Frances Egan (Routledge Encyclopedia of Philosophy Online)

Vision

Vision is the most studied sense. It is our richest source of information about the external world, providing us with knowledge of the shape, size, distance, colour and luminosity of objects around us. Vision is fast, automatic and achieved without conscious effort; however, the apparent ease with which we see is deceptive. Ever since Kepler characterized the formation of the retinal image in the early seventeenth century, vision theorists have known that the image on the retina does not correspond in an obvious manner to the way things look. The retinal image is two-dimensional, yet we see three dimensions; the size and shape of the image that an object casts on the retina varies with the distance and perspective of the observer, yet we experience objects as having constant size and shape. The primary task of a theory of vision is to explain how useful information about the external world is recovered from the changing retinal image.

Theories of vision fall roughly into two classes. Indirect theories characterize the processes underlying visual perception in psychological terms, as, for example, inference from prior data or construction of complex percepts from basic sensory components. Direct theories tend to stress the richness of the information available in the retinal image, but, more importantly, they deny that visual processes can be given any correct psychological or mental characterization. Direct theorists, while not denying that the processing underlying vision may be very complex, claim that the complexity is to be explicated merely by reference to non-psychological, neural processes implemented in the brain.

The most influential recent work in vision treats it as an information-processing task, hence as indirect. Many computational models characterize visual processing as the production and decoding of a series of increasingly useful internal representations of the distal scene. These operations are described in computational accounts by precise algorithms. Computer implementations of possible strategies employed by the visual system contribute to our understanding of the problems inherent in complex visual tasks such as edge detection or shape recognition, and make possible the rigorous testing of proposed solutions.

1 Historical background

Theorists of vision have proposed various accounts of the nature of the processing responsible for our perception of size, shape and distance. Geometric models, popular among optic theorists in the seventeenth century, and suggested in some of René Descartes' work on vision (particularly, his Sixth Set of Replies 1641: §9), construe visual processing as a species of mathematical calculation (see Molyneux problem). Geometric models can therefore be seen as precursors of modern-day computational models of vision (see §§4–8 below). According to one geometric model, the visual system computes the distance of an object in the visual field from the angles at which the light from the object strikes each eye, and the distance between the two eyes. Some of the knowledge required for the calculation, including knowledge of the relevant mathematical theorems, was thought to be provided by innate mechanisms, rather than acquired from experience (see Innate knowledge). A significant defect of geometric models is that they failed to provide an account of how the requisite knowledge is made available to the visual system (including, for this calculation, the distance between the eyes, which is not itself perceived, and which changes as the subject grows) or how it is deployed in calculations that were presumed to be unconscious.

The philosopher George Berkeley, in his influential 'Essay Towards a New Theory of Vision' (1709), questioned the psychological reality of the geometric models, arguing in effect that the information upon which the postulated calculations are based is not available to the visual system. Berkeley agreed with the geometric theorists that retinal information alone is insufficient to account for our perception of distance and size, but, consistent with his more general empiricism, he claimed that the process by which we acquire such knowledge is not a species of calculation based on innately specified information, but rather associative and learned.

According to Berkeley, our ideas of distance and size, unlike our ideas of colour, are not really visual ideas at all. Whereas light reflected at different wavelengths affects the retina differentially, and so (the special case of *metamers* aside) gives rise to different colour sensations, light reflected from different distances does not. There is no characteristic retinal pattern associated with something's being 10 feet away. As Berkeley put it in his famous 'one point' argument, 'distance being a line directed end-wise to the eye it projects only one point in the fund of the eye, which point remains invariably the same whether the distance is larger or shorter' (1709: §2). Similarly for size: there is no characteristic retinal pattern produced by our looking at an object that is 6 cubic feet in volume. A larger object placed at a greater distance along the line of sight will have the same geometric effect on the retina. Our ideas of distance and size, Berkeley concluded, derive not from visual experience, but from touch and movement, from the time and effort it takes to make contact with objects, and from the way they feel in our hands. We can tell the distance and size of objects by sight only because we learn to associate visual cues, including sensations caused by the convergence of the eyes and the accommodation of the lens, with ideas originally derived from our tactual sense.

