd et al., 2001). There are tantalisingly few synapses between these two neuronal stages, raising the hope that understanding the processes between them may be a tractable problem. The role of these cells in depth perception will only be fully understood, however, when their function can also be described at the levels of an algorithm and a computational theory.

Box 2

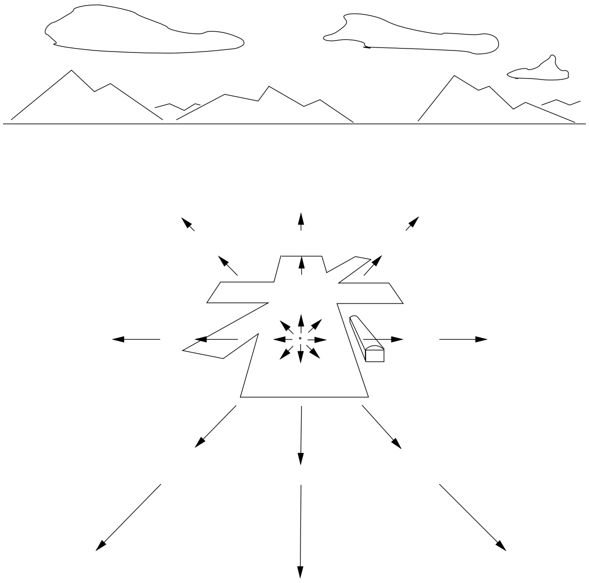

Figure 2: **Using optic flow to land a plane.** As the pilot approaches the air field, visible points move outwards as shown by the arrows. Gibson described this as ‘optic flow’. The point from which the flow emerges indicates the plane’s direction of heading (in computer vision, this is called the ‘epipole’). One strategy for landing the plane is to keep the epipole centred on the runway. (Adapted from Gibson (1979).)

In some ways, the strategy described here is deceptively simple. Gibson said that optic flow was only generated by translation of the observer (movement through space). Rotating the eye does not change the ’optic array’ and so generates no optic flow. (This is not quite enough, however, to define optic flow: a choice must be made about how to relate different optic arrays. A sensible choice is to choose a coordinate frame in which distant objects remain stationary, e.g. the mountains). Retinal flow is more complicated than optic flow. The eyes often rotate with respect to the world as, for example, when you fixate a nearby object and walk past it. The resulting ’rotational flow’ is added to the ’pure’ translational flow (which would be generated if you walked past the object fixating a distant point). Thought of in this way, retinal flow is a confusing mixture of two types of signal. Gibson himself was rather unclear on how the visual system was supposed to use ’optic flow’ given that it must start off with retinal flow.

There are two diverging hypotheses about how the brain deals with retinal flow. One assumes that the visual system extracts the ‘translational’ flow shown by the arrows in this figure, by subtracting the rotational flow component (reviewed by Lappe et al., 1999). Another assumes that the visuo-motor system uses task-specific strategies that avoid computing translational flow (e.g. Cutting et al., 1992; Glennerster et al., 2001). For example, Cutting et al. (1992) suggest that observers could fixate on different points as they walk and, using a simple rule, change their direction of gaze until it is aligned with their direction of heading. The difference between these approaches is at the level of computational theory: they have different goals.

Box 3

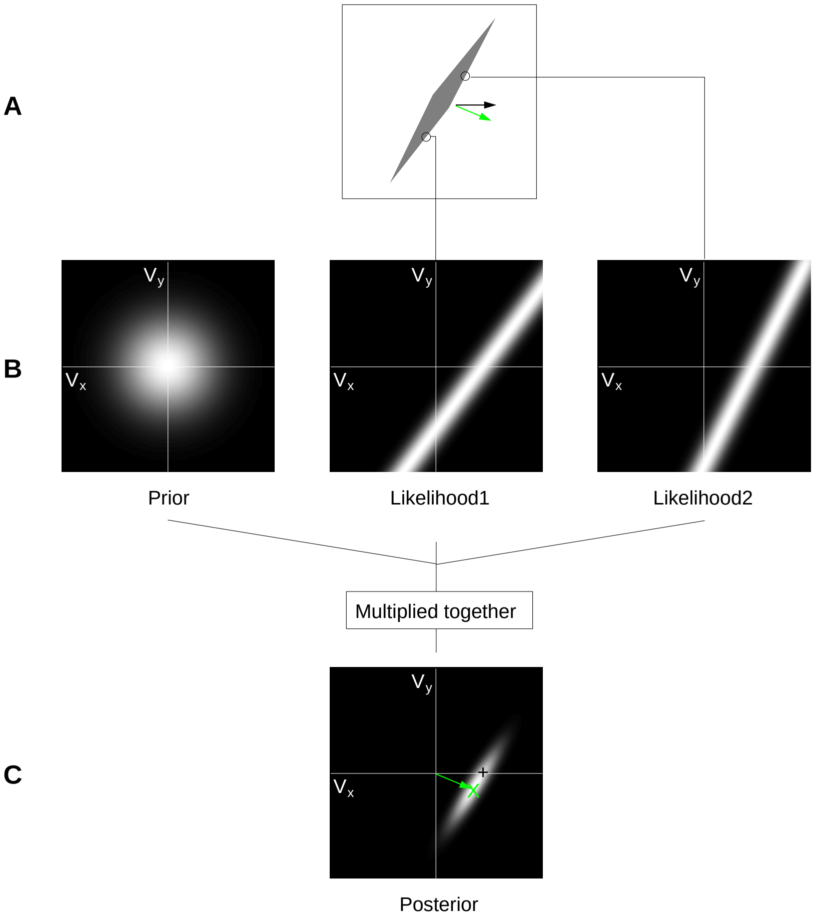

Figure 3: **A Bayesian model of a motion illusion.** When a narrow, low-contrast rhombus as shown in **A** is moved to the right it appears to move down as well (as shown by the green arrow). This can be understood by (i) considering the set of stimuli that *could* have produced the edge motion signals the observer receives and (ii) including a 'prior' assumption that objects tend to move slowly (adapted from Weiss et al., 2002).

The left hand plot in **B** shows the assumed 'prior' probability of velocities in horizontal and vertical directions (V_x, V_y), where intensity is proportional to probability: low velocities are favoured. The centre and right hand plots show the likelihood that an edge moving at each velocity (V_x, V_y) generated the motion observed at the points marked by the circles. For a high contrast edge, these velocities would all fall along a line. (The fact that the velocity is not known uniquely is known as the 'aperture problem': it arises because movement in the direction of the edge produces no local motion signal, e.g. within the circle.) The line is blurred for the low-contrast stimulus because, with some noise in the system, edges moving at other velocities can give rise to the same motion signal.

The plot in **C** is obtained by multiplying the three plots in **B** together. This amounts to following Bayes' rule to calculate the probabilities that the real object had a velocity (V_x, V_y) which gave rise to the two motion signals measured at the circles (known as the 'posterior' probability). (Weiss et al. (2002) did this for all points in the stimulus, giving a very similar result.) The mean (or peak) of the distribution is shifted, as shown by the green 'X', away from the true velocity of the rhombus (shown by the '+'). It predicts the direction of motion seen by subjects when presented with this stimulus.