Central to Berkeley's account of distance and size perception is the empiricist doctrine that there are no meaningful abstract ideas, that is, ideas not reducible to sensation (see Empiricism; Sense-data). He rejected the possibility that we might possess abstract spatial ideas that are shared by visual and tactual experience (see Molyneux problem). Later theorists of vision who do not share Berkeley's epistemological and metaphysical assumptions have found his claim that our ideas of distance and size are derived from our sense of touch uncompelling, and recent work on object perception in infants has independently undermined this claim. Nonetheless, Berkeley's discussion of the phenomena to be explained by a theory of vision shaped the field well into the twentieth century.

2 Direct v. indirect perception

The claim that visual perception is not direct or immediate involves more than the truism that some processing of the retinal image is necessary to account for what we see. Ideas or perceptions are thought not to be 'direct' if they are produced by psychological processes. While the notion of a psychological process admits of no precise definition, examples come readily to mind. Any process that occurs in consciousness, such as the association of ideas, is a psychological process, as is any process that involves learning. Mathematical calculation of distance and size based on the prior representation of lines and angles, whether accessible to consciousness or not, is a psychological process (see Intentionality). Since Berkeley's theory and the models of the geometric writers both posit psychological processing of the image (albeit of different sorts), they are considered indirect theories of vision.

The difference between direct and indirect accounts of perception is sometimes characterized as a disagreement over the richness of the stimulus, with direct theorists typically arguing that the stimulus contains more information than indirect theorists have been willing to allow. For example, James J. Gibson (1904–79), a prominent direct theorist, claimed that the input to

the visual system is not a series of static 'time slices' of the retinal image, but rather, the smooth transformations of the optic array as the subject moves about its environment (what Gibson (1979) called 'retinal flow'). But to characterize the fundamental difference between direct and indirect theories as a disagreement over the richness of the stimulus is to misplace the dispute. The issue that separates the two camps concerns neither the amount of information contained in the stimulus, nor even the precise character of this information, but, rather, how the information in the stimulus is accessed and used by the visual system to produce knowledge that is useful to the organism. In other words, it concerns the character of the intervening processes. Direct theorists deny that visual processes can be characterized in terms of ideas, beliefs, representations, knowledge or memories. In other words, they deny that visual processes have any true psychological description. A direct theory explicates any intervening or supplementary processing that occurs in perception in terms of neural structures and processes directly implemented in the brain. Indirect theorists, of course, do not deny that perceptual processes are implemented in neural structures, but they argue that such processes should be characterized at a distinct, psychological, level of description.

Direct theories of perception are sometimes explicitly contrasted with accounts that treat perception as a species of inference, akin to the drawing of a conclusion from premises according to a principle or rule. The nineteenth-century German physicist and physiologist Hermann von Helmholtz argued that the processes underlying visual perception are of the same general sort as inductive generalization employed in scientific reasoning (see Inductive inference; Inference to the best explanation). We will consider below a contemporary approach modeled on Helmholt's idea. The perceptual psychologist Irvin Rock (1983) advanced a view that explicitly treats much of perception as a process of hypothesis generation and testing. But the use of 'inferential' as a blanket term to refer to indirect theories of perception is somewhat misleading. The various processes that can be thought of as psychological (for example, conscious inference, unconscious calculation, habit-based association, and so on) seem too heterogeneous a collection to justify characterizing the entire class in terms of the drawing of conclusions from antecedently established premises.

3 Direct theories of vision

The 'Gestalt' movement of the early twentieth century rejected the view, prevalent since Berkeley, that complex percepts can be analysed into simple sensory components (see Gestalt psychology). According to the Gestalt theorists, perception is holistic: perceptual wholes are not built up out of more basic sensory elements, in the way, for example, that a painting is just the combination of all the paint-covered segments of the canvas. Gestalt theorists claimed further that perception is direct – perceptual processing is not correctly described in terms of psychological or mental processes. The structure of a visual experience is to be explicated in terms of the structure of the underlying brain states, that is, in neurophysiological terms. The Gestalt psychologist Wolfgang Köhler characterized as a physical gestalt any dynamic system that settles into an equilibrium state of minimal energy. A soap bubble forming a perfect sphere is an example of a physical gestalt, as is, Köhler argued, the brain producing an organized percept. Köhler proposed a theory that appealed to electrical fields within the brain to account for perception (and all other mental processes). Gestalt speculative physiology was not borne out by subsequent brain research, which failed to discover evidence of Gestalt mechanisms implicated in perceptual processing. That said, however, work by Gestalt theorists to characterize perception in terms of very general organizational principles (such as *proximity*, the idea that nearby elements in the image tend to be grouped together, or *similarity*, the idea that visually similar elements in the image tend to be grouped together) has proved useful in motivating the search for computational mechanisms that realize and explain these principles.

The psychologist James J. Gibson shared with the Gestalt theorists the belief that visual perception is not mediated by processes characterizable in psychological terms. Gibson argued that indirect theorists have mischaracterized the information in the optical array. If the effective stimulus for the visual system is taken to be retinal flow (the smooth transformations of the optic array as we move about), then, according to Gibson, there are important constancies in the stimulus that indirect theorists have typically missed. There is therefore no need to posit inferences, calculations, memories, association of ideas, or any other intervening psychological process, to explain our perception of size and shape constancy. In addition to brightness and colour, properties directly picked up in the stimulus include, according to Gibson, higher-order properties that remain invariant through movement and changes in orientation. These higher-order invariants specify not only structural properties such as 'being a cube', but also what Gibson called 'affordances', which are functionally significant aspects of the distal scene, like the fact that an object is edible or could be used for cutting.

Two fundamental assumptions underlie Gibson's 'ecological optics': (1) that functionally significant aspects of the environment structure the ambient light in characteristic ways; and (2) that the organism's visual system has evolved to detect these characteristic structures in the light. Both assumptions are controversial. With respect to (2), indirect theorists have complained that Gibson provides no account of the mechanism that allegedly detects salient higher-order invariants in the optical array. His claim that the visual system 'resonates', like a tuning fork, to these properties is little more than a metaphor. But it should be noted that in claiming that perception of higher-order invariants is direct, Gibson is simply advocating that the mechanism be treated as a black box, from the point of view of psychology, because no inferences, calculations, memories or beliefs mediate the processing. (The physiological account of the mechanism's operation will no doubt be very complex.) This claim might be plausible if assumption (1) is true – if there is a physically specifiable property of the light corresponding to every affordance. But for all but the simplest organisms it seems unlikely that the light is structured in accordance with the organism's goals and purposes. More likely, the things that appear to afford eating or cutting or fleeing behaviour structure the light in all kinds of different ways. This likelihood has led indirect theorists to claim that something like categorization – specifically, the bringing of an object identified initially by its shape, colour or texture under a further concept – is at work when an organism sees an object as food, as a cutting implement, or as a predator.

4 Computational models of vision: general approach

The predominant theoretical approach in cognitive psychology in recent years has been computationalism, which treats human cognitive processes, including perceptual processes, as a species of information processing (see Mind, computational theories of). Computational theories of vision attempt to specify the aspects of the external world that are represented by the visual system, and to characterize the operations that derive these representations from the information contained in the retinal image.

One of the most prominent early computational vision theorists was David Marr (1945–80), a researcher in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. While the details of Marr's specific computational model have been challenged by later theorists, his work is of continuing interest to philosophers and psychologists concerned with the foundations of the computational approach to vision. Accordingly, I will use Marr's theory to highlight significant features of the computational approach.

Marr argued in his book Vision (1982) that an information-processing capacity can be analysed at three distinct levels of description. The 'theory of the computation' is a precise

specification of the function computed by the mechanism, in other words, what the mechanism does. For example, the theory of the computation for a particular device may tell us that it adds numbers, or computes averages when given a list of numbers as input. The algorithm specifies the procedure or rule for computing the function, and the implementation level describes how the computation is carried out in neural or computer hardware. The first two levels in the hierarchy – the abstract characterization of the problem and the rule for its solution – exemplify a fundamental commitment of the computational approach: that cognitive processes can be understood in a way that is independent of the particular mechanisms that implement them in the brain.

Computational models treat the visual system as computing from the retinal image a representation of the three-dimensional structure of the distal scene. Marr's theory divides this process into three distinct stages, positing at each stage the construction of a representation that makes explicit (some of) the information contained in the image and represents it in a way that is efficient for later use. Various computational processes, some running in parallel, are defined over these representations. The algorithmic level of description characterizes the procedures the visual system uses to produce increasingly more useful representations of the scene.

Most of the processes that Marr describes are data driven, or 'bottom up' – they operate on information contained in the image, without supplementation by information or beliefs about specific objects and features in the scene. These processes use information about intensity changes across the visual field, or the orientation of surfaces, not such facts as that objects of a particular shape typically make good cutting implements. Marr advocated 'squeezing every ounce of information out of the image' before positing the influx of supplementary knowledge.

Data-driven models of perception have a number of advantages over hypothesis-driven models which appeal to high-level knowledge very early in visual processing. Data-driven processes are generally faster – the visual system does not have to retrieve the relevant piece of specialized knowledge before processing the information in the image – and tend to be more reliable. In Marr's model, the point at which high-level information is available to the visual system marks a distinction between early and late vision. Early visual processes are said to be 'cognitively impenetrable' by the subject's beliefs about the world (see Modularity of mind). As a consequence, they cannot be influenced by learning.

Marr emphasized the importance of the 'topmost' level of description – the theory of the computation – in developing accounts of human cognitive capacities. He noted that there is no point attempting to describe how a mechanism works before knowing what it does. A crucial first step in constructing a theory of a perceptual capacity is discovering very general constraints on the way the world is structured that enable adapted organisms to solve perceptual problems in their normal environments. An example should make the point clear. Marr's student and colleague Shimon Ullman (1979) proved that three distinct orthographic views of four non-coplanar points are sufficient to determine the three-dimensional structure of a rigid body (the 'structure from motion' theorem). If a body is not rigid, much more information is required to compute its shape. In a world such as ours, where most things are relatively rigid, a visual system built (that is, adapted) to assume that the objects in its environment are rigid would be able to compute the structure of those objects more easily and quickly than a visual system that had to consider the many non-rigid interpretations consistent with the data. Accordingly, Marr posited a mechanism that given three views of four non-coplanar points as input computes the unique rigid interpretation consistent with the data.

Recall Berkeley's objection to the geometric theorists' accounts of size and distance perception. He claimed that the information required for the postulated calculations was not generally available to the visual system, nor to the organism. Such a criticism, if true, is devastating for a computational account of a cognitive capacity. Any computational theory that posits processing beyond the computing capabilities of the mechanism, or that relies on information unavailable to the mechanism, is a non-starter as a biological model. An important lesson of Marr's work is that the theorist must attend to the general structure of the organism's environment before attempting to characterize computational mechanisms, because the environment determines the nature of the computational problems that the organism's visual system needs to solve. The perceptual systems of adapted organisms can be assumed to 'exploit' very general information about the environment. Consequently, the problems they have to solve may be simpler and computationally more tractable than might initially be assumed.

The work by Gestalt theorists to characterize perception in terms of general organizational principles, mentioned above, can be seen as the articulation of general environmental constraints and hence as contributing to the specification of theory at the top-most level in Marr's hierarchy. These principles are justified by reference to very general features of the environment. For example, *proximity*, the idea that nearby elements tend to be grouped together, reflects the fact that objects are cohesive.

5 Computational models of vision: modularity

Another characteristic feature of Marr's theory is that it treats the visual system as comprising a number of individual components or modules that can be analysed independently of the rest of the system. A 'module' is, by definition, cognitively impenetrable: its operation is not influenced by information external to it that may be available to the cognitive system as a whole, for example, information in the system's memory (see Modularity of mind). Marr posited a module responsible for computing three-dimensional structure from apparent motion, another for computing depth from disparity information available in stereo images, a third for computing shape from shading. Each of these modules is designed to exploit general environmental constraints in the manner that the 'structure from motion' module, described above, incorporates the rigidity assumption.

The various modules operate in parallel, and since they yield information about the depth of the distal scene from different input data, they may give inconsistent results. This is an advantage for the organism, because in cases where the general environmental constraints assumed by a processing module do not hold, the output of the module is subject to correction by another module operating on different data, and exploiting different environmental constraints. For example, imagine a non-rigid mass of jelly moving through space. Since the 'structure from motion' module is built to assume rigidity it will probably give an incorrect interpretation of the jelly's structure. But its output is then likely to be inconsistent with, and correctable by, the output of modules operating on shading or disparity information, which, though they exploit other environmental constraints, do not assume rigidity.

The principle of modular design has an evolutionary rationale. Modular processes are typically fast, because a time-consuming search of general memory is avoided. And assuming that the constraints governing a module's operation are generally true, the process will normally be reliable. Commitment to the principle of modular design makes the computational theorist's job easier, since modular processes can be studied and modeled without the theorist knowing how more central reasoning systems work. For all their theoretical advantages, however, modules do pose a general problem. The theorist has to explain how the outputs of various modular processes are combined in a single

representation of the structure of the scene. The possibility of inconsistent results from different modules suggests that this is a non-trivial problem.

In general, then, the visual processes posited in Marr's theory have three important features. They are data-driven, adapted to exploit general environmental constraints, and modular. The visual system, according to Marr, computes a series of intermediate representations of distal information, culminating in a representation of the three-dimensional structure of the scene. The input to the system is the image on the retina, in effect, a grey-level intensity array. The initial processing of the image produces what Marr called the 'primal sketch', a representation of the way that light intensities change over the visual field. The primal sketch makes explicit precisely the information that is required for subsequent processing. Discontinuities in intensity tend to be correlated with significant features of the scene, that is, object boundaries, although it is too early at this stage to assume that all sharp intensity changes in the image indicate edges in the world. Some may be produced by changes in illumination or surface reflectance (see Colour and qualia).

The various processing modules described above operate on aspects of the information contained in the primal sketch. The results are encoded in a representation that Marr called the '2.5-dimensional sketch'. It makes explicit the depth and surface orientation of the scene, and is the input representation for later visual processing. The visual system is assumed to be cognitively impenetrable up to the production of the 2.5-dimensional sketch, hence its operation to this point cannot be influenced by learning.

6 Computational models of vision: object recognition

Late or high-level visual processes use the representations of depth and surface orientation produced by early vision for tasks such as object recognition, locomotion and visually guided manipulation. Marr's own account of late visual processing is rather sketchy. His concrete proposals concern the computational level of description, with little or no detail supplied at the algorithmic level. In general, computational models of high-level vision are not as well developed as accounts of early visual processes. The difficulty is due in part to the fact that later processing is hypothesis- (or goal-) driven, and hence cognitively penetrable. The input to these processes is not limited to information contained in the image. Object recognition, for example, makes use of specific knowledge about objects in the world. This knowledge is usually characterized as a catalogue of object types stored in long-term memory. It is worth noting that only at this rather late stage does the visual system do anything like identify what Gibson calls 'affordances', and in computational accounts such identification is typically treated as a process of categorization, in other words, as a psychological process (see Concepts §1).

According to the simplest models, recognizing an object currently in view involves comparing it with previously stored views of objects and selecting the one that most resembles it. A problem with this approach is that it fails to explain our ability to recognize objects from novel views that do not straightforwardly resemble any previously stored views.

More promising are accounts that treat object recognition as associating with the current view of the object a description of the object type, perhaps in addition to previously stored views of representative examples. Here again, different approaches are possible. 'Invariant-property' accounts assume that the set of possible retinal projections of objects typically have higher-level invariant properties that are preserved across the various transformations that the object may undergo. Such proposals face the same problem as Gibson's account of higher-order invariants. For most object types it has proved impossible to find specifiable properties of the image that are common to all possible recognizable views.

The ‘decomposition’ approach to object recognition maintains that objects are identified on the basis of prior recognition of their component parts. An assumption of this approach is that the relevant part–whole relations are invariant and detectable in all possible views where the subject would recognize the object. According to Irving Biederman’s ‘recognition by components’ theory (1990), a given view of an object can be represented as an arrangement of simple primitive volumes called ‘geons’ (for ‘geometric icons’). Geons can themselves be characterized in terms of viewpoint-invariant properties, and, proponents of the theory claim, are recognizable even in the presence of visual noise. In general, though, the decomposition approach to object recognition has proved to be fairly limited in its application. Many objects do not decompose in a natural way into easily characterizable parts; and for many of those that do the decomposition is insufficient to specify the object in question.

A third strategy, known as the ‘alignment’ approach, suggests that the visual system detects the presence of transformations between the current view of an object and a stored model, and can ‘undo’ the transformation to achieve a correspondence between the two. For example, suppose that the current view of the object differs from the model stored in memory because the object has undergone a three-dimensional rotation and moved further away from the viewer. On the current proposal, the visual system first detects the nature of the transformations, and then performs them in reverse on the current view to bring it into ‘alignment’ with the stored model (assuming that the object is rigid). The main problem for this approach, as for the other proposals, is its limited applicability. It is only feasible for a small range of possible transformations that an object can undergo (for example, rotation and scaling) and then only for a limited range of objects. (Imagine detecting and ‘undoing’ the rotation of a crumpled piece of newspaper.)

‘Mixed’ approaches to object recognition attempt to extend the range of applicability of the decomposition and alignment approaches by combining elements of the two, positing separate identification systems that operate in parallel. While mixed accounts appear promising, they face the additional burden of explaining how the outputs of the two recognition systems are combined.

7 Bayesian models of vision

Computational models based on Bayesian decision principles are currently very popular. This approach follows Helmholtz in treating vision as a species of unconscious inference, in particular, as probabilistic inference. Bayesian theories treat the visual system as an ideal observer that uses prior knowledge about visual scenes and information in the image to infer the most probable interpretation of the image.

The fundamental idea underlying Bayesian perceptual models is that the *posterior probability* of a possible real world structure S is proportional to the product of the *prior probability of S* (that is, the probability before receiving the stimulus I) and the *likelihood* (the probability of I given S). Prior probability distributions in typical applications of the Bayesian strategy represent knowledge of the regularities governing object shapes, constituent materials, and illumination, and likelihood distributions represent knowledge of how images are formed through projection on the retina. Some examples of prior knowledge that figure in Bayesian models are that solids are more likely to be convex than concave and that the light source is above the viewer.

The Bayesian approach provides a framework for taming the ambiguity and complexity in natural images. Perception in the Bayesan framework is explicitly seen as a trade-off between image reliability – p(I/S) – and the prior p(S). The less likely the image is given a structure –

in other words, the more ambiguous the image – the greater the influence of prior knowledge in yielding a non-ambiguous percept. Some perceptions may be more data-driven, and others more knowledge driven. The Bayesian framework provides a schema for explicitly comparing the relative contributions of image data and prior knowledge in alternative proposals. Perceptual constancies – the fact that the visual system is able to detect a fixed structure in successive retinal transformations due to movement, or viewpoint or illuminant changes – are modeled in the Bayesian framework as *discounting*, where the confounding variables – motion, or viewpoint, or illuminant – are discounted in the computation by integrating them out, or summing over them.

Bayesian models rely on statistical analysis of natural images and their real-world causes to arrive at plausible hypotheses concerning image formation (p(I/S)) and prior knowledge about naturally occurring structures (p(S)). Uncovering statistical regularities relating image features to object or scene properties has enabled theorists to design systems that group images consistent with the natural constraints noted above (such as that nearby edges with similar orientations belong to the same contour). Such work has yielded computer vision solutions for edge detection, face recognition, interpretation of bodily movement, and provided insight into the functional nature of certain kinds of visual illusions.

The Bayesian framework affords several advantages for the study of human vision. Perhaps most obviously, it provides a convenient and natural framework for studying all aspects of perception in a unified manner, by treating perception as a Bayesian decision problem. Secondly, Bayesian methods allow the development of quantitative theories at Marr's topmost level, avoiding premature commitment to specific neural mechanisms. Thirdly, Bayesian theories explicitly model uncertainty, and hence are an important tool in understanding how the visual system might combine large amounts of objectively ambiguous information to arrive at percepts that are rarely ambiguous. Finally, as noted above, it provides an explicit account of the interaction between information in the stimulus and prior knowledge of the world.

Nonetheless, Bayesian visual modeling raises some pressing questions about the appropriateness of particular Bayesian models and of the Bayesian approach more generally for understanding human vision. One question concerns how the visual system knows the relevant priors. Some priors, or strategies for learning priors, are assumed to be innate, encoded in our genes. For the reasons discussed in §4, the idea of innate knowledge available to the visual system is quite plausible for general natural constraints but less so for specific knowledge concerning visible properties (shapes, textures, etc.) of specific objects. For those priors which are not plausibly innate, the question is whether the visual system can learn the relevant probability distributions p(S) and p(I/S) from the available data. There are further issues concerning priors. It is likely that some probability distributions will be context sensitive; compare, for example, a forest with a city scene. It is conceivable (in fact, likely) that more than one prior will be applicable in a given context. What does system do when priors are inconsistent? How are priors updated when the environment changes? These questions suggest directions for further research.

Another issue concerns the idealization inherent in the Bayesian framework itself. As mentioned above, this has certain advantages, notably, generality and simplicity. But because human vision is limited not just by the partial nature of the information available – a feature nicely modeled in the Bayesian framework – but also by the available neural hardware, we might expect significant departures from optimality. The assumption of logical omniscience central to Bayesian epistemology – that degrees of belief satisfy the probability

laws – is an issue for Bayesian perceptual theories as well. Is it reasonable to assume that the visual system knows the probability calculus and operates according to it?

8 Computational models of vision: problems and prospects

The most common criticism of computational models of human cognitive capacities, including accounts of our perceptual abilities, is that they are unable to approximate actual human performance. It is true that many impressive computer models fail miserably in the real world. Sometimes they fail because the information required is not available to the mechanism. As Marr emphasized, the computational theorist can try to avoid this problem by first attempting to characterize the computational problems that perceptual mechanisms, in their natural context, are required to solve, a process that involves discovering general environmental constraints that perceptual mechanisms of adapted organisms can be expected to exploit.

But the study of biological visual systems faces additional hurdles. Even if the information on which a posited process runs is in some abstract sense 'in the data,' the input may be too 'noisy' for the mechanism to make use of it. Computational theorists are of course aware of this problem. Some of the processing posited by computational accounts, especially in early vision, involves the elimination of extraneous or irrelevant information in the image. (For example, the primal sketch in Marr's account, which represents intensity changes in the image, does not preserve the absolute values of intensity gradients at every point in the grey-level array.) Bayesian models, in particular, attempt to isolate and discount confounding varlables. Additionally, the theorist must eventually find neural hardware capable of doing the computationally characterized job, before being confident that the model is biologically feasible. Given the difficult nature of the task it is unlikely that a complete computational account of vision is just around the corner. None the less, computational theorists make an important contribution to our understanding of vision by their careful study of the nature of the problems to be solved by visual mechanisms, although the solutions they offer are properly evaluated by their performance in the real world.

An alternative style of computational model may ultimately prove better suited to explicating human vision than models, such as Marr's, that treat perceptual processing as rule-governed operations defined over representations. In 'connectionist' computational architectures information is typically represented by patterns of activation over a connected network of units or nodes. Connectionist processes are explicated at a level distinct from the neurological or implementational. Connectionist cognitive models typically appeal to representations, memory and learning, hence they qualify as indirect; although connectionist accounts of representation, memory and learning differ in significant respects from more traditional computational accounts (see Connectionism). Connectionist theorists have claimed that their models are better able to handle noisy input and 'multiple simultaneous constraints' characteristic of real-world processing situations, though traditional computationalists have disputed this claim. Many Bayesian models lend themselves to implementation in parallel networks. Indeed, despite the significant idealization imposed by the Bayesian framework itself, Bayesian models may prove more amenable to integration with neurological accounts than traditional 'representationalist' models such as Marr's. Some Bayesian models are designed specifically to be consistent with known neural mechanisms, with the prior and likelihood functions implemented in the model by synaptic weights. Whether the 'transparency' of these models from the neurological perspective proves ultimately to be a virtue will depend on whether the empirical predictions the models make possible are borne out.

References and further reading

Berkeley, G. (1709) ‘An Essay Towards a New Theory of Vision’, in The Works of George Berkeley, Bishop of Cloyne, vol. 1, ed. A.A. Luce and T.E. Jessop, Edinburgh: Thomas Nelson, 9 vols, 1948–57. (Referred to in §1.)

Biederman, I. (1990) ‘Higher-Level Vision’, in D.N. Osherson et al. (eds) Visual Cognition and Action: An Invitation to Cognitive Science, vol. 2, Cambridge, MA: MIT Press. (An example of the ‘decomposition’ approach to object recognition.)

Descartes, R. (1637) ‘Optics’, in The Philosophical Writings of Descartes, trans. J. Cottingham, R. Stoothoff and D. Murdoch, Cambridge: Cambridge University Press, 1985, vol. 1, 152–75. (Discourses 5 and 6 are particularly relevant.)

Descartes, R. (1641) ‘Author’s Replies to the Sixth Set of Objections’, in The Philosophical Writings of Descartes, trans. J. Cottingham, R. Stoothoff and D. Murdoch, Cambridge: Cambridge University Press, 1984, vol. 2, esp. §9: 294–6. (Referred to in §1 – Descartes’ ‘intellectualist’ theory of vision.)

Fodor, J.A. and Pylyshyn, Z. (1981) ‘How Direct is Visual Perception?: Some Reflections on Gibson’s "Ecological Approach"’, Cognition 9: 139–96. (A critical discussion of Gibson’s direct theory of perception. Includes detailed argument but no technicality.)

Gibson, J. (1979) The Ecological Approach to Visual Perception, Boston, MA: Houghton Mifflin. (The most developed statement of Gibson’s theory of perception.)

Helmholtz, H. von (1950) Treatise on Physiological Optics, ed. J. Southall, New York: Dover, 3 vols. (Influential nineteenth-century account of perceptual processing as a species of inference.)

Hinton, G.E. (1992) ‘How Neural Networks Learn from Experience’, Scientific American 267 (3): 144. (Includes a discussion of connectionist models of shape recognition.)

Kersten, D., Mamassian, P., and Yuille, A. (2004) ‘Object Perception as Bayesian Inference’, Annual Review of Psychology 55: 271-304. (A general discussion of the Bayesian framework applied to object perception.)

Kersten, D. and Yuille A. (2003) ‘Bayesian Models of Object Perception’, Current Opinion in Neurobiology 13: 1-9. (A useful short introduction to Bayesian models of vision.)

Marr, D. (1982) Vision, New York: Freeman Press. (Somewhat technical, but includes a clear account of the rationale behind the computational approach to vision.)

Paragios, N., Chen, Y., and Faugeras, O. eds. (2006) Handbook of Mathematical Models in Computer Vision, New York: Springer. (A comprehensive survey of recent work in computational vision. Very technical.)

Rao, R., Olshausen, B., and Lewicki, M. eds. (2002) Probabilistic Models of the Brain, Cambridge, MA: MIT Press. (A survey of probabilistic models of perception and neural function, including Bayesian models.)

Rock, I. (1983) The Logic of Perception, Cambridge, MA: MIT Press. (An account of perceptual processing as a form of hypothesis formation and testing.)

Schwartz, R. (1994) Vision: Variations on Some Berkelian Themes, Oxford: Blackwell. (A useful discussion of historical work on the problems of vision. Also includes a chapter on Gibson's theory.)

Ullman, S. (1979) The Interpretation of Visual Motion, Cambridge, MA: MIT Press. (A detailed analysis of the computations involved in visual motion perception. Cited in §4.)